Combinaisons markoviennes et semi-markoviennes de modèles de régression. Application à la croissance d'arbres forestiers.

Florence Chaubert-Pereira

Université Montpellier II
Sciences et techniques du Languedoc

Thèse

présentée en vue d'obtenir le grade de

Docteur de l'Université Montpellier II

Discipline : Mathématiques appliquées
Formation Doctorale : Biostatistique
École Doctorale : Information, Structures et Systèmes

par

Florence CHAUBERT-PEREIRA

le 05 Novembre 2008

Titre:

Combinaisons markoviennes et semi-markoviennes de modèles de régression. Application à la croissance d'arbres forestiers.

devant le jury composé de :

M.	Gilles Ducharme	Université Montpellier II	(Président)
M.	Eric Moulines	ENST	(Rapporteur)
M.	Stéphane Robin	AgroParisTech	(Rapporteur)
M.	Francis Colin	INRA	(Examinateur)
M.	Yann Guédon	CIRAD	(Directeur de thèse)
M.	Christian Lavergne	Université Montpellier III	(Co-directeur de thèse)
M^{me}	Catherine Trottier	Université Montpellier III	(Membre invitée)

Hier est derrière, demain est un mystère, mais aujourd'hui est un cadeau,
C'est pourquoi on l'appelle le "présent"...
Oogway

Remerciements

Ce travail de thèse a été réalisé au sein de l'Unité Mixte de Recherche CIRAD *Développement et Amélioration des Plantes* et plus particulièrement au sein de l'équipe-projet INRIA Virtual Plants. Bien que je sois l'auteur principal de cette thèse, avant d'en commencer l'exposé, je souhaite remercier celles et ceux qui y ont apporté leur contribution.

Mes premiers remerciements s'adressent à Yann Guédon. Il m'a encadré et accompagné durant ces trois années tout en me laissant une certaine liberté dans mes choix. Je le remercie pour la confiance qu'il m'a accordée. J'ai bénéficié au cours de ma thèse d'un encadrement de qualité. En effet, Yann a toujours pris du temps pour m'expliquer, discuter et échanger. Sa grande disponibilité, la passion et l'enthousiasme qui l'animent dans ses activités de recherche, son sens du perfectionnement et son efficacité ont fait de cette thèse un vrai moment de plaisir. Je tiens à lui témoigner ma plus sincère gratitude pour tout ce qu'il m'a appris et pour tout le temps qu'il m'a consacré.

Ces remerciements sont à partager avec Catherine Trottier, qui s'est impliquée de manière remarquable dans cette thèse grâce à ses conseils experts et ses relectures approfondies. Son sens aigu du perfectionnisme a donné sa dimension actuelle à cette thèse. J'ai découvert au travers de nos nombreux échanges une personne avec de grandes qualités humaines et professionnelles. Merci.

Je remercie également Christian Lavergne pour avoir encadré mes travaux. Merci pour son optimiste, son enthousiasme et ses remarques avisées.

Ce travail s'inscrivait dans la continuité de la thèse de Carine Véra. Je tiens à la remercier pour son aide lors du passage de relais.

J'adresse de sincères remerciements aux membres de mon jury de thèse :

- à Gilles Ducharme pour avoir accepté de présider ce jury et pour m'avoir fait confiance au niveau de l'enseignement. Il fut au démarrage de cette aventure en m'acceptant au DEA de Biostatistique en cours d'année. Merci.
- à Stéphane Robin et Éric Moulines, pour avoir accepté d'être rapporteurs de ma thèse et pour leurs remarques avisées et constructives.
- à Francis Colin, pour avoir examiné ce travail et pour avoir apporté son expertise biologique. Merci pour l'intérêt porté à ce travail ainsi que pour les critiques avisées.

J'adresse également de sincères remerciements aux membres de mon comité de thèse :

- à Frédéric Mortier pour son enthousiasme, sa gentillesse et ses conseils. Il m'a fait partagé sa passion pour les statistiques au cours de mon stage de DEA. C'est éga-

lement lui qui a trouvé les mots pour me motiver à me lancer dans cette aventure. Je le remercie tout simplement d'avoir cru en moi.
– Christèle Robert-Granié pour les conseils avertis qu'elle m'a donnés et pour les discussions enrichissantes que nous avons eues lors de ces comités.
– Yves Caraglio pour son expertise biologique.

Cette thèse étant motivée par des applications biologiques, j'ai eu la chance de rencontrer des biologistes qui m'ont ouvert leur monde et appris beaucoup de choses. Mes remerciements s'adressent tout particulièrement à Yves Caraglio. Il a toujours su se rendre disponible et pédagogue pour m'éclaircir sur la botanique. Cette collaboration fut pour moi une expérience très enrichissante et j'espère vivement qu'elle ne s'arrêtera pas là. Merci également à Patrick, Sylvie, Olivier et Evelyne. Ils ont vraiment tous su se mettre à ma portée.

Je tiens à exprimer toute ma gratitude à Christophe Godin pour m'avoir accueilli au sein de l'équipe-projet Virtual Plants pendant ces trois années de thèse. Merci pour ses nombreux conseils. Merci aussi à la joyeuse bande de l'équipe : Christophe P. (pour sa gentillesse et son aide), David et Szymon (mes compagnons de fortune) ainsi que tous les autres : Fred, Jérome, Romain, Mohamad, Damien, Samuel, JB, Mikaël, ... Merci également aux secrétaires de l'équipe : Valérie et Annie.

Je remercie les professeurs Altman et Albert pour leur disponibilité. Ils n'ont pas hésité à me fournir les données médicales utilisées dans leurs travaux. Je remercie également Météo France pour nous avoir gracieusement donné les données météorologiques.

Durant ces années de thèse, j'ai eu la chance d'enseigner à l'université Montpellier 2. Je remercie donc les responsables du CIES pour les formations qu'ils m'ont apportées ainsi que les différentes personnes pour lesquelles j'ai fait mes enseignements.

Je profite de cette occasion pour remercier l'ensemble de mes professeurs de mathématiques et de statistique qui m'ont donné le goût pour ces disciplines. Je remercie plus particulièrement Jean-Jacques Téchéné qui m'a fait découvrir le vaste monde des statistiques.

Un grand merci à ma famille pour la confiance qu'elle m'accorde. Sans leur soutien, je n'aurais pas pu faire de longues études. Je leur en serai éternellement reconnaissant : *Sans eux, je ne serai rien.* Je tiens également à remercier tous mes amis de Bugnein (et alentours) et plus particulièrement Mathieu, un témoin dont tout le monde rêverait. L'exemple de la p.63 leur est dédicacé...

Bruno et moi avons partagé nos trois années de thèse. Aussi, je tiens à le remercier pour sa disponibilité, son soutien et ses encouragements. Il sait toujours trouver les bons mots dans les moments de doute et m'a apporté cette sérénité nécessaire à la réalisation d'un travail de thèse. Nous terminons ensemble l'étape de la thèse mais restons côte à côte pour la suite du chemin...

Table des matières

Glossaire

Sigles

GHSMC : Gaussian Hidden Semi-Markov Chain (semi-chaîne de Markov cachée gaussienne)

GLM : Generalized Linear Model (modèle linéaire généralisé)

GLMM : Generalized Linear Mixed Model (modèle linéaire généralisé mixte)

HMC : Hidden Markov Chain (chaîne de Markov cachée)

HMM : Hidden Markov Model (modèle de Markov caché)

HSMC : Hidden Semi-Markov Chain (semi-chaîne de Markov cachée)

LMM : Linear Mixed Model (modèle linéaire mixte)

MS-GLM : Markov Switching Generalized Linear Model (combinaison markovienne de modèle linéaire généralisé)

MS-GLMM : Markov Switching Generalized Linear Mixed Model (combinaison markovienne de modèle linéaire généralisé mixte)

MS-LM : Markov Switching Linear Model (combinaison markovienne de modèle linéaire)

MS-LMM : Markov Switching Linear Mixed Model (combinaison markovienne de modèle linéaire mixte)

SMS-GLM : Semi-Markov Switching Generalized Linear Model (combinaison semi-markovienne de modèle linéaire généralisé)

SMS-GLMM : Markov Switching Generalized Linear Mixed Model (combinaison semi-markovienne de modèle linéaire généralisé mixte)

SMS-LM : Semi-Markov Switching Linear Model (combinaison semi-markovienne de modèle linéaire)

SMS-LMM : Markov Switching Linear Mixed Model (combinaison semi-markovienne de modèle linéaire mixte)

Notations

a	Individu
N	Nombre d'individus
t	Temps
d	Date (ou année)
h_a	Date de la première mesure pour l'individu a
H_a	Date de la dernière mesure pour l'individu a
$T_a = H_a - h_a + 1$	Longueur de la séquence observée pour l'individu a
$D = \max_a H_a - \min_a h_a + 1$	Nombre de dates différentes
$t_a(d) = d - h_a + 1$	Temps associé à la date d pour l'individu a
$T = \sum_{a=1}^{N} T_a$	Longueur cumulée de toutes les observations
J	Nombre d'états
Y	Variable aléatoire observée
y	Réalisation de la variable aléatoire observée Y
X	Covariables
S	Variable aléatoire cachée
s	Réalisation de la variable aléatoire cachée S
y_{at}	Observation de l'individu a au temps t
s_{at}	État dans lequel se trouve l'individu a au temps t
$Y_{a1}^{T_a} = y_{a1}^{T_a}$	Suite des variables observées et de leurs réalisations
	$Y_{a1} = y_{a1}, \ldots, Y_{aT_a} = y_{aT_a}$ relatives à l'individu a
$S_{a1}^{T_a} = s_{a1}^{T_a}$	Suite des variables cachées et de leurs réalisations
	$S_{a1} = s_{a1}, \ldots, S_{aT_a} = s_{aT_a}$ relatives à l'individu a
ξ	Effet aléatoire "individuel"
ξ_a	Effet aléatoire "individuel" pour l'individu a
λ	Effet aléatoire "temporel"
λ_d	Effet aléatoire "temporel" pour la date d

θ	Ensemble des paramètres du modèle
π_j	Probabilité initiale d'être dans l'état j
p_{ij}	Probabilité de transition de l'état i à l'état j
β	Paramètres des effets fixes
σ^2	Variance résiduelle
τ^2	Variance aléatoire associée à ξ
ς^2	Variance aléatoire associée à λ
Γ^2	Variance totale

$\phi(y; \mu, \sigma^2)$	Densité de la distribution gaussienne de moyenne μ et de variance σ^2
$I()$	Fonction indicatrice
U'	Transposée de la matrice U

Introduction

L'ARCHITECTURE des plantes est devenue un objet d'étude à part entière dans les années 70 avec les travaux fondateurs de Hallé et Oldeman (1970). L'étude de l'architecture des plantes fait nécessairement appel à une connaissance approfondie de la morphologie végétale, de la systématique et de l'écophysiologie des plantes. Au début des années 90, la croissance et la structure de la plante n'avaient quasiment pas fait l'objet de développements de méthodes d'analyse spécifiques à l'exception du domaine de la dendrochronologie qui consiste en l'étude des largeurs ou surfaces de cernes successifs du bois en relation avec les conditions climatiques (Fritts, 1976). Aussi, ce sujet était très en retard par rapport à d'autres champs de la biologie (biologie moléculaire avec les méthodes d'analyse de séquences d'ADN et de séquences protéiques, écologie avec par exemple les modèles de capture-recapture ou encore le domaine biomédical avec l'application des méthodes d'analyse de survie ou d'analyse de données longitudinales). A partir des années 90, l'étude de la plante à une échelle macroscopique a connu son essor à divers niveaux tels que l'architecture de la plante (Durand et al., 2005), le rendement des forêts (Candy, 1997), les composantes de la croissance d'arbres (Véra, 2004; Guédon et al., 2007) ou encore le processus de croissance (Ninomiya et Yoshimoto, 2008). Le présent travail a pour objet de répondre à des problèmes actuels de recherche sur le développement des plantes à l'échelle macroscopique en proposant de nouvelles méthodes statistiques pour l'analyse de données de croissance.

Problématique biologique

Le développement d'un arbre peut être reconstruit rétrospectivement à une date d'observation donnée à l'aide de marqueurs morphologiques (tels que des cicatrices de cataphylles ou de branches qui aident à délimiter les pousses annuelles successives) correspondant à des événements passés (Barthélémy et Caraglio, 2007). La croissance observée, caractérisée par exemple par la longueur de pousses annuelles successives le long du tronc d'un arbre, est supposée être principalement le résultat de trois composantes : une composante ontogénique, une composante environnementale et une composante individuelle. L'analyse de données de croissance d'arbres par Guédon et al. (2007) a mis en évidence des hypothèses biologiques pour la modélisation statistique de ces données :

• La composante ontogénique d'un arbre est supposée être structurée comme une succession de phases de croissance stationnaires liées à sa morphogénèse. Les changements de phase sont supposés spécifiques à chaque arbre, c'est-à-dire que les phases de crois-

sance sont asynchrones entre arbres. Les études sur l'architecture des plantes ont souligné le rôle central joué par la composante ontogénique dans le développement des plantes (Barthélémy et Caraglio, 2007).

• La composante environnementale est supposée prendre la forme de fluctuations synchrones entre arbres. Cette composante peut être vue comme une composante "population" par opposition à la composante individuelle. Hanson et al. (2001) ont étudié en détail la réponse de l'arbre aux changements de conditions environnementales. Les facteurs environnementaux qui modulent le développement d'un arbre sont principalement d'origine climatique telles que la température ou la pluviométrie. La sensibilité de la croissance d'un arbre à ces facteurs climatiques est plus ou moins importante selon l'âge de l'arbre.

• La composante individuelle correspond à la croissance propre à chaque arbre. En effet, même si des arbres sont soumis aux mêmes conditions environnementales, d'autres facteurs peuvent entraîner une hétérogénéité de la croissance entre ces arbres. La composante individuelle est supposée être principalement due à des facteurs génétiques, à des attaques de parasites, à des maladies ou encore à la compétition entre arbres pour les ressources (lumière, eau et nutriments).

Le but de ce travail est d'une part de modéliser conjointement les phases de croissance, l'influence des facteurs environnementaux et l'hétérogénéité inter-individuelle sur la base des hypothèses biologiques présentées précédemment et d'autre part de répondre aux questions suivantes : Quels rôles jouent les facteurs environnementaux sur la croissance ? Ces rôles sont-ils différents selon l'âge des arbres ? Quelle est la part de variabilité induite par l'hétérogénéité inter-individuelle ? Cette variabilité change-t-elle avec les phases de croissance ? Le statut d'un arbre par rapport à l'arbre "moyen" change-t-il avec les phases de croissance ?

Des modèles de regression et modèles markoviens cachés aux combinaisons markoviennes et semi-markoviennes de modèles de regression

Nos données prennent la forme de séquences ou séries temporelles structurées en zones homogènes. Bien qu'en analyse de la croissance d'arbres forestiers, ces zones peuvent se succéder de manière transitoire ou récurrente, nous nous plaçons dans un cadre générique. L'analyse de telles séquences s'appuie le plus souvent sur des modèles markoviens cachés. Les chaînes de Markov cachées sont historiquement le premier modèle de la famille des modèles de Markov cachés. Introduites à la fin des années 60 (Baum et Petrie, 1966), les chaînes de Markov cachées ont connu un grand succès au milieu des années 70 (Baker, 1975; Jelinek, 1976) en reconnaissance automatique de la parole pour la modélisation de zones de signal homogènes. Ce succès est à l'origine de la diffusion de ces modèles dans de nombreux champs d'application tels que l'analyse de séquences génomiques (Durbin et al., 1998) ; voir Ephraim et Merhav (2002) pour d'autres exemples d'applications.

Proposées au début des années 80 (Ferguson, 1980), les semi-chaînes de Markov cachées généralisent les chaînes de Markov cachées en s'affranchissant de l'hypothèse de lois géométriques d'occupation des états (ou de temps de séjour dans les états). Cette hypothèse est en effet peu réaliste notamment si les états sont supposés représenter des zones homogènes, les temps de séjour représentant alors les longueurs de zones. Bien que les semi-chaînes de Markov cachées permettent une modélisation plus réaliste d'un grand nombre de structures, ces modèles ont principalement été utilisés pour la détection de gènes (Burge et Karlin, 1997; Lukashin et Borodovsky, 1998), l'analyse de structures de ramification (Guédon et al., 2001) ou de phases de croissance d'arbres forestiers (Guédon et al., 2007).

Les états cachés n'ont pas *a priori* d'existence physique dans le phénomène observé mais peuvent avoir deux rôles distincts : soit d'états instrumentals soit d'états modélisant des zones ayant un sens par rapport à l'application. Aussi, après analyse des données observées en regard du modèle, elles trouvent souvent une interprétation concrète *a posteriori* : zone codante en analyse de génome, phonème en reconnaissance de la parole ou phase de croissance en analyse de la croissance d'arbres. Enfin, une interprétation fine des modèles de Markov cachés nécessite le plus souvent la restauration des séquences d'états cachés. L'analyse des variables observées à l'aide des paramètres estimés du modèle et de la restauration des états cachés fournit souvent une interprétation concrète du processus caché.

On s'intéresse également au rôle joué par les conditions environnementales sur la croissance d'arbres et à la part d'hétérogénéité inter-individuelle au sein d'un peuplement. L'analyse de tels phénomènes s'appuie sur des modèles de régression. Introduits au début du $19^{ème}$ siècle par Legendre et Gauss, les modèles linéaires permettent d'analyser et de caractériser l'influence de covariables sur les données observées. Ces modèles, qui reposent sur l'utilisation de la loi gaussienne, se sont imposés dans de nombreuses situations et sont devenus des outils usuels de modélisation statistique.

Cependant, de très nombreux phénomènes observés sont difficilement modélisables par la loi gaussienne. Citons, par exemple, la cas de relevés de durée de vie de composants, de l'observation du nombre d'individus dans une population ayant une certaine caractéristique, ou encore de la caractérisation dichotomique d'un phénomène. Aussi, afin de permettre une analyse de ces données non gaussiennes, une extension en termes de loi des modèles linéaires classiques a conduit, au début des années 70, au développement de la classe plus large de modèles que sont les modèles linéaires généralisés (Nelder et Wedderburn, 1972). Depuis, les modèles linéaires généralisés (GLM, Generalized Linear Model) ont pris une place tellement importante dans de nombreux domaines d'application que de nombreux ouvrages leur sont entièrement consacrés dont celui de McCullagh et Nelder (1989) ou encore celui de Dobson et Barnett (2008).

La modélisation des effets pouvant intervenir dans l'explication du phénomène étudié s'est enrichie depuis les travaux de Legendre et Gauss. La notion d'effet aléatoire a été

introduite en la distinguant de la notion d'effet fixe. En effet, le caractère répété des données mesurées pour un même individu nécessite d'introduire dans la modélisation deux niveaux de lecture du comportement des individus : un niveau global traduit par les effets fixes et un niveau individuel traduit par les effets aléatoires. Ceci vient principalement du fait que l'ensemble des observations relatives à un individu est généralement corrélé. L'introduction d'effets aléatoires permet de séparer les différentes sources de variation : celles dues aux effets aléatoires et celle due aux erreurs (Searle et al., 1992). La combinaison des modèles linéaires et des modèles à effets aléatoires a donné naissance en 1959 aux modèles linéaires mixtes (LMM, Linear Mixed Model) pour l'analyse de données de génétique animale (Henderson et al., 1959). Actuellement, le modèle linéaire mixte a des applications dans de nombreux domaines, notamment en économie, en biologie, en agronomie et en médecine ; voir Verbeke et Molenberghs (2000) et McCulloch et al. (2008) pour des exemples d'application.

Afin d'identifier et de caractériser les composantes de la croissance d'arbres, il est nécessaire de combiner la famille des modèles markoviens cachés et la famille des modèles de regression. Les combinaisons markoviennes de modèles linéaires (MS-LM, Markov Switching Linear Model) ont été introduites en 1978 par Lindgren. Ces modèles étendent la famille des chaînes de Markov cachées gaussiennes en incorporant l'influence de covariables comme effets fixes dans le processus d'observation. Ces combinaisons peuvent être vues comme des mélanges finis de modèles linéaires avec des dépendances markoviennes. Les combinaisons markoviennes de modèles linéaires ont depuis été utilisées dans de nombreux domaines telles qu'en économie (Frühwirth-Schnatter, 2006) ou pour l'analyse de réseaux de gènes en biologie (Gupta et al., 2007). Il y a seulement dix ans que Turner et al. (1998) ont introduit les combinaisons markoviennes de modèles linéaires généralisés (MS-GLM, Markov Switching Generalized Linear Model) qui reposent sur la généralisation en termes de lois du processus d'observation des combinaisons markoviennes de modèles linéaires. Ces modèles ont été principalement utilisés en écophysiologie (Turner et al., 1998) et en médecine (Wang et Puterman, 1999). Bien que les modèles à effets aléatoires soient largement connus et exploités, ce n'est que depuis quelques années que les premiers travaux sur l'introduction d'effets aléatoires dans les processus d'observation de modèles markoviens cachés ont vu le jour. Véra (2004) a introduit les combinaisons markoviennes de modèles linéaires mixtes (MS-LMM, Markov Switching Linear Mixed Model) pour l'analyse des composantes de la croissance des arbres. Ces modèles combinent les modèles mixtes de manière markovienne et étendent les MS-LM en incorporant des effets aléatoires "individuels" dans le processus d'observation. Altman (2007) a généralisé en termes de lois les processus d'observation des MS-LMM en introduisant les combinaisons markoviennes de modèles linéaires généralisés mixtes (MS-GLMM, Markov Switching Generalized Linear Mixed Model). Ces modèles ont été jusqu'à maintenant uniquement utilisés en médecine (Altman, 2007; Rijmen et al., 2008). Il est important de noter qu'il existe dans la littérature de nombreux modèles de type Markov caché qui combinent de manière mar-

kovienne des modèles tels que les combinaisons markoviennes de modèles autorégressifs pour lesquels les observations sont modélisées dans chaque état par un processus autorégressif (Ephraim et Merhav, 2002). D'autres exemples de combinaisons markoviennes sont données dans l'ouvrage de Frühwirth-Schnatter (2006).

Contrairement aux approches basées sur les semi-chaînes de Markov cachées proposées par Guédon et al. (2007) pour l'étude des composantes de la croissance d'arbres, les longueurs des phases de croissance ne sont pas modélisées explicitement par les combinaisons markoviennes de modèles de regression. Les combinaisons semi-markoviennes de modèles de régression, où la semi-chaîne de Markov sous-jacente représente à la fois la succession de phases de croissance et leurs longueurs, permettent de pallier ce problème. Les semi-chaînes de Markov cachées et les combinaisons markoviennes de modèles de regression étant relativement peu exploitées, hormis ces dernières années, le nombre de travaux sur les combinaisons semi-markoviennes de modèles de regression est très limité. Russell (1993) a introduit les combinaisons semi-markoviennes de modèles linéaires (SMS-LM, Semi-Markov Switching Linear Model) qui étend la famille des combinaisons markoviennes de modèles linéaires au cas semi-markovien. Il a appliqué ces modèles à la reconnaissance de la parole. Kim et Smyth (2006) ont étendu les MS-LMM au cas semi-markovien en introduisant les combinaisons semi-markoviennes de modèles linéaires mixtes (SMS-LMM, Semi-Markov Switching Linear Mixed Model) de type "gauche-droite", c'est-à-dire constitué d'une succession d'états transitoires suivie par un état final absorbant, avec un effet aléatoire associé à chaque état et modélisant l'hétérogénéité inter-individuelle. La structure "gauche-droite" entraîne un ordre sur les états et chaque état ne peut être visité au maximum qu'une fois. Ils ont utilisé ces modèles en traitement du signal pour analyser les formes d'ondes.

Ce travail de thèse s'est donc intéressé à des modèles statistiques émergeants : les combinaisons markoviennes et semi-markoviennes de modèles de régression.

Inférence dans les combinaisons markoviennes et semi-markoviennes

Bien que les combinaisons markoviennes et semi-markoviennes de modèles de régression soient très intéressantes en pratique, la présence de variables cachées complique l'inférence statistique. L'estimation des paramètres basée sur la vraisemblance est rendue difficile par l'absence de formule explicite pour le maximum de vraisemblance, et requiert en général des algorithmes itératifs comme l'algorithme EM (McLachlan et Krishnan, 2008).

Lindgren (1978) et Cosslett et Lee (1985) ont transposé l'algorithme EM pour chaînes de Markov cachées classiques aux combinaisons markoviennes de modèles linéaires. L'étape E est implémentée par l'algorithme "avant-arrière" propre aux chaînes de Markov cachées classiques. L'étape M repose sur la maximisation directe de l'espérance de la log-vraisemblance des données complètes sachant les données observées. Il existe cependant des méthodes alternatives à l'algorithme EM. Chopin et Pelgrin (2004) ont proposé d'uti-

liser une approche bayésienne pour estimer les paramètres des combinaisons markoviennes de modèles linéaires.

L'hypothèse de linéarité des processus d'observation dans chaque état n'étant pas respectée dans le cas des MS-GLM, la maximisation n'est pas aussi simple. Turner et al. (1998) ont proposé une méthode basée sur l'algorithme EM sous la condition d'équilibre de la chaîne de Markov sous-jacente avec une étape de maximisation pour les paramètres des GLM qui repose sur l'algorithme des moindres carrés pondérés itératifs. Leur approche peut s'avérer vite lourde et coûteuse si le nombre d'états est élevé, si l'influence des covariables est supposée différente d'un état à l'autre ou si la chaîne de Markov sous-jacente n'est pas en équilibre. Une alternative pour estimer les MS-GLM a été proposée par Wang et Puterman (1999) où la maximisation à l'étape M se fait à l'aide de méthodes de quasi-Newton (Lange, 2004). Cependant, les travaux de Wang et Puterman (1999) reposent sur l'hypothèse d'une unique séquence observée (i.e. d'un seul individu). Dans le cas de plusieurs individus observés, Wang et Puterman (1999) les traitent indépendamment avec un processus caché propre à chaque individu ; c'est-à-dire que les probabilités initiales et les probabilités de transition sont différentes d'un individu à un autre.

L'estimation des combinaisons (semi-)markoviennes de modèles linéaires mixtes est un problème plus difficile dans la mesure où cela nécessite de prendre en compte, non plus une unique structure cachée mais deux types de structures cachées : les états du processus markovien sous-jacent d'une part, et les effets aléatoires des modèles linéaires mixtes d'autre part. Véra (2004) a étudié un algorithme itératif de type "restauration-maximisation" pour estimer les paramètres des MS-LMM avec des effets aléatoires "individuels" pouvant être associés à chaque état. L'étape de restauration des séquences d'états sachant les effets aléatoires repose sur une restauration déterministe par l'algorithme de Viterbi (Forney, 1973). L'étape de prédiction des effets aléatoires sachant les séquences d'états repose pour chaque individu sur la séquence d'états la plus probable restaurée de manière déterministe. La restauration déterministe n'est adaptée que dans le cadre de chaîne de Markov sous-jacente de type "gauche-droite" où la séquence d'états la plus probable concentre une bonne partie de la vraisemblance de toutes les séquences d'états possibles. Kim et Smyth (2006) ont proposé une méthode pour estimer les paramètres des SMS-LMM de type "gauche-droite" avec un effet aléatoire modélisant l'hétérogénéité inter-individuelle dans chaque état. La méthode proposée, qui est en fait une application de l'algorithme EM avec une étape E basée sur l'algorithme "avant-arrière", repose fortement sur les deux hypothèses spécifiques au modèle (état visité au maximum une fois et un effet aléatoire différent pour chaque état).

Altman (2007) a proposé une méthode déterministe et une méthode stochastique pour estimer les paramètres des MS-GLMM. L'approche déterministe repose sur une combinaison de méthodes d'intégration numérique de type quadrature de Gauss et de méthodes de quasi-Newton sous l'hypothèse que la vraisemblance d'une chaîne de Markov cachée peut s'écrire sous la forme d'un produit de matrices. L'approche stochastique est fondée

sur l'algorithme MCEM (Monte Carlo EM) où l'étape de maximisation se fait à l'aide de méthodes de quasi-Newton. Altman a souligné quelques limites aux deux méthodes proposées : une sensibilité aux valeurs initiales, une convergence lente et un fort coût en calculs. Comme les relations d'indépendance conditionnelle d'une combinaison markovienne de modèles linéaires généralisés mixtes peuvent être représentées sous forme de graphe orienté acyclique, Rijmen et al. (2008) ont proposé de réaliser l'étape E de l'algorithme EM par l'algorithme d'arbre de jonction (Smyth et al., 1997; Cowell et al., 1999). L'étape de maximisation nécessite l'utilisation d'algorithmes des scores de Fisher et de méthodes d'intégration numérique de type quadrature de Gauss. L'algorithme d'arbre de jonction a quelques inconvénients. Les calculs ne sont pas effectués à partir de probabilités mais à partir de potentiels de cliques, ce qui a l'inconvénient de ne pas prendre en compte les paramètres naturels du modèle (à savoir les probabilités de transition dans le cas d'une structure orienté acyclique). Du coup, il est difficile d'interpréter les calculs et les quantités intermédiaires de l'algorithme de manière probabiliste. Les approches de Altman (2007) et Rijmen et al. (2008) reposent sur l'hypothèse que les effets aléatoires "individuels" sont indépendants des états. De plus, les méthodes proposées par Altman (2007) et Rijmen et al. (2008) ne peuvent pas se transposer au cas semi-markovien.

Objectifs

Le premier objectif de ce travail est de définir des classes de combinaisons markoviennes et semi-markoviennes de modèles de régression permettant la modélisation de données de type séquence ou série chronologique présentant les caractéristiques suivantes :

- les données observées sont structurées en phases successives, asynchrones entre individus,
- les données observées sont influencées par des covariables pouvant varier dans le temps et communes aux individus,
- les données observées présentent une hétérogénéité inter-individuelle pouvant être modulée sur les différents états.

Cette problématique statistique est issue d'une problématique biologique : identifier et caractériser les trois principales composantes de la croissance d'arbres forestiers que sont la composante ontogénique, la composante environnementale et la composante individuelle, sous les hypothèses biologiques introduites précédemment.

Le deuxième objectif de ce travail est de développer des algorithmes d'inférence sous la contrainte qu'ils ne soient pas contraints par le type de structure sous-jacente (ergodique, "gauche-droite" ...), qu'ils se transposent facilement d'un processus markovien sous-jacent à un processus semi-markovien et qu'ils soient stables numériquement.

Le troisième objectif de ce travail est de montrer l'intérêt, d'un point de vue applicatif, des modèles statistiques et des algorithmes d'inférence proposés. En particulier, nous montrons que la modélisation proposée permet de confirmer les hypothèses biologiques

introduites précédemment et nous examinons le rôle joué par chaque composante sur la croissance d'arbres de différentes espèces dans différentes conditions de croissance.

Plan du mémoire

La motivation biologique de ce travail de thèse est présentée en détail dans le chapitre 1. Nous revenons sur les notions botaniques nécessaires à une bonne compréhension du contexte applicatif et sur la problématique biologique. Les jeux de données sur lesquels s'appuieront nos travaux sont présentés. L'analyse des longueurs de pousses annuelles de pins Laricio par une semi-chaîne de Markov cachée gaussienne permet de mettre en évidence les hypothèses biologiques décrites précédemment.

Dans le chapitre 2, nous proposons une revue de la littérature traitant des fondements statistiques de ce travail. Dans un premier temps, nous décrivons l'algorithme EM et ses nombreuses variantes et extensions. Nous présentons ensuite les chaînes et semi-chaînes de Markov cachées permettant de prendre en compte la structure en zones homogènes des données observées. Nous terminons par la présentation des modèles linéaires généralisés et des modèles linéaires mixtes permettant de modéliser les effets intervenant sur le phénomène observé. Les propriétés asymptotiques de chaque modèle concluent chaque partie.

Le chapitre 3 est consacré aux combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés pour le cas de données appartenant à la famille exponentielle. Après avoir défini ces modèles, nous traitons les problèmes d'inférence statistique. Nous proposons des techniques d'estimation par maximum de vraisemblance qui s'appuient sur l'algorithme du gradient EM. L'étape E consiste soit en une restauration probabiliste de toutes les séquences d'états possibles par l'algorithme "avant-arrière", soit en une restauration par simulation de séquences d'états par l'algorithme "avant-arrière" de simulation. L'étape M pour les paramètres du processus markovien sous-jacent s'écrit sans difficulté quel que soit le type de restauration. Les paramètres des modèles linéaires généralisés sont obtenus par l'algorithme des scores de Fisher où les matrices hessiennes et les gradients sont pondérés soit par les comptages extraits des séquences d'états simulées, soit par les pseudo-comptages calculés par l'algorithme "avant-arrière".
Pour terminer ce chapitre, nous traitons deux cas particuliers des combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés : les données de comptage et les données binaires. Nous étudions la robustesse des techniques d'estimation proposées sur simulations et sur données réelles.

Les combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes sont présentées dans le chapitre 4. Nous proposons une revue de littérature sur les techniques d'estimation existantes. Nous distinguons deux types d'effet aléatoire dans le modèle linéaire mixte associé à chaque état : soit un effet aléatoire "individuel" soit un effet aléatoire "temporel". Quel que soit le type d'effet aléatoire, nous sommes en présence de deux types de structures cachées : les états du processus markovien sous-jacent d'une part et les effets

aléatoires des modèles linéaires mixtes d'autre part. De plus, les observations successives pour un individu ne sont plus indépendantes sachant les séquences d'états mais sachant les séquences d'états et les effets aléatoires. C'est pourquoi, nous proposons comme alternative à l'algorithme EM des algorithmes de type MCEM (Monte Carlo EM) où les quantités calculées dans l'étape E classique sont approchées en utilisant des méthodes de Monte Carlo. L'étape E des algorithmes MCEM proposés se décompose en deux étapes de restauration conditionnelle, une pour les effets aléatoires sachant les séquences d'états (et les données observées) et une pour les séquences d'états sachant les effets aléatoires (et les données observées). Les estimations des paramètres sont obtenues à l'étape M par maximisation de l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées. L'étape de restauration des séquences d'états sachant les effets aléatoires peut se faire soit de manière probabiliste, soit par simulation. L'étape de restauration des effets aléatoires sachant les séquences d'états peut se faire soit par prédiction soit par simulation.

Nous discutons pour chaque type d'effet aléatoire des combinaisons possibles de ces étapes de restauration conditionnelle à l'étape E. Chaque partie est complétée par une étude des propriétés des algorithmes MCEM proposés : initialisation, convergence, ordre des étapes et choix du nombre d'échantillons à simuler à chaque itération. Nous étudions le comportement des algorithmes d'estimation proposées et les comparons sur des données simulées et réelles.

Le chapitre 5 est dédié à l'application des combinaisons semi-markoviennes de modèles linéaires mixtes sur les données de longueur de pousses annuelles des jeux de données présentés au chapitre 1. D'une part, nous analysons conjointement les trois composantes de la croissance d'arbres forestiers en modélisant dans chaque phase de croissance, l'influence du climat par des effets fixes et l'hétérogénéité inter-individuelle par un effet aléatoire. D'autre part, nous analysons conjointement la composante ontogénique et la composante environnementale en modélisant dans chaque phase de croissance, l'environnement commun à tous les arbres par un effet aléatoire. Nous présentons pour chaque analyse les étapes dans l'ordre suivant : élaboration du modèle, caractérisation de la technique d'estimation, étude des propriétés de la population, étude du comportement individuel et étude des effets aléatoires. L'intérêt de ces applications est de vérifier si la modélisation proposée permet de confirmer les hypothèses biologiques introduites précédemment et de quantifier le rôle joué par chaque composante sur la croissance d'arbres.

Dans le chapitre 6, nous discutons des méthodes proposées d'un point de vue statistique, informatique et biologique en revenant à la fois sur leurs points forts et sur leurs limites. Nous soulignerons les développements futurs pouvant être envisagés.

Chapitre 1

Présentation de la problématique biologique et des données

Ce chapitre se propose de présenter la problématique biologique à laquelle nous nous sommes intéressés. Nous introduisons dans un premier temps les termes botaniques les plus utiles à la compréhension du contexte applicatif. Les définitions présentées visent essentiellement à aider à une compréhension des mécanismes qui affectent la croissance de l'axe principal d'une plante, c'est-à-dire la tige (ou le tronc dans le cas d'un arbre). Nous abordons également les notions d'âge d'une plante et parlons des protocoles de mesure et d'échantillonnage. Dans un second temps, nous explicitons la problématique biologique. Dans un troisième temps, nous présentons les jeux de données botaniques et les données climatiques associées. Enfin, pour terminer, nous analysons de manière exploratoire les données du pin Laricio.

1.1 Notions botaniques

La plante a depuis longtemps été à la fois très utilisée (aliment, matériau, médicament ...) et en grande partie méconnue. Aujourd'hui, les connaissances acquises sont importantes et ne cessent de s'améliorer. Une bonne part des espèces existantes est maintenant connue. Cependant, la construction des végétaux et notamment des arbres n'a été abordée que récemment par la botanique avec les travaux de Hallé et Oldeman (1970). Cette démarche tardive est liée aux dimensions de l'arbre, à sa longévité et au coût pour la prise de mesure. Auparavant, les scientifiques se limitaient à la description de parties d'arbres tels que les bourgeons, les feuilles ou les rameaux (approche essentiellement morphologique).

Les plantes sont des êtres vivants pluricellulaires[1] à la base de la chaîne alimentaire. Elles se caractérisent par une structure organisée à la fois dans l'espace et dans le temps.

[1] Organisme vivant comportant plusieurs cellules et ayant des cellules différenciées assurant des fonctions spécifiques

Cette structure peut se décomposer en un certain nombre d'entités botaniques qui correspondent à différents niveaux d'organisation emboîtés les uns dans les autres. Ces entités élémentaires se répètent dans le temps au cours de trois processus fondamentaux : la croissance, la ramification et la mortalité. Elles caractérisent le développement de la plante.

Bien que la croissance puisse être envisagée de diverses manières (croissance d'une branche, croissance d'une plante ou croissance des racines), nous ne considérerons dans la suite de ces travaux de thèse que la croissance en longueur de la tige principale, appelée **croissance primaire** en botanique. La tige (ou axe feuillé) principale est l'organe qui croît à l'opposé de la racine et porte des feuilles et des bourgeons ; ces derniers pouvant évoluer en rameaux ou en fleurs. La tige, appelée également axe principal ou tronc chez les arbres, correspond à une succession d'organes et de tissus engendrés par un seul et même méristème[2] ; une revue critique sur l'architecture des plantes est exposée par Barthélémy et Caraglio (2007). Cet axe assure les fonctions de support des organes qu'il porte et de transport des substances d'un secteur à l'autre de l'individu. La croissance d'une tige est le résultat de deux mécanismes (Champagnat et al., 1986) : l'organogénèse et l'allongement (également appelée élongation dans la littérature botanique).

Définition 1.1 *L'**organogénèse** se déroule au niveau de l'apex (ou sommet) d'une tige. C'est au niveau du méristème terminal (Nougarède et Rembur, 1985) que se produisent des phases d'intense activité mitotique[3] au cours desquelles sont initiés de nouveaux éléments de la tige ou métamères.*

Définition 1.2 *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 situé au sommet de la tige.*

En fonction de la durée de son déroulement et de ses modalités d'expression dans le temps, la croissance en longueur d'une tige peut être qualifiée de continue ou de rythmique, de monocyclique ou de polycyclique. D'autres caractères permettent de qualifier cette croissance primaire ; voir dans Barthélémy et Caraglio (2007).

Croissance continue ou rythmique

Hallé et al. (1978) distinguent :

- des axes qui ne présentent pas de périodicité d'allongement marquée et qui sont dits à croissance continue,
- des axes qui montrent une périodicité d'allongement marquée et qui sont dits à croissance rythmique.

[2]Le méristème est l'ensemble des cellules situées à l'extrémité d'un axe qui par leur aptitude à se diviser gênèrent les différentes parties du végétal (feuille, tige, racine, organe de reproduction).

[3]Qui se rapporte à la mitose, mode de division de la cellule vivante, au cours duquel le noyau se dédouble avant le corps cellulaire.

En des termes plus simples, la croissance d'une plante est qualifiée de **croissance continue** si les deux aspects de la croissance, l'organogénèse et l'allongement sont ininterrompus tout au long de la croissance (figure 1.1.a). C'est notamment le cas du Cocotier (*Cocos nucifera* L., Arecaceae), du palmier à huile (*Elaeis guineensis* Jacq., Arecaceae) et du Palétuvier[4] (*Rhizophora mangle* L., Rhizophoraceae).

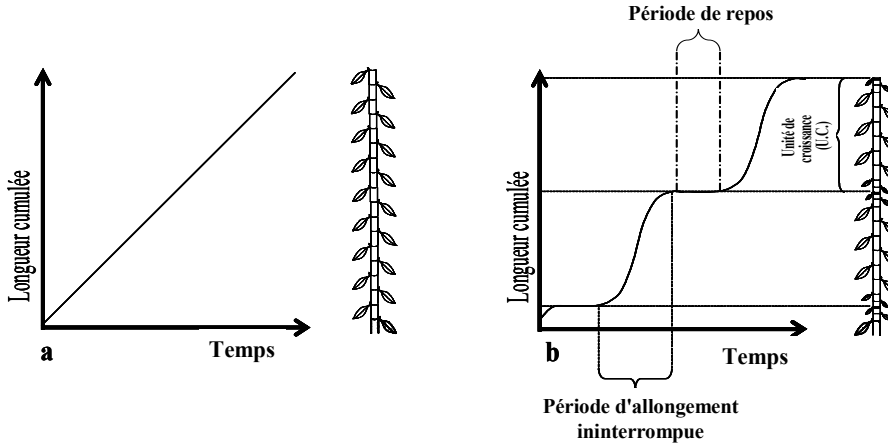

FIG. 1.1 – *(a) Croissance continue ; (b) Croissance rythmique de l'axe feuillé chez l'Hévéa (Heuret, 2002).*

Par opposition, la croissance d'une plante est qualifiée de **croissance rythmique** si l'organogénèse et l'allongement de la tige se réalisent en alternance avec des phases de repos et se manifestent par vagues (figure 1.1.b). La portion de tige mise en place au cours d'une phase d'allongement ininterrompue est appelée **unité de croissance (U.C.)**. Chez l'Hévéa (*Hevea brasiliensis* L., Euphorbiaceae), l'alternance de phases d'allongement et de repos est facilement identifiable sur le terrain ; chaque U.C. correspondant à une portion de tige qui porte une série de cataphylles[5] basales suivie par une série de feuilles (figure 1.1.b).

Croissance monocyclique ou polycyclique

Introduits en 1951 pour qualifier la structure des tiges de diverses espèces du genre *Pinus* L., les termes de monocyclisme et de polycyclisme ont été, par la suite, utilisés pour désigner la structure des tiges de multiples autres végétaux. Au cours d'une même année de végétation, une portion de tige peut résulter d'une unité de croissance (figure 1.2.a) ou de plusieurs unités de croissance (figure 1.2.b). La croissance de la tige sera alors respectivement qualifiée de monocyclique ou polycyclique.

Marqueurs morphologiques du mode de croissance

Des marqueurs morphologiques, qui traduisent le fonctionnement passé des méristèmes, permettent au botaniste de reconstituer la vie d'une plante en repérant *a posteriori* les

[4]Plante à fleurs tropicale, vivant dans la mangrove.

[5]Feuilles réduites ou écailles, précédant les feuilles assimilatrices.

arrêts de croissance. Une tige est constituée d'une succession d'entre-noeuds séparés par des noeuds ; les noeuds étant les lieux d'insertion des feuilles (figure 1.3.a). Les marqueurs morphologiques visuels d'arrêt de croissance peuvent être caractérisés par la diminution de la longueur des entre-noeuds autour de la zone d'arrêt ou par des cicatrices laissées par les cataphylles (figure 1.3.b), petites écailles protégeant les bourgeons.

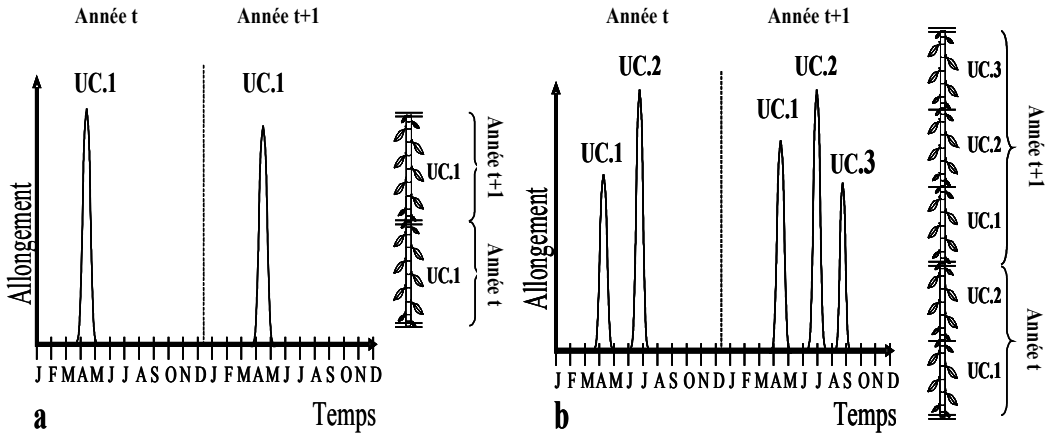

FIG. 1.2 – *(a) Croissance annuelle monocyclique ; (b) Croissance annuelle polycyclique (Heuret, 2002).*

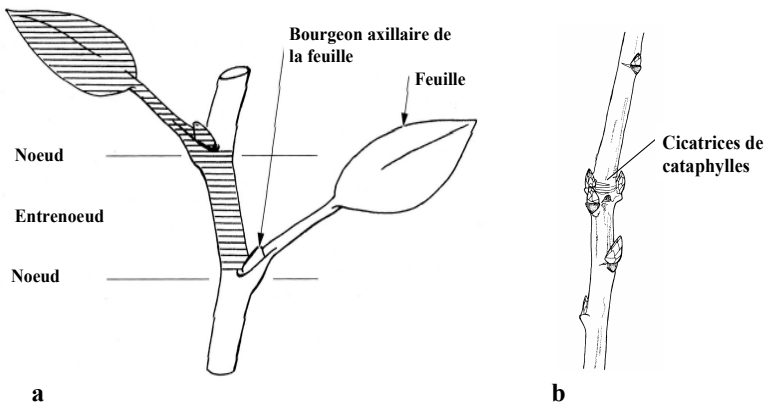

FIG. 1.3 – *(a) Organisation fondamentale de la tige. L'ensemble [entrenoeud + feuille + bourgeon axillaire] (en grisé) est l'élément modulaire de base, seul responsable, par superpositions et modifications, de l'édification de l'appareil aérien des plantes (Raynal-Roques, 1994). (b) Arrêt de croissance inter-annuel délimité par les cicatrices de cataphylles chez le chêne sessile (Heuret, 2002).*

Les cicatrices laissées par les cataphylles permettent à la fois de repérer les changements d'année de végétation et de délimiter les unités de croissance sur les portions d'axe relativement jeunes. Sur des portions de tige plus âgées, les cicatrices laissées par les différents organes foliaires s'estompent à cause notamment de la croissance en diamètre des arbres et du changement de texture de l'écorce. Il est alors parfois difficile de localiser tous les arrêts de croissance et les changements d'année de végétation. Une lecture de ces

arrêts reste possible mais elle est plus coûteuse en temps car il est nécessaire de mesurer la trajectoire et la dimension de la moelle[6]. A partir des définitions introduites précédemment, nous pouvons définir une des notions fondamentales dans la suite de ces travaux de thèse, la notion de pousse annuelle.

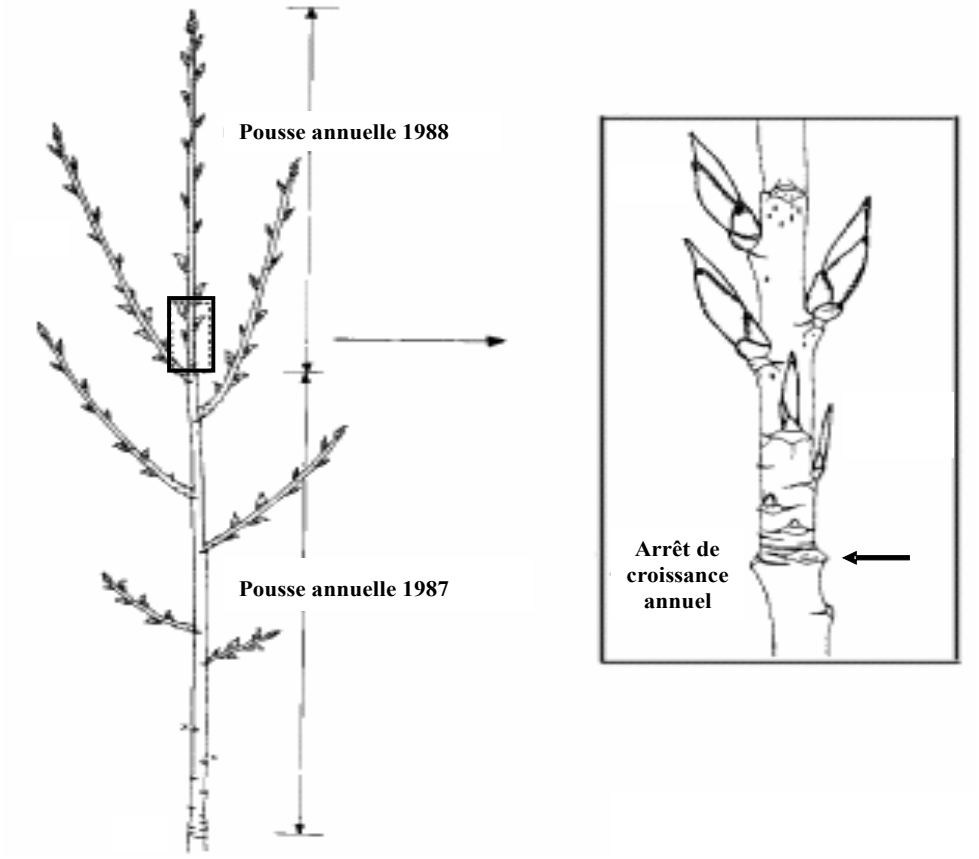

FIG. 1.4 – *Succession de pousses annuelles délimitées par un arrêt de croissance (image fournie par Y. Caraglio).*

Définition 1.3 *La **pousse annuelle**, dénotée P.A., représente la portion de tige mise en place au cours d'une année de végétation de la plante (figure 1.4).*

Nous nous intéresserons principalement par la suite à des caractéristiques telles que la longueur de pousses annuelles en cm, le nombre de branches par pousse annuelle, le nombre d'unités de croissance par pousse annuelle et la présence/absence de ramification latérale[7].

[6]Tissu au centre des tiges produit par le méristème primaire.

[7]La ramification latérale résulte de la différenciation, sur les flancs de l'apex d'une tige, d'un territoire de cellules à caractères embryonnaires. Ce territoire se développe à l'aisselle d'une ébauche foliaire et

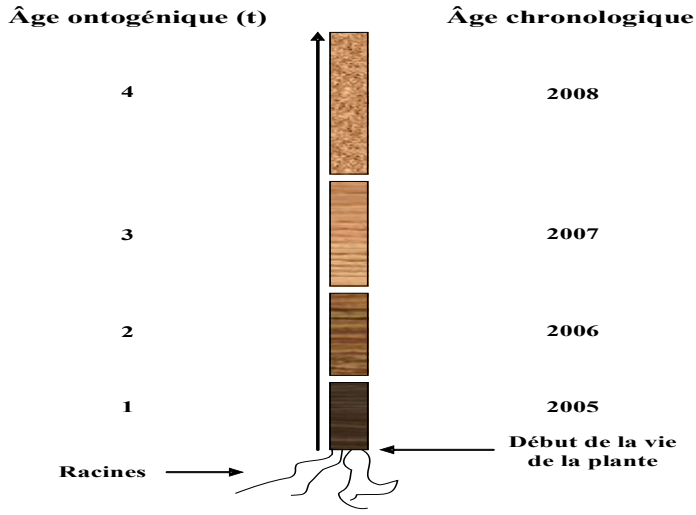

FIG. 1.5 – *Schéma d'une tige constituée d'une succession de quatre pousses annuelles représentées sous forme de rectangle.*

Âge ontogénique et âge chronologique

La vie d'une plante peut être reconstituée *a posteriori* à partir des marqueurs morphologiques annuels. Une plante peut être caractérisée par deux âges différents : l'**âge ontogénique** ou l'**âge chronologique** (ou âge du calendrier). L'âge chronologique d'une pousse annuelle correspond à l'année de formation (et d'allongement) de cette pousse (figure 1.5).

Il existe une relation entre l'âge ontogénique t et l'âge chronologique d : $t(d) = d - h + 1$ où h représente l'année de formation de la première pousse annuelle d'une tige.

Ces différentes notions botaniques étant introduites, nous allons vous présenter le protocole de mesure classique à l'échelle de la plante entière et de la pousse annuelle.

Protocole de mesure

Un ensemble de plantes est échantillonné selon divers critères : afin de couvrir l'ensemble des classes de diamètre, de hauteur, d'âge ou encore parmi les arbres dominants[8] ou co-dominants[9]. Les arbres sont alors abattus et découpés en tronçons pour pouvoir être manipulés plus facilement. La prise de mesures s'effectue du sommet vers la base de

forme un méristème latéral. Par son fonctionnement, ce méristème latéral pourra alors édifier à son tour un nouvel axe feuillé, dont il deviendra le méristème terminal et qui sera qualifié d'axe latéral ou rameau. La ramification latérale correspond au mode de ramification le plus répandu chez les végétaux vasculaires. Chez les Angiospermes, ce processus aboutit à la formation de rameaux, généralement axillaires et qu'il est possible d'identifier, la plupart du temps, par la présence du (ou des deux) premier(s) organe(s) foliaire(s) qui occupe(nt) une position précise : les préfeuilles.

[8] Arbre au-dessus du niveau moyen de la canopée (ensemble des cimes des arbres, constituant le couvert d'une forêt), la couronne reçoit pleinement la lumière.

[9] Arbre généralement au niveau de la canopée, la couronne est entourée en partie au moins par d'autres couronnes et reçoit peu de lumière latérale.

l'arbre pour être sûr de la date de mise en place des différentes unités de croissance. Une coupe transversale est faite à la base de chacune des unités de croissance repérées visuellement, afin de déterminer précisément l'âge de chaque arbre par lecture de son nombre de cernes[10]. Chaque tronçon de bois sépare deux coupes et comporte donc un arrêt de croissance. Ces tronçons sont alors fendus longitudinalement en passant par la moelle. Si, sur un tronçon donné, un arrêt supplémentaire n'ayant pas été identifié visuellement est repéré par l'analyse de la moelle, la longueur des unités de croissance est alors mesurée entre deux arrêts successifs visualisés sur la moelle, et les données sont bien-entendus corrigées. Le caractère monocyclique ou polycyclique d'une pousse annuelle est déduit de son nombre d'UC constitutives. Les longueurs des pousses annuelles sont quant à elle obtenues en faisant la somme des longueurs des UC les constituant. L'âge ontogénique et l'âge chronologique sont relevés pour chaque arbre en sachant que comme il n'est pas toujours possible de remonter entièrement toute la vie de l'arbre, h correspondra à la date de formation de la première pousse annuelle observée de l'arbre. Du fait d'abattre et de tronçonner chaque arbre, ce protocole s'avère vite coûteux en temps et en énergie. Aussi, le nombre d'arbres échantillonnés est en moyenne de l'ordre de 30-40 arbres par campagne de mesure.

1.2 Problématique biologique

La croissance d'une plante (figure 1.6) est principalement le résultat de trois composantes : une composante ontogénique, une composante environnementale et une composante individuelle.

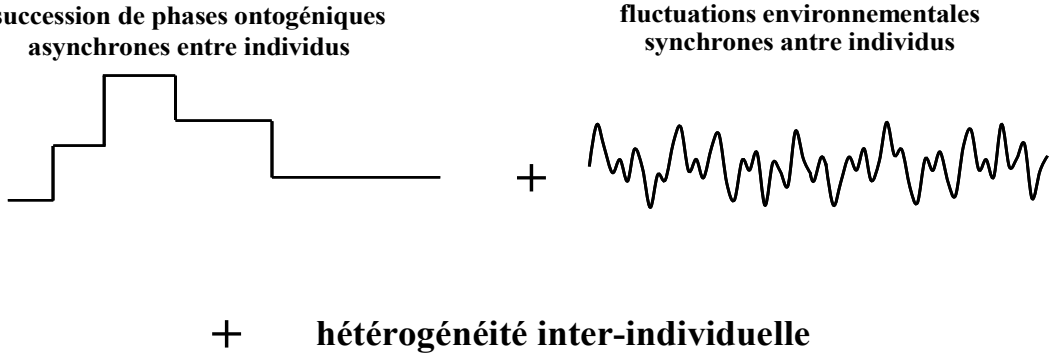

Fig. 1.6 – *Composantes de la croissance d'une plante.*

[10]Cercle concentrique apparent sur la coupe d'un arbre, correspondant à un an d'âge et représentant la rythmicité de l'activité du cambium, tissu mérismatique responsable de la croissance en diamètre des axes (formation du bois entre autres).

Composante ontogénique

Selon Raynal-Roques (1994), l'**ontogénie** (ou l'ontogenèse) correspond à l'ensemble des transformations successives intervenant au cours du développement d'un individu de sa conception à sa mort. Elle comprend le développement embryonnaire, la croissance, la différenciation des organes et la possible apparition de la floraison. Deux points de vue s'opposent quant à la structure de la composante ontogénique de la croissance d'une plante : une structure sous forme de tendance ou une structure sous forme de successions de phases ; voir Guédon et al. (2007) et les références citées dans son article sur les deux points de vue. Guédon et al. (2007) montrent que la composante ontogénique est structurée comme une succession de phases de croissance (figure 1.6) supposées liées à la morphogénèse[11] : une augmentation rapide de la croissance, une croissance maximale stable, puis une diminution lente de la croissance. Selon Guédon et al. (2007), les changements de phase de croissance sont nets et leurs moments varient d'une plante à l'autre. Par exemple, la croissance d'un arbre peut passer en phase de stabilisation à 3 ans tandis que ce changement de phase de croissance pour un autre arbre peut avoir lieu à 5 ans. On parlera alors de changements de phase asynchrones entre individus.

Composante environnementale

La croissance d'une plante est fortement influencée par les conditions environnementales dans lesquelles elle pousse. On peut distinguer deux types de facteurs environnementaux :

- des facteurs qui ne varient pas dans le temps : par exemple, le type de sol, la situation géographique (altitude, au sommet d'une colline, en plaine, pente du terrain...) ;
- des facteurs qui varient dans le temps : par exemple, la température, la pluviométrie, l'ensoleillement...

Nous nous intéresserons principalement à l'effet des variables environnementales variant dans le temps qui ont un effet à court terme sur la croissance des plantes contrairement aux autres variables environnementales.
Dans la suite de ce travail de thèse, nous faisons l'hypothèse biologique que la composante ontogénique, la composante environnementale et la composante individuelle se combinent de manière additive (figure 1.6). Cette hypothèse vient du fait que l'ontogénie a généralement un effet à plus long terme que les variables environnementales variant dans le temps. De plus, Barthélémy et Caraglio (2007) ont montré que la structure et les caractéristiques principales de la composante ontogénique n'étaient pas affectées (hormis dans des conditions extrêmes) par les facteurs environnementaux.

Guédon et al. (2007) font l'hypothèse que la composante environnementale, caractérisée globalement, prend la forme de fluctuations synchrones entre individus. Par exemple, si les conditions climatiques sont favorables une certaine année sur une parcelle alors les arbres de cette parcelle auront tous une croissance plus forte cette même année. Les

[11] Correspond au développement des formes et des structures d'un organisme

botanistes se posent de nombreuses questions sur l'influence des conditions environnementales sur la croissance des plantes : Cette influence est-elle modulée sur chaque phase de croissance ? Peut-on quantifier explicitement cette influence ? Quelles sont les conditions favorables et/ou défavorables ? La réponse de la plante aux variations des conditions de croissance est-elle immédiate ou différée dans le temps ?

Composante individuelle

La composante individuelle correspond principalement à l'environnement "local" de chaque plante. Les différences entre plantes (hétérogénéité inter-individuelle) sont supposées principalement dues au potentiel individuel, à l'origine génétique (Danusevicius, 2001), à des maladies (Lefèvre et al., 1994) ou encore à la compétition entre arbres pour les ressources en lumière, eau et nutriments (Dolezal et al., 2004). Les botanistes s'intéressent à cette composante à deux niveaux bien différents et se posent des questions complémentaires :

- au niveau de la population : quelle est la part de variabilité due à l'hétérogénéité inter-individuelle ? Cette hétérogénéité est-elle modulée sur chaque phase de croissance ?
- au niveau de l'arbre : comment un individu évolue-t-il par rapport à l'arbre "moyen" ?

Il est important de noter que les facteurs influençant différemment chaque individu sont difficilement mesurables rétrospectivement (voir section 1.1, paragraphe protocole de mesure).

L'objectif biologique de ce travail de thèse est de proposer une approche statistique permettant de modéliser conjointement les trois composantes de la croissance basées sur les hypothèses biologiques suivantes :

- la composante ontogénique est supposée structurée en phases de croissance asynchrones entre arbres ;
- la composante environnementale prend la forme de fluctuations synchrones entre arbres et dont l'amplitude moyenne peut être modulée d'une phase de croissance à une autre ;
- la part d'hétérogénéité inter-individuelle peut varier entre les phases de croissance.

1.3 Jeux de données

1.3.1 Pins Laricio de Corse (*Pinus nigra* Arn. ssp. *laricio* Poir., Pinaceae)

Dans le cadre d'une thèse en biologie forestière, Meredieu (1998) a étudié la croissance et la ramification du pin Laricio. Ses travaux portaient sur un jeu de données de pins Laricio échantillonnés dans la forêt domaniale d'Orléans (département du Loiret). Ce jeu de données a été de nouveau en partie analysé par Véra (2004) et en globalité par Guédon et al. (2007). Nous comparerons notamment par la suite les approches et résultats de Guédon et al. (2007) avec les nôtres.

Le jeu de données comprend quatre sous-échantillons de pins Laricio : 31 arbres âgés de 6 ans (première année non mesurée), 29 arbres âgés de 12 ans (première année non mesurée), 30 arbres âgés de 18 ans (première année non mesurée) et 13 arbres âgés de 23 ans (deux premières années non mesurées). Les arbres dans les deux premiers sous-échantillons (6 et 12 ans) ont été sélectionnés en fonction de la répartition de la hauteur des arbres à l'intérieur du peuplement alors que les arbres dans les deux autres sous-échantillons (18 et 23 ans) ont été sélectionnés en fonction de la répartition du diamètre des arbres à hauteur de poitrine à l'intérieur du peuplement. Les arbres du premier sous-échantillon (6 ans) sont restés en pépinière pendant deux ans avant d'être transplantés alors que les arbres des trois autres sous-échantillons sont restés en pépinière pendant trois ans avant d'être transplantés. La densité de plantation était de 1800 tiges/ha pour le premier sous-échantillon (6 ans) et de 2200 tiges/ha pour les trois autres sous-échantillons.

Les données observées ont été collectées rétrospectivement, c'est-à-dire que le développement de chaque arbre a été reconstitué à une date d'observation donnée à partir de marqueurs morphologiques (comme par exemple des cicatrices de ramification) correspondant aux événements passés. Les arbres ont été décrits du sommet vers la base et la longueur (en cm), le nombre de branches par étages et le nombre d'U.C. ont été notés pour chaque pousse annuelle. Ces arbres n'ont été sujets à aucune intervention sylvicole. Un résumé de ces jeux de données est fourni dans le tableau 1.1. L'intervalle d'âge ontogénique correspond à l'âge ontogénique du plus jeune arbre et à l'âge ontogénique du plus vieil arbre.

À partir des notions botaniques introduites dans la section 1.1, nous allons donner certaines caractéristiques du pin Laricio. Le pin Laricio de Corse est une espèce monocyclique (avec de courtes périodes de polycyclisme en début de vie ; voir le tableau 1.1) à croissance rythmique. Chaque année, l'organogénèse a lieu de début juillet à fin octobre et l'allongement, consécutif à l'organogénèse de l'année précédente, a lieu de début mai à fin juillet ; d'après les travaux de Lanner (1976).

Les données climatiques associées aux jeux de données de pins Laricio ont été fournies

	âgés de 6 ans	âgés de 12 ans	âgés de 18 ans	âgés de 23 ans
nombre d'arbres (intervalle d'âge ontog.)	31 (6)	29 (9 → 11)	30 (15 → 17)	13 (20 → 21)
première et dernière année mesurée	1990 → 1995	1985 → 1995	1979 → 1995	1975 → 1995
longueur de P.A. (cm) (plage, moyenne, écart-type)	1 → 72 19.74, 17.65	3 → 64 26.84, 16.62	1 → 83 31.28, 19.08	4 → 85 46.95, 21.16
nombre de branches (plage, moyenne, écart-type)	1 → 9 4.65, 2.06	1 → 9 4.97, 1.73	1 → 11 4.91, 1.91	0 → 10 5.59, 1.80
nombre d'U.C. (plage, moyenne, écart-type)	1 → 2 1.09, 0.28	1 → 2 1.03, 0.16	1 → 2 1.03, 0.16	1 → 2 1.02, 0.12

TAB. 1.1 – *Caractéristiques des sous-échantillons de pins Laricio de Corse.*

par Météo France et sont issues de la station de Chambon-la-forêt, localisée en pleine forêt, non loin de la forêt domaniale d'Orléans. Les températures maximales (en °C), les températures minimales (en °C) et les précipitations cumulées (en mm) ont été recueillies à l'échelle journalière sur une période allant de 1975 à 1996.

1.3.2 Chênes sessiles (*Quercus petraea* Matt. Liebl., Fagaceae)

Dans le cadre d'une étude botanique, Heuret et al. (2000) ont établi un modèle descriptif des différentes composantes de la croissance en hauteur du chêne sessile. Ces travaux sont basés sur un jeu de données de chênes sessiles échantillonnés dans la forêt privée de Louppy-le-Château (département de la Meuse). Comme pour le jeu de données sur les pins Laricio de Corse, ces données ont fait l'objet d'une analyse en partie par Véra (2004) et en globalité par Guédon (2007).

Le jeu de données est constitué de deux sous-échantillons : 46 arbres âgés de 15 ans et 19 arbres de 29 ans (seulement les 24 dernières années ont été mesurées). Ces arbres sont issus de régénération naturelle (glandée[12] de 1983 pour le premier échantillon, glandée de 1969 pour le second). Les individus échantillonnés ont été choisis sans défaut de forme majeur et prélevés parmi les arbres dominants et codominants, définis d'après la hauteur des arbres et le développement important de leur houppier par rapport aux arbres voisins. La densité de plantation était d'environ 2000 tiges/ha. Ces arbres ont subi des interventions sylvicoles : quatre coupes de régénération[13] pour les deux sous-échantillons et une éclaircie[14] au sein du peuplement pour le second échantillon (29 ans). Il convient de noter que ces pratiques sylvicoles favorisent la germination synchrone des arbres les années

[12]Production massive de glands une année donnée.

[13]Opération qui consiste à récolter du bois "mûr" de grande qualité.

[14]Opération qui a pour but de réduire la densité du peuplement. Elle consiste à couper certains arbres afin de laisser aux arbres sélectionnés suffisamment de lumière et de place pour accroître leur diamètre et leur hauteur.

suivant la fructification ; voir Heuret et al. (2000) pour plus de détails. Les troncs sont décrits par pousse annuelle où deux variables quantitatives ont été notées pour chaque pousse annuelle, à savoir la longueur (en cm) et le nombre d'U.C. (tableau 1.2).

	âgés de 15 ans	âgés de 29 ans
nombre d'arbres (intervalle d'âge ontog.)	46 ($10 \rightarrow 15$)	19 ($15 \rightarrow 24$)
première et dernière année mesurée	$1983 \rightarrow 1997$	$1974 \rightarrow 1997$
longueur de P.A. (cm) (plage, moyenne, écart-type)	$1 \rightarrow 128$ $30.72, 25.41$	$7 \rightarrow 148$ $48.98, 24.05$
nombre d'U.C. (plage, moyenne, écart-type)	$1 \rightarrow 4$ $1.74, 0.68$	$1 \rightarrow 4$ $1.91, 0.65$

TAB. 1.2 – *Caractéristiques des sous-échantillons de chênes sessiles.*

Le chêne sessile est caractérisé par une croissance rythmique polycyclique, ce qui signifie que chaque pousse annuelle peut être constituée de plusieurs U.C. successives. Les périodes d'organogénèse, pour toutes les U.C. hormis la première, et les périodes d'allongement couvrent globalement chaque année une période allant de début mars à fin septembre. La période d'organogénèse pour la première U.C. de l'année suivante s'étend d'août à octobre (Champagnat et al., 1986; Fontaine et al., 1999).

Les données climatiques relatives aux chênes sessiles ont été fournies par Météo France et proviennent de la station de Saint-Dizier (département de la Haute-Marne) non loin de Louppy-le-Château. Les températures maximales (en °C), les températures minimales (en °C), la durée d'ensoleillement (en min) et les précipitations cumulées (en mm) ont été recueillies à l'échelle journalière sur une période allant de 1973 à 1997.

1.3.3 Noyers communs (*Juglans regia* L., Juglandaceae)

Dans le cadre de son stage[15], Olivier Taugourdeau a étudié la croissance de noyers communs situés dans un bois de Montferrier-le-Lez (département de l'Hérault).

Le jeu de données est constitué de 138 noyers sélectionnés en fonction d'une absence de traumatisme important sur leur axe principal. Aucune information n'est connue sur l'origine des plants et sur l'occurrence d'interventions sylvicoles. Le bois dans lequel poussent ces noyers n'est pas entretenu et présente une grande diversité d'espèces (pins, chênes, tilleuls...). La morphologie de chaque arbre a été observée, sans les abattre, par lecture *a posteriori* de la croissance annuelle sur les pousses annuelles successives formant le tronc, et ce grâce à la présence d'un anneau de cicatrice au niveau de l'arrêt de croissance hivernal. La longueur (en cm) et la présence/absence de ramification ont été notées pour

[15]stage de master 1 BGAE (Biologie, Géosciences, Agroressources, Environnement) spécialité FENEC (Fonctionnement des Ecosystèmes Naturels Et Cultivés).

chaque pousse annuelle (tableau 1.3). Les noyers communs qui ont ramifié au moins une fois au cours de leur vie ont souvent des pousses annuelles de longueurs supérieures à celles des noyers communs qui n'ont jamais ramifié.

	ayant ramifié	n'ayant pas ramifié
nombre d'arbres (intervalle d'âge ontog.)	22 (5 → 24)	116 (2 → 30)
première et dernière année mesurée	1983 → 2006	1977 → 2006
longueur de P.A. (cm) (plage, moyenne, écart-type)	1 → 35 6.63, 4.84	2 → 49 14.01, 9.31

TAB. 1.3 – *Caractéristiques des sous-échantillons de noyers communs.*

Le noyer commun est une espèce à croissance rythmique. Bien que Sabatier et al. (2003) montrent que la croissance du noyer commun peut être bicyclique (deux U.C. par pousse annuelle), les arbres échantillonnés dans le sous-bois de Montferrier-le-Lez sont monocycliques. L'organogénèse et l'allongement se produisent chaque année entre début avril et fin mai. L'organogénèse de la pousse de l'année suivante a lieu durant l'allongement de la pousse de l'année.

Les données climatiques relatives au noyer commun proviennent d'une station météorologique située au Nord de Montpellier. La température maximale (en °C), la température minimale (en °C), le rayonnement (en J/cm^2) et les précipitations cumulées (en mm) ont été recueillies à l'échelle journalière sur une période allant de 1975 à 2006.

1.3.4 Pins sylvestres (*Pinus sylvestris* L., Pinaceae)

Guédon et al. (2007) ont analysé des pins sylvestres dans une étude de la structure de la composante ontogénique de la croissance. Le jeu de données comprend six sous-échantillons de pins sylvestres plantés dans la forêt d'Hanau, non loin de Bitche (département de la Moselle) : 32 arbres âgés de 3 ans, 10 arbres âgés de 12 ans, 10 arbres âgés de 15 ans, 16 arbres âgés de 30 ans, 6 arbres âgés de 40 ans et 2 arbres âgés de 70 ans. Les arbres des deux premiers sous-échantillons (3 et 12 ans) sont restés en pépinière pendant deux ans avant d'être transplantés alors que les arbres des trois sous-échantillons suivants (de 15 à 40 ans) sont restés en pépinière seulement un an avant d'être transplantés. La densité de plantation était de 4500 tiges/ha pour les deux premiers sous-échantillons tandis que les trois suivants avaient de fortes densités de plantation (7500 à 9000 tiges/ha). Les troncs des arbres ont été décrits par pousse annuelle où deux variables ont été notées pour chaque pousse annuelle, la longueur (en cm) et le nombre de branches par étages (tableau 1.4). Les arbres des trois premiers sous-échantillons (3 à 15 ans) n'ont été sujets à aucune intervention sylvicole. Une éclaircie a été pratiquée en 1993 pour le quatrième

sous-échantillon (30 ans) et en 1984 pour le cinquième sous-échantillon (40 ans). Nous n'avons pas d'informations précises pour le dernier sous-échantillon (70 ans).

	âgés de 3 ans	âgés de 12 ans	âgés de 15 ans	âgés de 30 ans	âgés de 40 ans	âgés de 70 ans
nb d'arbres (interv. d'âge ont.)	32 (3)	10 (7 → 10)	10 (11 → 15)	16 (21 → 27)	6 (33 → 36)	2 (47 → 67)
prem. et dern. année mesurée	1999 → 2001	1992 → 2001	1987 → 2001	1975 → 2001	1966 → 2001	1934 → 2000
long. de P.A. (cm) (plage, moy., et.)	1 → 32 6.08, 5.58	3 → 67 35.05, 17.52	4 → 91 54.8, 20.93	10 → 86 55.4, 14.74	14 → 87 52.15, 13.14	5 → 81 40.05, 18.80
nb de branches (plage, moy., et.)	0 → 16 3.42, 2.00	1 → 18 6.43, 2.63	1 → 12 6.72, 1.83	2 → 13 6.17, 1.72	2 → 12 5.98, 1.50	0 → 8 4.13, 1.73

TAB. 1.4 – *Caractéristiques des sous-échantillons de pins sylvestres.*

Le pin sylvestre est une espèce monocyclique à croissance rythmique dont les périodes d'organogénèse et d'allongement sont identiques à celles du pin Laricio. L'organogénèse a lieu de début juillet à fin octobre et l'allongement, consécutif à l'organogénèse de l'année précédente, a lieu de début mai à fin juillet, d'après Lanner (1976).

Météo France nous a fourni les précipitations cumulées (en mm) journalières de 1970 à 2001. Elles proviennent de la station météorologique de Mouterhouse (département de la Moselle).

1.3.5 Analyse exploratoire des pins Laricio

L'objet de cette analyse exploratoire est de justifier les hypothèses biologiques présentées dans la section 1.2 à partir de données observées de croissance. Cette partie s'appuie sur des modèles statistiques présentés ultérieurement dans la section 2.2. Nous avons estimé une semi-chaîne de Markov cachée gaussienne sur la base des longueurs de pousses annuelles (en cm) des 4 sous-échantillons de pins Laricio (section 1.3.1 et figure 1.7). Cette semi-chaîne de Markov cachée gaussienne supposée de type "gauche-droite" à 3 états est composée de 2 états successifs transitoires suivis par un état final absorbant. On s'est appuyé sur les travaux de Guédon et al. (2007) pour le choix du nombre d'états et du type de structure sous-jacente.

Sur la base des paramètres estimés obtenus, la séquence d'états la plus probable a été restaurée pour chaque séquence (ou arbre) observée à l'aide d'une adaptation de l'algorithme de Viterbi (Guédon, 2003) aux semi-chaînes de Markov cachées gaussiennes. Les segmentations optimales des pins Laricio âgés de 18 ans sont représentées par une fonction en escalier (il y a au plus 3 marches correspondant aux 3 phases de croissances identifiées) sur la figure 1.8. Le niveau de chaque marche correspond à la moyenne empirique estimée

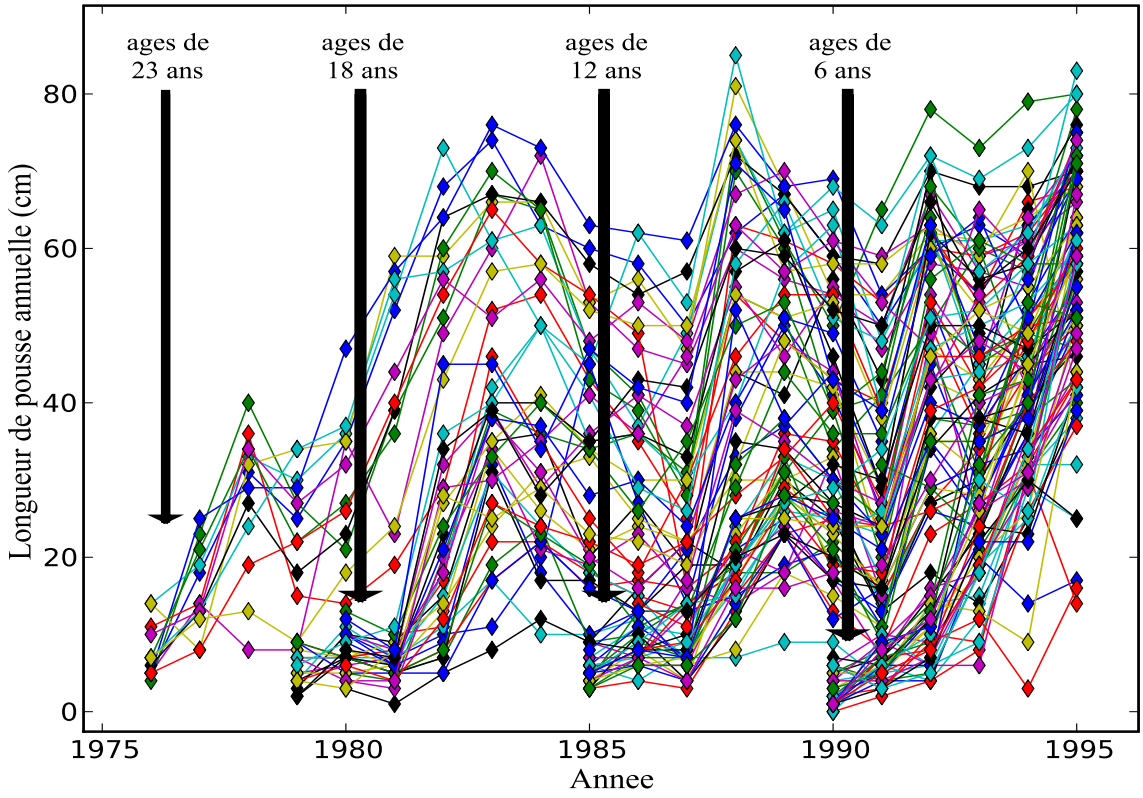

FIG. 1.7 – *Longueur des pousses annuelles successives le long du tronc des 4 sous-échantillons de pins Laricio en fonction des années.*

sur la base du segment extrait de la séquence d'origine. Ces fonctions montrent le degré de synchronisme entre les arbres pour les années de changement de phase de croissance. Mais de manière plus intéressante, ces fonctions soulignent une hétérogénéité inter-individuelle modulée d'une phase de croissance à une autre.

Pour chaque arbre, une séquence résiduelle a été calculée en soustrayant de la séquence originale, la fonction en escalier déduite de la semi-chaîne de Markov cachée gaussienne estimée. Les moyennes (et les intervalles de confiance associés) calculées à partir de ces séquences résiduelles sont significativement supérieures à 0 pour les années 1992 et 1995 et inférieures à 0 pour les années 1986, 1987 et 1991 (figure 1.9.a). Les moyennes (et les intervalles de confiance associés) calculées pour les pins Laricio de 30 ans sont significativement supérieures à 0 pour les années 1988, 1992 et 1995 et inférieures à 0 pour les années 1986, 1987 et 1991 (figure 1.9.b). Nous pouvons faire l'hypothèse que ces résidus moyens traduisent des fluctuations synchrones entre les individus en réponse au climat commun à tous les arbres chaque année.

À partir de cette analyse exploratoire des 4 sous-échantillons de pins Laricios, nous avons pu mettre en avant les principales hypothèses auxquelles devront répondre les modèles statistiques que nous proposerons par la suite :

- la succession de phases de croissance est asynchrone entre les arbres et correspond à des changements marqués de rythme de croissance,
- la croissance des arbres est influencée par des fluctuations annuelles synchrones entre arbres et dont le poids varie entre les phases de croissance,
- la variabilité augmente au cours de la vie des arbres,
- la différence entre arbres est modulée d'une phase de croissance à une autre.

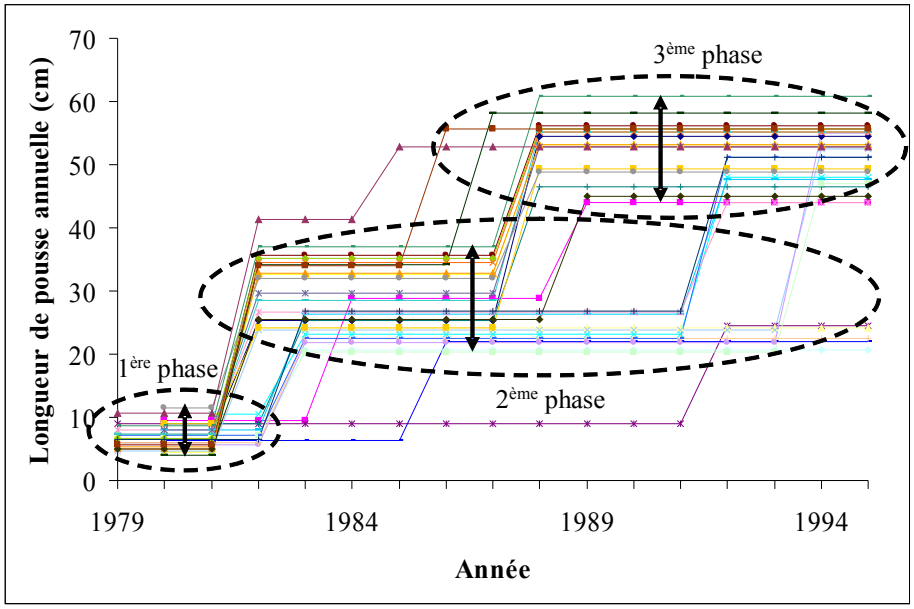

FIG. 1.8 – *Pins Laricio de 18 ans : segmentations optimales calculées à partir de la semi-chaîne de Markov cachée gaussienne vues comme des fonctions en escalier.*

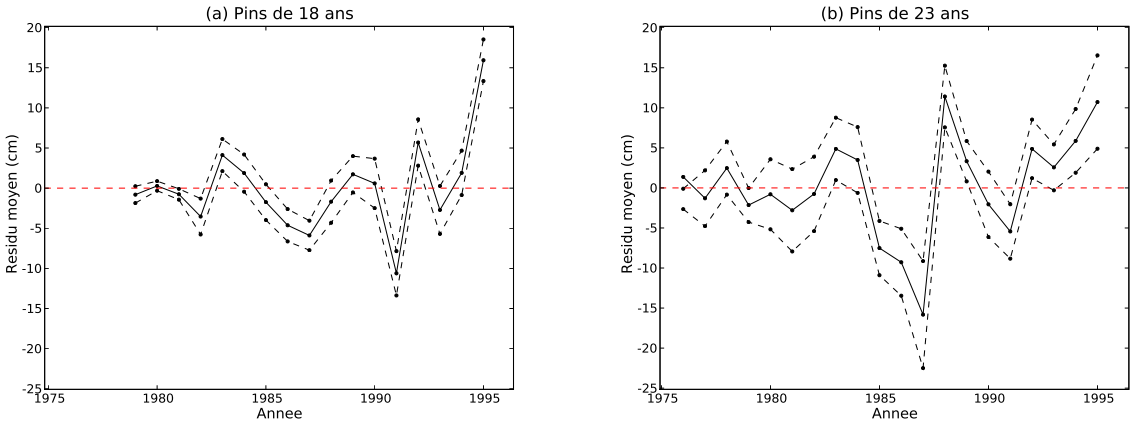

FIG. 1.9 – *Moyennes et intervalles de confiance calculés à partir des séquences résiduelles (a) pins âgés de 18 ans, (b) pins âgés de 23 ans.*

1.4 Discussion

Dans ce travail de thèse, l'objectif est donc de modéliser conjointement les trois principales composantes de la croissance des arbres : ontogénique, environnementale et individuelle en se basant sur les hypothèses biologiques présentées précédemment. La composante ontogénique est structurée comme une succession de phases de croissance asynchrones entre individus tandis que la composante environnementale prend la forme de fluctuations synchrones entre individus. Ce type de données trouvent, dans notre cas, son intérêt dans la modélisation de la croissance d'un arbre : longueur de pousses annuelles, nombre de branches ou nombre d'unités de croissance. Cependant, il existe de nombreux autres domaines dans lesquels nous pouvons rencontrer de telles données :

- en biologie avec les réseaux de gènes (Gupta et al., 2007) ou avec les coliformes fécaux dans l'eau de mer (Turner et al., 1998),
- en médecine avec les tumeurs du cerveau (Rijmen et al., 2008), avec les données d'IRM de patients atteints de sclérose en plaques (Altman et Petkau, 2005) ou avec les crises d'épilepsie (Wang et Puterman, 1999),
- en économie avec les courbes de rendement (Chopin et Pelgrin, 2004),
- en traitement du signal avec les formes d'ondes (Kim et Smyth, 2006).

Chapitre 2

Fondements statistiques

Avant de s'intéresser aux modèles statistiques étudiés au cours de cette thèse, nous allons en donner les fondements statistiques. Ce chapitre débutera par une présentation de l'algorithme EM sur lequel seront basées les méthodes d'estimation des modèles étudiés. Différentes variantes de l'algorithme EM telles que l'algorithme Monte Carlo EM et l'algorithme du gradient EM seront présentées. Les combinaisons markoviennes et semi-markoviennes de modèles de regression reposent sur deux familles de modèles statistiques : les modèles markoviens cachés et les modèles de régression. La seconde partie de ce chapitre sera consacrée à la présentation des chaînes de Markov cachées et des semi-chaînes de Markov cachées. Les propriétés d'indépendance conditionnelle et les méthodes d'estimation accompagneront cette présentation. Une troisième partie traitera des modèles linéaires généralisés. Nous terminerons par une dernière partie sur les modèles linéaires mixtes.

2.1 Algorithme EM

L'algorithme EM (Expectation-Maximization) est une méthode d'estimation qui permet d'obtenir les estimateurs des paramètres dans les problèmes à données incomplètes pour lesquels l'approche classique d'estimation n'est pas toujours envisageable. La notion de données incomplètes couvre de très nombreuses situations : données manquantes, données censurées, effets aléatoires ou variables latentes.

Le nom d'algorithme EM a été donné par Dempster et al. (1977) dans un article célèbre publié dans le journal de la "Royal Statistical Society". Leur travail a été avant tout le résultat d'une synthèse et d'une unification de nombreux travaux antérieurs. Dans cet article, une formulation générale de l'algorithme EM a été présentée, ses principales propriétés ont été établies et de nombreux exemples et applications ont été fournis. Le monde de la statistique appliquée a, depuis 1977, largement contribué par le nombre de ses publications au succès de l'algorithme EM (Meng et van Dyk, 1997). En effet,

l'algorithme EM s'est imposé comme un outil majeur de la statistique et les applications de cette méthode d'estimation ne se comptent plus.

La suite de cette section va largement s'appuyer sur le livre de McLachlan et Krishnan (2008) "The EM algorithm and extensions", le meilleur ouvrage complètement dédié à ce sujet, qui donne à la fois une très bonne introduction au sujet ainsi qu'une ouverture sur les nombreuses extensions actuellement proposées et sur la multitude des applications.

2.1.1 Principe

Supposons que l'on observe une variable aléatoire Y dont la loi dépend d'un paramètre θ (appartenant au domaine Θ) que l'on souhaite estimer par maximum de vraisemblance. La vraisemblance de θ pour les données observées sera notée $L(\theta)$. Par abus de langage, on parlera par la suite de vraisemblance des données incomplètes ou observées. La variable aléatoire observée Y peut être complétée par une variable aléatoire Z correspondant aux données "manquantes" telle que la variable aléatoire $W = (Y, Z)$ corresponde à des données complètes pour lesquelles une solution simple existe pour l'estimation de θ par maximum de vraisemblance.

L'idée de base de l'algorithme EM consiste donc à associer à un problème aux données incomplètes un problème aux données complètes pour lequel une solution simple existe pour l'estimation par maximum de vraisemblance.

Soit $f(y, z; \theta)$ la vraisemblance des données complètes. La vraisemblance $L(\theta)$ des données observées peut s'écrire :

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

Notons que dans la suite de la présentation de l'algorithme EM, nous nous plaçons dans le cadre d'une variable aléatoire Z discrète. La transposition au cas d'une variable aléatoire Z continue est naturelle ; c'est d'ailleurs de cette façon que McLachlan et Krishnan (2008) présentent l'algorithme EM.

L'algorithme EM se décompose à l'itération k en une succession de deux étapes : l'étape E (Expectation) et l'étape M (Maximisation).

Étape E :

L'étape E consiste à concevoir un problème aux données complètes tel que l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées soit manipulable (ce qui suppose d'étudier la relation entre la vraisemblance des données complètes et la vraisemblance des données incomplètes).

Soit $Q(\theta|\theta^{(k)})$ l'espérance de la log-vraisemblance des données complètes à l'itération k conditionnellement aux données observées, en utilisant la valeur courante estimée $\theta^{(k)}$

du paramètre θ. À l'étape E, on calcule

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= \mathrm{E}\left(\log f(Y,Z;\theta)|Y=y;\theta^{(k)}\right) \\ &= \sum_y \mathrm{P}(Z=z|Y=y;\theta^{(k)})\ \log f(y,z;\theta^{(k)}). \end{aligned} \tag{2.1}$$

Du fait de devoir connaître la valeur courante $\theta^{(k)}$ du paramètre θ pour estimer la log-vraisemblance des données complètes, l'algorithme EM est nécessairement un algorithme itératif. L'étape E peut également être vue comme le calcul de la loi conditionnelle $\tilde{\mathrm{P}}^{(k+1)} = \mathrm{P}(Z=z|Y=y;\theta^{(k)})$ pour toutes les valeurs possibles z prises par Z (Neal et Hinton, 1998).

Étape M :

La prochaine valeur du paramètre θ, $\theta^{(k+1)}$ est choisie telle que

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta|\theta^{(k)}), \qquad \forall\theta\in\Theta.$$

Ceci se traduit par le fait de choisir $\theta^{(k+1)}$ dans l'ensemble des valeurs $M(\theta^{(k)})$ qui maximisent la quantité $Q(\theta|\theta^{(k)})$ calculée à l'étape E :

$$M(\theta^{(k)}) = \arg\max_{\theta\in\Theta}\Big\{Q(\theta|\theta^{(k)})\Big\}.$$

Les étapes E et M sont itérées jusqu'à la convergence de l'algorithme. Les propriétés de l'algorithme EM et les critères de convergence sont présentés dans la partie suivante.

2.1.2 Propriétés

L'algorithme EM maximise la vraisemblance des données observées $L(\theta)$ en maximisant itérativement $Q(\theta|\theta^{(k)})$ sur Θ. De ce fait, la vraisemblance des données observées ne peut décroître. Cette propriété d'accroissement monotone de la vraisemblance des données observées démontrée par Dempster et al. (1977) caractérise l'algorithme EM et se traduit par la propriété suivante.

Propriété 2.1 *[Accroissement monotone de la vraisemblance.] À chaque itération de l'algorithme EM, $L(\theta)$ ne peut décroître :*

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)})$$

pour $k = 0, 1, 2, \ldots$.

Propriété 2.2 *[Convergence vers un point stationnaire.] La conséquence principale de cette propriété est la convergence monotone de la séquence de vraisemblances $\{L(\theta^{(k)})\}$ vers une valeur stationnaire L^* telle que*

$$\partial L(\theta)/\partial\theta = 0,$$

ou de manière équivalente

$$\partial \log L(\theta)/\partial\theta = 0.$$

En général, si $L(\theta)$ a plusieurs points stationnaires qui peuvent être soit des maxima locaux, soit un maximum global, soit des points selles. La convergence d'une suite d'itérations EM vers l'une ou l'autre de ces valeurs stationnaires dépend du choix de la valeur initiale du paramètre θ notée $\theta^{(0)}$. Selon le type de point stationnaire (maximum local ou maximum global), on parlera respectivement de convergence locale ou de convergence globale. Dans le cas où la vraisemblance $L(\theta)$ est unimodale, la séquence des paramètres estimés convergera vers l'unique paramètre maximisant la vraisemblance et ceci indépendamment du choix de $\theta^{(0)}$.

Remarques :

1. *Foulley (2002) précise que si notamment la condition de continuité de $\partial Q(\theta|\theta^{(k)})/\partial\theta$ n'est pas respectée, la convergence de $L(\theta^{(k)})$ vers L^* n'implique pas forcément la convergence de $\theta^{(k)}$ vers θ^*.*
2. *McLachlan et Krishnan (2008) discutent de ces propriétés dans leur livre et évoquent le fait que pour les variantes de l'algorithme EM, où l'étape E se fait par simulation (notamment les algorithmes Stochastic EM ou Monte Carlo EM), la propriété d'accroissement monotone de la vraisemblance des données observées n'est pas conservée. Nous en rediscuterons par la suite.*

La vitesse de convergence de l'algorithme EM dépend de la proportion d'information manquante sur θ du fait que l'on observe seulement des réalisations de Y au lieu d'observer conjointement Y et Z. Plus la part d'information manquante est élevée, plus la vitesse de convergence est lente. Cette proportion peut varier pour les différentes composantes de θ et donc expliquer une vitesse de convergence variable pour ces composantes. Si la log-vraisemblance des données observées $\log L(\theta)$ se calcule aisément, celle-ci constitue le moyen le mieux fondé pour évaluer la convergence de l'algorithme EM.

Depuis que Dempster et al. (1977) ont développé la théorie de l'algorithme EM, de nombreuses variantes ont vu le jour. Ces variantes permettent de répondre aux difficultés qui peuvent se rencontrer soit dans le calcul de $Q(\theta|\theta^{(k)})$ à l'étape E, soit dans la maximisation de $Q(\theta|\theta^{(k)})$ à l'étape M, soit dans l'obtention de meilleures performances dont par exemple, l'augmentation de la vitesse de convergence. Nous allons présenter en détail deux variantes qui nous seront utiles par la suite : l'algorithme du gradient EM et l'algorithme de Monte Carlo EM (MCEM). Nous terminerons par une présentation rapide d'autres variantes.

2.1.3 Variantes et extensions de l'algorithme EM

2.1.3.1 Algorithme du gradient EM

L'algorithme du gradient EM est une méthode utilisée lorsqu'il n'y a pas de solution analytique à l'étape M de maximisation ; c'est-à-dire s'il n'est pas possible de maximiser directement $Q(\theta|\theta^{(k)})$ (équation (2.1)). Il fournit alors une alternative itérative au calcul d'une solution explicite. Dans la version décrite par Lange (1995a,b), l'étape M de l'algorithme du gradient EM repose principalement sur des itérations de type Newton-Raphson.

Dans l'algorithme du gradient EM, l'étape M se fait à partir de l'équation itérative suivante à l'itération k (Lange, 2004) :

$$\begin{aligned} \theta^{(k+1)} &= \theta^{(k)} - \left(\frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta\partial\theta'} \right)^{-1} \frac{\partial Q(\theta|\theta^{(k)})}{\partial\theta} \\ &= \theta^{(k)} - \Big(\frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta\partial\theta'} \Big)^{-1} \frac{\partial \log L(\theta)}{\partial\theta} \end{aligned} \tag{2.2}$$

où $\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'$ est la matrice hessienne de $Q(\theta|\theta^{(k)})$, $\partial Q(\theta|\theta^{(k)})/\partial\theta$ est le vecteur gradient de $Q(\theta|\theta^{(k)})$ et $\partial \log L(\theta)/\partial\theta$ est le vecteur gradient de la log-vraisemblance des données observées $L(\theta)$. Par abus de langage, l'algorithme du gradient EM avec une étape M basée sur des itérations de type Newton-Raphson sera appelé algorithme du gradient EM-NR.

Il est nécessaire d'évaluer explicitement la matrice hessienne de la fonction objectif (fonction à maximiser) ; ce qui peut parfois s'avérer difficile (Lange, 2004). En effet, la matrice hessienne $\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'$ peut ne pas être définie négative, et par conséquent, l'algorithme du gradient EM-NR peut ne pas être un algorithme ascendant, ce qui est contraire à la propriété fondamentale de l'algorithme EM. En effet, la définie négativité de la matrice hessienne de $Q(\theta|\theta^{(k)})$ assure un caractère ascendant à l'algorithme du gradient EM-NR. Dans l'algorithme du gradient EM-NR, Titterington (1984) remplace la matrice $\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'$ par la matrice d'information de Fisher dans l'équation (2.2). En effet, dans certaines situations, l'expression de $\mathrm{E}\Big(-\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'\Big)$ prise par rapport à la distribution de Y, données observées, est beaucoup plus simple que celle de $\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'$ et on aura alors recours à un algorithme du gradient EM avec une étape M basée sur des itérations de type scores de Fisher (Titterington, 1984; Foulley, 2002) appelé algorithme du gradient EM-Fisher.

McLachlan (1995) souligne que dans le cas où les données complètes appartiennent à la famille exponentielle, ces deux matrices coïncident et par suite, l'algorithme du gradient EM-NR coincide avec l'algorithme ascendant de Titterington (1984). L'avantage de l'algorithme de Titterington est que cet algorithme est nécessairement un algorithme ascendant, contrairement à l'algorithme du gradient EM-NR. Quand les données complètes

n'appartiennent pas à la famille exponentielle, une reparamétrisation permet fréquemment de rendre cette matrice définie négative (McLachlan et Krishnan, 2008).

Il peut être avantageux numériquement de ne pas aller jusqu'à la convergence en réduisant le nombre d'itérations internes à l'étape M jusqu'à une seule comme l'envisage Lange (1995a). Dans ce cas, il importe toutefois de bien vérifier qu'on augmente la fonction $Q(\theta|\theta^{(k)})$ et qu'on reste ainsi dans le cadre d'un EM dit généralisé (McLachlan et Krishnan, 2008), et par suite de vérifier la définie négativité de la matrice hessienne de $Q(\theta|\theta^{(k)})$. L'algorithme du gradient EM converge quadratiquement tandis que l'algorithme EM converge linéairement. Ces considérations suggèrent qu'une seule itération de l'algorithme du gradient EM à chaque étape M peut être adéquate pour assurer la convergence de cet algorithme approché (Lange, 1995b). C'est cet argument qui est à la base de l'algorithme du gradient EM car Lange a initialement proposé cette variante de l'algorithme EM pour augmenter la vitesse de convergence de l'algorithme EM.

Puisque l'algorithme du gradient EM ressemble étroitement à l'algorithme EM (Lange, 1995a), il tend à partager les propriétés souhaitables de stabilité et de convergence de l'algorithme EM. Sous condition de définie négativité de la matrice hessienne $\partial^2 Q(\theta|\theta^{(k)})/\partial\theta\partial\theta'$, les propriétés d'accroissement monotone de la vraisemblance des données observées (propriété 2.1) et de convergence locale et globale vers un point stationnaire (propriété 2.2) sont respectées.

2.1.3.2 Algorithmes SEM et MCEM

La calcul de $Q(\theta|\theta^{(k)})$ à l'étape E de l'algorithme EM peut s'avérer problématique. Pour contourner cette difficulté, Celeux et Diebolt (1985) ont introduit l'algorithme Stochastic EM (SEM) en vue de l'estimation du maximum de vraisemblance des paramètres d'un mélange de lois exponentielles. Cette variante est une version stochastique de l'algorithme EM. L'intérêt de l'algorithme SEM réside dans l'introduction d'une perturbation aléatoire à chaque itération de l'algorithme EM pour tenter d'éviter une "mauvaise" convergence de l'algorithme vers des points stationnaires stables mais indésirables (points selles, maxima locaux "peu intéressants"). La maximisation de la log-vraisemblance des données complètes $\log f(y, z; \theta)$ se fait à partir d'une évaluation numérique de celle-ci via le calcul de $\log f(y, z^{(k)}; \theta)$ où $z^{(k)}$ est un échantillon de données manquantes simulé à partir de la distribution conditionnelle $Z|Y = y, \theta^{(k)}$ à l'itération k.

Wei et Tanner (1990) ont proposé d'étendre l'algorithme SEM en introduisant l'algorithme Monte Carlo EM (MCEM). Cet algorithme repose sur une approximation de l'espérance des données complètes conditionnellement aux données observées (équation (2.1)) par une approche de Monte Carlo (Robert et Casella, 2004) :

$$\mathrm{E}\left(\log f(Y, Z; \theta)|Y = y; \theta\right) \approx \frac{1}{M}\sum_{m=1}^{M} \log f(y, z_m; \theta) \tag{2.3}$$

où $z_1, \ldots, z_M$ sont i.i.d. et simulées selon la distribution conditionnelle $Z|Y = y, \theta$ et M est le nombre de simulations.

L'algorithme EM peut donc être modifié en remplaçant $Q(\theta|\theta^{(k)})$ par la quantité de l'équation (2.3) pour $\theta = \theta^{(k)}$. Plus formellement, si $\theta^{(k)}$ est la valeur courante du paramètre θ, l'algorithme MCEM consiste en une procédure itérative alternant les deux étapes suivantes à l'itération k :

- **Étape Monte Carlo E :**
  Simulation de M répétitions de z soient $z_1, \ldots, z_M$, selon la distribution conditionnelle $Z|Y = y, \theta^{(k)}$.
- **Étape M :**
  Maximisation par rapport à θ de

$$Q(\theta|\theta^{(k)}) \approx \frac{1}{M} \sum_{m=1}^{M} \log f(y, z_m; \theta).$$

  Ceci se traduit par le fait de choisir $\theta^{(k+1)}$ dans l'ensemble des valeurs $\tilde{M}(\theta^{(k)})$ qui maximisent la quantité approximée à l'étape E :

$$\tilde{M}(\theta^{(k)}) = \arg\max_{\theta \in \Theta} \Big\{ \frac{1}{M} \sum_{m=1}^{M} \log f(y, z_m; \theta) \Big\}.$$

Nous pouvons remarquer que pour $M = 1$ à chaque itération, l'algorithme MCEM est équivalent à l'algorithme SEM. L'algorithme MCEM est la solution naturelle pour résoudre les problèmes de calcul à l'étape E. Cependant, il est nécessaire d'être prudent dans le choix du nombre d'échantillons à simuler et du critère de convergence.

Choix du nombre d'échantillons à simuler M

Wei et Tanner (1990) ont souligné le fait qu'il était inefficace de commencer l'algorithme MCEM avec un très grand nombre d'échantillons à simuler M à l'étape E. Ils recommandent, au lieu de choisir et de fixer M pour tout l'algorithme, que de petites valeurs de M soient utilisées au début de l'algorithme et que ce nombre augmente au fur et à mesure que l'algorithme MCEM se rapproche du maximum de vraisemblance. Des règles déterministes ont été proposées afin de déterminer comment augmenter le nombre d'échantillons à simuler à chaque itération. Par exemple, McCulloch (1994) a augmenté linéairement ce nombre avec le nombre d'itérations. McCulloch (1997) a choisi, dans le cadre des modèles linéaires généralisés mixtes, de simuler 50 échantillons pour les itérations 1 à 19, puis 200 pour les itérations 20 à 39 et 5000 à partir de la $40^{ème}$ itération. Booth et Hobert (1999) et Levine et Casella (2001) ont proposé une règle basée sur l'erreur de Monte Carlo à chaque itération pour augmenter automatiquement le nombre d'échantillons à simuler. Eickhoff et al. (2004) ont défini un critère basé sur des distances par rapport à la vraisemblance des données observées et la taille M est choisie de manière adaptative à chaque itération.

Le nombre d'échantillons à simuler peut également augmenter de manière géométrique au fil des itérations (Caffo et al., 2005).

Convergence de l'algorithme MCEM

McLachlan et Krishnan (2008) ont montré que contrairement à l'algorithme EM, l'algorithme MCEM (et par suite, l'algorithme SEM) ne respecte pas la propriété (2.1) d'accroissement monotone de la vraisemblance des données observées. Caffo et al. (2005) ont proposé une stratégie adaptative pour retrouver la propriété d'accroissement de l'algorithme EM.
Afin de contrôler la convergence de l'algorithme MCEM, Wei et Tanner (1990) ont proposé de surveiller visuellement l'évolution des paramètres au fil des itérations et de s'arrêter quand ces évolutions prennent la forme de fluctuations autour de valeurs stationnaires. Ce critère s'avérant vite fastidieux, McCulloch (1994, 1997) a choisi de stopper l'algorithme MCEM après un nombre prédéterminé d'itérations. Une autre possibilité est de contrôler la convergence en suivant l'évolution de la log-vraisemblance des données observées. Cependant, cette quantité n'étant pas toujours calculable, Shi et Lee (2000) ont utilisé la méthode du "bridge sampling" pour obtenir une approximation du rapport des vraisemblances observées entre deux itérations consécutives et contrôler visuellement la convergence de l'algorithme MCEM. Eickhoff et al. (2004) ont défini un critère basé sur des distances par rapport à la vraisemblance des données observées.

2.1.3.3 Autres variantes et extensions

Il existe diverses autres variantes ou extensions de l'algorithme EM où soit l'étape E, soit l'étape M est modifiée. Une brève description sera donnée dans cette section. On trouvera des explications plus détaillées dans l'ouvrage de McLachlan et Krishnan (2008).

Les algorithmes SEM et MCEM amènent une solution au problème du calcul explicite de l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées (équation (2.1)). Delyon et al. (1999) proposent l'**algorithme Stochastic Approximative EM** (SAEM) pour résoudre les problèmes de convergence et de coût de calcul de l'algorithme MCEM. Cet algorithme requiert la simulation de plusieurs réalisations des données manquantes à chaque itération puis l'actualisation du paramètre courant $\theta^{(k)}$ par l'algorithme SAEM se fait par combinaison des valeurs actualisées par les algorithmes MCEM et EM avec des poids donnés tels que la somme des poids soit égale à 1. Delyon et al. (1999) recommandent de diminuer le poids des valeurs actualisées par l'algorithme MCEM et/ou d'augmenter le nombre de réalisations simulées au fil des itérations. Kuhn et Lavielle (2004) ont également proposé de combiner l'algorithme SAEM avec une procédure MCMC (Monte Carlo Markov Chain, Robert et Casella (2004)) dans le cadre de modèles mixtes non-linéaires.

Dans les cas où le calcul à l'étape E ne pose pas de souci, il peut s'avérer que l'étape

M soit plus délicate. Des variantes de l'algorithme EM sont utilisées lorsque le système à résoudre en le paramètre θ à l'étape M est trop complexe. Par exemple, lors de l'étape M de l'**algorithme Generalized EM** (GEM), on choisit maintenant la valeur du paramètre $\theta^{(k+1)}$ telle que

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta^{(k)}|\theta^{(k)}).$$

Le paramètre $\theta^{(k+1)}$ ne maximise pas $Q(\theta|\theta^{(k)})$, il l'augmente simplement. Dans l'**algorithme Expectation Conditional Maximisation** (ECM), l'objectif étant de simplifier la phase de maximisation, on partitionne le vecteur des paramètres en sous vecteurs puis on maximise successivement par rapport à un sous vecteur, les autres sous vecteurs étant fixés. Dans la version de l'**algorithme Expectation Conditional Maximisation Either** (ECME), une des étapes précédentes de maximisation conditionnelle est réalisée par maximisation directe de la vraisemblance des données observées.

Nous avons évoqué précédemment le problème de vitesse de convergence de l'algorithme EM. L'**algorithme Parameter Expanded EM** (PX-EM) est utilisé pour pallier le fait que la vitesse de convergence de l'algorithme EM peut s'avérer lente dans certains types de problème. Son principe repose sur le concept d'extension paramétrique à un espace plus large des paramètres que l'espace d'origine par adjonction d'un vecteur de paramètres de travail. Dans le cas des modèles mixtes par exemple, il a été proposé des procédures de normalisation des effets aléatoires (Foulley, 2002). L'**algorithme Alternating ECM** (AECM, Meng et van Dyk (1997)) est une extension de l'algorithme ECM, où la spécification des données complètes est différente à chaque étape de Maximisation Conditionnelle (CM), ce qui permet d'accélérer la convergence de l'algorithme. L'algorithme AECM permet aux données complètes de varier dans et entre les itérations et à la stratégie de réduction de modèle d'aller au delà d'une simple partition des paramètres. L'algorithme AECM augmente la vraisemblance des données observées à chaque cycle et à chaque itération. D'autres approches (**algorithme Incremental EM, algorithme Sparse EM, algorithme Sparse Incremental EM**) reposent sur l'algorithme EM vu comme un algorithme de Maximisation-Maximisation (Neal et Hinton, 1998). Elles permettent d'améliorer la vitesse de convergence de l'algorithme EM, l'étape E se faisant maintenant bloc par bloc.

La présentation de l'**algorithme EM supplémenté** introduit par Meng et Rubin (1991) termine la revue non exhaustive des nombreuses variantes ou extensions de l'algorithme EM. Cet algorithme est utilisé pour compléter l'algorithme EM classique, en vue d'obtenir la précision des estimations du maximum de vraisemblance de θ sous la forme de la matrice de variance covariance asymptotique de $\hat{\theta}$. Il ne s'applique que sur des jeux de données où les observations sont i.i.d.. En résumé, cet algorithme permet de calculer la précision asymptotique des estimations par maximum de vraisemblance obtenues via l'algorithme EM.

2.2 Chaînes et semi-Chaînes de Markov cachées (HMC/HSMC)

Baum et Petrie ont introduit en 1966 les modèles de chaînes de Markov cachées, notés plus communément HMM (Hidden Markov Model). Cette classe de modèles repose sur l'hypothèse qu'une séquence n'est pas directement générée par une chaîne de Markov mais indirectement par des lois de probabilités attachées aux états de la chaîne de Markov. Ephraim et Merhav (2002) soulignent que l'une des premières applications de ces modèles fut la reconnaissance automatique de la parole à partir des années 70. Leur champ d'application s'est depuis beaucoup élargi, allant du traitement du signal à l'analyse de séquences d'ADN.

Ces modèles sont utilisés pour deux raisons principales. La première est la possibilité d'expliquer les variations du processus observé à partir des variations d'un processus sous-jacent caché. Pour exemple, Albert (1991) a étudié la distribution des fréquences de crise épileptique selon que le patient soit dans un état élevé ou bas d'activité de saisie informatique. La seconde raison d'utiliser les HMM est la possibilité de prédire un processus non observé à partir d'un processus observé. Dans le contexte de la reconnaissance de la parole, les données observées (le signal acoustique) peuvent être modélisées comme fonction des configurations articulatoires non-observées telles que le mouvement de la langue. Le principe repose sur deux phases : une phase d'apprentissage où les paramètres du HMM sont estimés à partir d'échantillons de signaux enregistrés, et une phase de reconnaissance où est choisi le modèle préalablement estimé expliquant au mieux un signal inconnu.

Les articles de Rabiner (1989) et Ephraim et Merhav (2002) constituent d'excellents tutoriaux sur les chaînes de Markov cachées (HMC, Hidden Markov Chain). Rabiner (1989) présente l'intérêt des chaînes de Markov cachées en reconnaissance de la parole et leurs possibles extensions. Ephraim et Merhav (2002) se placent dans un cadre beaucoup plus large au niveau des applications de ces modèles. Dans la suite de cette partie, nous nous appuierons également sur l'ouvrage de Cappé et al. (2005), ouvrage le plus complet sur les chaînes de Markov cachées et sur leur estimation. Contrairement aux deux autres articles de références qui se placent dans un espace d'états fini, Cappé et al. (2005) développent également la théorie des HMM dans un espace d'états continu.

Ferguson (1980) a proposé au début des années 1980 les semi-chaînes de Markov cachées (HSMC, Hidden Semi-Markov Chain) ; un résumé de ses travaux est fourni dans le tutorial de Rabiner (1989). Les semi-chaînes de Markov cachées généralisent les chaînes de Markov cachées en s'affranchissant de l'hypothèse de loi géométrique d'occupation des états (ou de temps de séjour dans les états). Les semi-chaînes de Markov permettent une modélisation beaucoup plus réaliste d'un grand nombre de structures dont par exemple la détection de gêne, l'étude de signaux acoustiques ou médicaux ou encore, l'analyse de séquences biologiques de différentes natures. Dans un article plus récent, Guédon (2003)

a posé de manière plus générale le problème des semi-chaînes de Markov cachées et de leur méthode d'estimation en prenant en compte la censure à droite.

2.2.1 Exemple introductif

Nous allons nous aider de l'exemple des sacs en papier[1] pour introduire les notions de chaîne de Markov et de chaîne de Markov cachée. Imaginons un jeu simple, avec des sacs en papier opaques contenant des jetons étiquetés. Nous disposons de 2 sacs :

- le sac C_1 contenant 1 jeton étiqueté a et 9 jetons étiquetés b,
- le sac C_2 contenant 4 jetons étiquetés a et 1 jeton étiqueté b.

Le jeu se déroule selon les règles suivantes :

- Nous commençons par piocher un jeton au hasard dans le sac C_1. Si l'on pioche un jeton a, on reste au sac C_1. Si l'on pioche un jeton b, on passe au sac C_2. On note également quel jeton a été tiré et on le remet dans son sac d'origine.
- On recommence 9 fois cette étape en prenant pour sac de départ, le sac associé au jeton pioché à l'étape précédente : le sac C_1 si le jeton pioché était a ou le sac C_2 si le jeton pioché était b.

En jouant plusieurs parties, nous pouvons obtenir les séquences suivantes : {a b a b a b a a b a} ou {a b b a b a b a b a} ou encore {a b b a b a b b a a}. Au vu du contenu des sacs, nous pouvons écrire les probabilités de passer d'un sac à un autre entre le $t^{ème}$ et le $(t+1)^{ème}$ tirage :

Sac	C_1	C_2
C_1	$\mathrm{P}(\mathrm{Sac}_{t+1} = C_1 \mid \mathrm{Sac}_t = C_1) =$ $\mathrm{P}(\mathrm{Jeton}_t = a \mid Sac_t = C_1) = 0.1$	$\mathrm{P}(\mathrm{Sac}_{t+1} = C_2 \mid \mathrm{Sac}_t = C_1) =$ $\mathrm{P}(\mathrm{Jeton}_t = b \mid \mathrm{Sac}_t = C_1) = 0.9$
C_2	$\mathrm{P}(\mathrm{Sac}_{t+1} = C_1 \mid \mathrm{Sac}_t = C_2) =$ $\mathrm{P}(\mathrm{Jeton}_t = a \mid Sac_t = C_2) = 0.8$	$\mathrm{P}(\mathrm{Sac}_{t+1} = C_2 \mid \mathrm{Sac}_t = C_2) =$ $\mathrm{P}(\mathrm{Jeton}_t = b \mid \mathrm{Sac}_t = C_2) = 0.2$

Ce premier jeu peut être modélisé par ***une chaîne de Markov*** d'ordre 1 à deux états : chaque sac représente un état, les probabilités de passer d'un sac à un autre entre deux tirages représentent les probabilités de transition et seul le dernier tirage a de l'influence sur le suivant d'où l'ordre 1.

Nous allons maintenant changer les règles du jeu et ajouter deux nouveaux sacs :

- le sac D_1 contenant 4 jetons rouges (R) et 1 jeton vert (V),
- le sac D_2 contenant 2 jetons rouges (R) et 3 jetons verts (V).

Les sacs C_1 et C_2 sont des sacs de transition permettant de savoir dans quels sacs effectuer le prochain tirage, les sacs D_1 et D_2 sont les sacs de sortie générant la séquence. Les règles du jeu deviennent :

[1] http ://fr.wikipedia.org/wiki/Modèle_de_Markov_caché

- On part du groupe de sac $\{C_1\ D_1\}$, on tire un jeton dans le sac D_1. On note sa couleur et on le replace dans son sac d'origine.
- On tire un jeton dans le sac C_1 pour savoir où se fera le prochain tirage (dans le groupe de sac $\{C_1\ D_1\}$ ou $\{C_2\ D_2\}$). On le replace.
- On réitère 3 fois le procédé en prenant pour groupe de sac de départ, le groupe de sac associé au jeton pioché à l'étape précédente : le groupe de sac $\{C_1\ D_1\}$ si le jeton pioché était a ou le groupe de sac $\{C_2\ D_2\}$ si le jeton pioché était b.

Deux séquences ont été générées par ce procédé : la séquence des couleurs (connue car noté à chaque étape) et la séquence des sacs (inconnue car pas noté à chaque étape). On peut noter que des séquences de sac différentes peuvent amener à de mêmes séquences de sortie. Par exemple :

Séquence de sacs inconnue	$C_1\ C_2\ C_2\ C_1$	$C_1\ C_2\ C_1\ C_2$	$C_1\ C_2\ C_1\ C_1$
Séquence de sortie connue	R V V R	R V V R	R V V R

Ce second jeu peut être modélisé par **une chaîne de Markov cachée** d'ordre 1 à deux états (figure 2.1) : chaque sac C_1 ou C_2 représente un état, la proportion de jeton d'une valeur dans un sac C_1 ou C_2 est la probabilité de transition, la proportion de jeton d'une certaine couleur dans un sac D_1 ou D_2 est la probabilité d'observation.

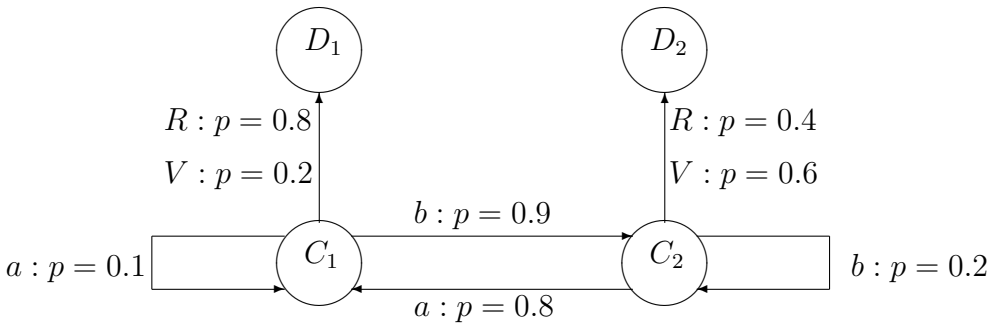

FIG. 2.1 – *Jeu des sacs en papiers cachés. Chaîne de Markov cachée d'ordre 1 à 2 états.*

2.2.2 Définitions

Nous allons introduire des notations pour la suite de cette partie. La suite des états et de leurs réalisations $S_1 = s_1, S_2 = s_2, \ldots, S_t = s_t$ sera notée $S_1^t = s_1^t$. Une chaîne de Markov d'ordre 1 est un processus stochastique à temps discret et à espace d'états discret caractérisé par la relation de dépendance suivante :

$$\mathrm{P}(S_t = s_t | S_1^{t-1} = s_1^{t-1}) = \mathrm{P}(S_t = s_t | S_{t-1} = s_{t-1}),$$

ce qui veut dire que tout le passé du processus est résumé dans l'état précédent, ou encore que, le présent étant connu (temps $t-1$), le futur est indépendant du passé.

Définition 2.1 *Soit $\{S_t\}$ une chaîne de Markov à valeurs dans l'espace d'états fini $\{1,\dots,J\}$. Une **chaîne de Markov** à J états d'ordre 1, homogène dans le temps, est définie par les paramètres suivants :*

- *probabilités initiales $\pi_j = P(S_1 = j); j = 1,\dots,J$ avec $\sum_j \pi_j = 1$.*
- *probabilités de transition $p_{ij} = P(S_t = j|S_{t-1} = i); i,j = 1,\dots,J$ avec $\sum_j p_{ij} = 1$.*

Dans le cas d'une chaîne de Markov, le temps de séjour dans l'état j est modélisé implicitement. Rester u fois dans un état consiste à boucler $(u-1)$ fois puis à sortir. La loi d'occupation ou du temps de séjour dans l'état j est donnée par

$$\begin{aligned} d_j(u) &= P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0,\dots,u-2|S_{t+1} = j, S_t \neq j) \\ &= P(S_{t+u+1} \neq j|S_{t+u} = j) \prod_{v=0}^{u-2} P(S_{t+u-v} = j|S_{t+u-v-1} = j) \\ &= (1-p_{jj})p_{jj}^{u-1}, \qquad u = 1,2,\dots \end{aligned}$$

La loi d'occupation de l'état j est la loi géométrique de paramètre $1-p_{jj}$. Les lois d'occupation sont déduites des paramètres du modèle et constituent seulement une caractéristique du modèle.

Une semi-chaîne de Markov est construite à partir d'une chaîne de Markov sous-jacente qui représente uniquement les transitions entre états distincts. Dans le cas d'une chaîne de Markov, le temps de séjour dans un état est modélisé implicitement par une loi géométrique tandis que dans le cas d'une semi-chaîne de Markov, le temps de séjour dans un état est explicitement modélisé par une loi de probabilité discrète quelconque (loi de Poisson, loi binomiale, loi binomiale négative, ...). Avant de donner la définition d'une semi-chaîne de Markov, nous allons introduire la notion d'état absorbant. Un état absorbant est un état dans lequel, une fois entré, il est impossible d'en sortir. Cette notion ne peut donc se traduire dans une loi de temps de séjour (ou d'occupation) d'un état. Les lois d'occupation ne sont donc définies que pour les états non-absorbants (c'est-à-dire tels que $p_{jj} < 1$).

Définition 2.2 *(Guédon, 2003) Une **semi-chaîne de Markov** se construit à partir d'une chaîne de Markov sous-jacente d'ordre 1. Cette chaîne de Markov d'ordre 1 à J états est définie par les paramètres suivants :*

- *probabilités initiales $\pi_j = P(S_1 = j); j = 1,\dots,J$ avec $\sum_j \pi_j = 1$.*
- *probabilités de transition*
  - *état i non-absorbant : $\forall j \neq i, \tilde{p}_{ij} = P(S_t = j|S_t \neq i, S_{t-1} = i)$ avec $\sum_{j\neq i} \tilde{p}_{ij} = 1$ and $\tilde{p}_{ii} = 0$,*
  - *état i absorbant : $p_{ii} = P(S_t = i|S_{t-1} = i) = 1$ et $\forall j \neq i, p_{ij} = 0$.*

Cette chaîne de Markov sous-jacente représente les transitions entre états distincts hormis le cas de l'état absorbant. À chaque état non-absorbant de cette chaîne de Markov est attachée une loi d'occupation explicite représentant le temps de séjour dans cet état

$$d_j(u) = P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-2 | S_{t+1} = j, S_t \neq j), u = 1, \ldots, U_j$$

où U_j est la borne supérieure du temps passé dans l'état non-absorbant j. Pour le cas particulier du dernier état visité, Guédon (2003) introduit la fonction de survie du temps de séjour dans l'état j, $D_j(u) = \sum_{v \geq u} d_j(v)$.

On peut se reporter à l'ouvrage de Kulkarni (1995) pour une référence générale sur les modèles semi-markoviens. Dans le cas d'une semi-chaîne de Markov, les lois d'occupation des états font parties de la définition du modèle. Les lois d'occupation des états sont des lois paramétriques discrètes telles que par exemple la loi binomiale $\mathrm{B}(d, n, p)$, la loi de Poisson $\mathrm{P}(d, \lambda)$ ou la loi binomiale négative $\mathrm{NB}(d, r, p)$ avec un paramètre additionnel de décalage $d \geq 1$; voir Guédon et al. (2007) pour une définition plus formelle de ces distributions.

Le mécanisme associé à une semi-chaîne de Markov peut s'interpréter comme suit : à un instant t donné, on passe de l'état i à un état j choisi selon la loi de transition de l'état i $(p_{i1}, \ldots, p_{iJ})$ puis l'on reste dans l'état j un temps u choisi selon la loi d'occupation de l'état j $\{d_j(u) ; u = 1, 2, \ldots\}$. Enfin, on effectue une nouvelle transition conformément à la loi de transition de l'état j $(p_{j1}, \ldots, p_{jJ})$.

Nous avons vu dans l'exemple introductif des sacs en papier que la séquence observée des couleurs n'est pas directement générée par une chaîne de Markov mais par des probabilités attachées aux états de la chaîne de Markov (les sacs C_1 et C_2). Par analogie avec les notations introduites précédemment, nous noterons par $Y_1^t = y_1^t$ la suite des variables observées et de leurs réalisations $Y_1 = y_1, Y_2 = y_2, \ldots, Y_t = y_t$.

Définition 2.3 *Une **(semi-)chaîne de Markov cachée** peut être vue comme un couple de processus stochastiques $\{S_t, Y_t\}$ tel que le processus $\{S_t\}$, appelé processus d'état ou caché (ce processus n'est pas observable), soit une (semi-)chaîne de Markov d'ordre 1 et que le processus $\{Y_t\}$, appelé processus d'observation ou d'émission, soit lié au processus d'état par une fonction probabiliste f telle que $Y_t = f(S_t)$. Le processus d'observation $\{Y_t\}$ est lié à la (semi-)chaîne de Markov sous-jacente $\{S_t\}$:*

- *si le processus $\{Y_t\}$ est discret, par les probabilités d'observation :*

$$b_j(y) = P(Y_t = y | S_t = j) \qquad \text{avec} \quad \sum_y b_j(y) = 1,$$

- *si le processus $\{Y_t\}$ est continu, par les densités d'observation :*

$$b_j(y) = f(y | S_t = j) \qquad \text{avec} \quad \int b_j(y) dy = 1.$$

La fonction f de l'espace d'états dans l'espace des observations est supposée telle qu'une observation peut être observée dans différents états. La définition des probabilités ou des densités d'observation exprime l'hypothèse selon laquelle le processus d'observation au temps t ne dépend que de la (semi-)chaîne de Markov sous-jacente au temps t.

Les chaînes de Markov cachées avec un processus d'observation discret ont été beaucoup utilisées, notamment pour l'analyse de séquence biologique d'ADN (Churchill, 1989) ou l'analyse des images à résonance magnétique chez les patients atteints de sclérose en plaques (Albert, 1991) ; de nombreux autres exemples sont donnés dans la monographie de MacDonald et Zucchini (1997). Les chaînes de Markov cachées gaussiennes où le processus d'observation $\{Y_t\}$ est lié au processus caché $\{S_t\}$ par la distribution gaussienne $Y_t|_{S_t=s_t} \sim \mathcal{N}(\mu_{s_t}, \sigma^2_{s_t})$ ont été introduites en reconnaissance de la parole au début des années 1980 et ont depuis été étendues à de très nombreuses applications dont la modélisation des canaux ioniques (Ephraim et Merhav, 2002; Cappé et al., 2005).
Les semi-chaînes de Markov cachées avec un processus d'observation discret ont trouvé leur application dans l'analyse de structures de ramification ou de floraison chez les plantes (Guédon et al., 2001) ou la reconnaissance de gènes (Burge et Karlin, 1997). Les semi-chaînes de Markov cachées gaussiennes ont été introduites en reconnaissance de la parole au début des années 1990 par Russell (1993).

Selon Leroux (1992), la loi marginale du processus $\{Y_t\}$ au temps t est un mélange fini de distributions pondérées par la probabilité d'être dans un état à un temps donné t

$$f(y_t) = \sum_{j=1}^{J} \mathrm{P}(S_t = j) b_j(y_t).$$

Une chaîne de Markov cachée peut donc être vue comme un modèle de mélange fini de distributions avec dépendances markoviennes (Ephraim et Merhav, 2002; Cappé et al., 2005). Par analogie, une semi-chaîne de Markov cachée peut être vue comme un modèle de mélange fini de distributions avec dépendances semi-markoviennes.

2.2.3 Propriétés d'indépendance conditionnelle et vraisemblances

La partie suivante est développée pour les chaînes et semi-chaînes de Markov cachées discrètes d'ordre 1. La généralisation aux processus d'observation continus est naturelle.

Avant de présenter les propriétés structurelles de ces processus stochastiques, nous allons définir la notion de graphe d'indépendance conditionnelle dérivée de la notion de modèles graphiques (Lauritzen, 1998). Soit A, B et C trois variables aléatoires sommets d'un graphe orienté sans cycle. Nous dirons que A est parent de B s'il existe un arc ayant pour origine A et pointant sur B. Nous dirons que C est enfant de B s'il existe un arc ayant pour origine B et pointant sur C. Le sommet A sera dit ancêtre de B s'il est soit parent de B, soit l'ancêtre d'un parent de B (définition récursive). Le sommet C sera dit descendant

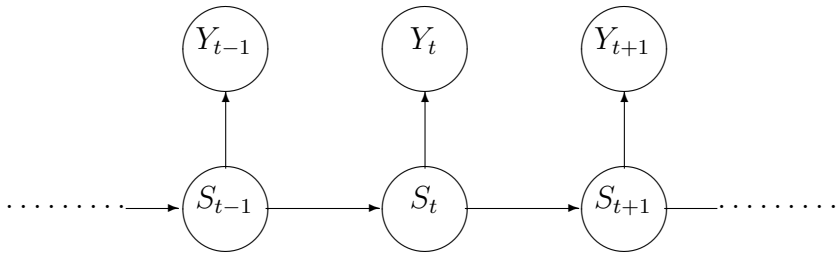

FIG. 2.2 – *Graphe des indépendances conditionnelles d'une chaîne de Markov cachée d'ordre 1.*

de A s'il est soit enfant de B, soit enfant d'un descendant de B (définition récursive). Pour les relations d'indépendance conditionnelle entre variables aléatoires, nous utiliserons la notation suivante :

$$A \perp B|C$$

qui exprime le fait que A est indépendant de B, sachant C.
Dans le cas d'une chaîne de Markov cachée, nous avons les deux relations d'indépendance conditionnelle (Smyth et al., 1997)

$$\begin{aligned} S_t &\perp \{S_1, Y_1, \ldots, S_{t-2}, Y_{t-2}, Y_{t-1}\} | S_{t-1}, \qquad t > 1, \\ Y_t &\perp \{S_1, Y_1, \ldots, S_{t-1}, Y_{t-1}\} | S_t, \qquad\quad t \geq 1. \end{aligned}$$

Dans le graphe orienté acyclique correspondant, une variable aléatoire est donc indépendante de toutes les autres variables aléatoires à l'exception de ses descendants connaissant ses parents. Ainsi, S_t "influence" directement Y_t (ou encore S_{t-1} "influence" directement S_t) alors que la réciproque n'est pas vraie. 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 dans le graphe de la figure 2.2, si l'on supprime le sommet S_t, les variables aléatoires S_{t-v} $(v > 0)$ sont séparées des variables aléatoires S_{t+v} $(v > 0)$ et la variable aléatoire Y_t est isolée, ce qui peut s'exprimer aussi en disant que l'état à l'instant t étant connu, la connaissance des valeurs prises par les variables aléatoires S_{t-v} n'influe pas sur les valeurs prises par les variables aléatoires S_{t+v} et Y_t.

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

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

Loi jointe et vraisemblance d'une chaîne de Markov cachée.

Les propriétés d'indépendance conditionnelle nous permettent à l'aide d'une factorisation de probabilités conditionnelles d'écrire la loi jointe d'une chaîne de Markov cachée discrète d'ordre 1, pour une séquence observée de longueur T :

$$\begin{aligned} P(S_1^T = s_1^T, Y_1^T = y_1^T) &= P(Y_1^T = y_1^T | S_1^T = s_1^T) P(S_1^T = s_1^T) \\ &= \mathrm{P}(S_1 = s_1)\mathrm{P}(Y_1 = y_1 | S_1 = s_1) \prod_{t=2}^{T} \mathrm{P}(S_t = s_t | S_{t-1} = s_{t-1})\mathrm{P}(Y_t = y_t | S_t = s_t) \\ &= \pi_{s_1} b_{s_1}(y_1) \prod_{t=2}^{T} p_{s_{t-1}s_t} b_{s_t}(y_t). \end{aligned} \tag{2.4}$$

Nous pouvons alors déduire de l'équation (2.4) la vraisemblance de la séquence des données observées $L(\theta)$ pour une chaîne de Markov cachée :

$$\begin{aligned} L(\theta) &= P(Y_1^T = y_1^T; \theta) \\ &= \sum_{s_1^T} P(S_1^T = s_1^T, Y_1^T = y_1^T; \theta) = \sum_{s_1^T} \Big(\pi_{s_1} b_{s_1}(y_1) \prod_{t=2}^{T} p_{s_{t-1}s_t} b_{s_t}(y_t) \Big) \end{aligned} \tag{2.5}$$

où $\sum_{s_1^T}$ signifie "somme sur toutes les séquences d'états possibles de longueur T" (J^T séquences possibles) et θ est l'ensemble des paramètres de la chaîne de Markov cachée (probabilités initiales, probabilités de transition et probabilités d'observation).

Loi jointe et vraisemblance d'une semi-chaîne de Markov cachée.

A partir des indépendances conditionnelles entre états distincts, on peut écrire la loi jointe d'une semi-chaîne de Markov cachée pour une séquence de longueur T (Guédon, 2003) :

$$\begin{aligned} P(S_1^T = s_1^T, Y_1^T = y_1^T) &= \mathrm{P}(Y_1^T = y_1^T | S_1^T = s_1^T) \mathrm{P}(S_1^T = s_1^T) \\ &= \pi_{s_1} d_{s_1}(u_1) \Big\{ \prod_{r=2}^{R-1} p_{s_{r-1}s_r} d_{s_r}(u_r) \Big\} p_{s_{R-1}s_R} d_{s_R}(u_R) I\Big(\sum_{r=1}^{R} u_r = T \Big) \prod_{t=1}^{T} b_{s_t}(y_t) \end{aligned} \tag{2.6}$$

où R est le nombre d'état visité, s_r est le $r^{ème}$ état visité, u_r est le temps passé dans l'état s_r et $I()$ représente la fonction indicatrice.

Par analogie avec les chaînes de Markov cachées, la vraisemblance de la séquence des données observées est la somme sur toutes les séquences d'états possible de longueur T des quantités de l'équation (2.6).

2.2.4 Méthodes d'estimation

2.2.4.1 Maximisation directe de la vraisemblance

Dans le cas d'une chaîne de Markov cachée, la vraisemblance des données observées (équation (2.5)) peut être réécrite sous la forme d'un produit de matrices (Altman, 2003; MacDonald et Zucchini, 1997, chap. 2). Soit A^1 le vecteur ligne de taille J constitué des éléments $A_j^1 = \pi_j b_j(y_1)$, soit A^t la matrice de dimension $J \times J$ constituée des éléments

$A^t_{ij} = p_{ij} b_j(y_t), t > 2$ et soit $\mathbf{1}$ le vecteur ligne unitaire de taille J. La vraisemblance des données observées s'écrit donc :

$$L(\theta) = \sum_{s_1^T} \Big(\pi_{s_1} b_{s_1}(y_1) \prod_{t=2}^{T} p_{s_{t-1} s_t} b_{s_t}(y_t) \Big) = (A^1)' \Big\{ \prod_{t=2}^{T} A^t \Big\} \mathbf{1}. \tag{2.7}$$

Cappé et al. (1998), Collings et Rydén (1998) et Turner (2008) ont proposé de maximiser directement la vraisemblance des données observées (équation (2.7)). Les approches proposées par Collings et Rydén (1998) et Cappé et al. (1998) reposent sur des méthodes numériques où le calcul du gradient de la vraisemblance des données observées est nécessaire. L'approche présentée par Cappé et al. (1998) pour estimer les paramètres d'une chaîne de Markov cachée consiste à utiliser des méthodes de quasi-Newton (Lange, 2004) où la log-vraisemblance et son gradient sont évalués récursivement. Collings et Rydén (1998) ont également proposé de maximiser récursivement la vraisemblance à partir du calcul du gradient et ont développé leur méthodologie pour les chaînes de Markov cachées gaussiennes. Les méthodes de Newton pouvant s'avérer peu fiable lorsque la matrice hessienne devient singulière, Turner (2008) a proposé une alternative basée sur l'algorithme de Levenberg-Marquardt. Cette approche a l'avantage d'être plus fiable et plus rapide.

La maximisation directe des semi-chaînes de Markov cachées s'avère plus délicate. En effet, l'écriture de la vraisemblance des données observées d'une semi-chaîne de Markov cachée sous la forme d'un produit de matrices est impossible. Par conséquent, le gradient et la matrice hessienne de la log-vraisemblance des données observées ne pouvant être calculés, les approches proposées dans la cadre des chaînes de Markov cachées ne peuvent pas se transposer aux semi-chaînes de Markov cachées.

Le processus caché $\{S_t\}$ n'étant pas observable, le problème d'estimation des paramètres peut être vu comme un problème aux données incomplètes. Dans ce cas, l'algorithme EM (section 2.1) est le candidat naturel pour l'estimation des paramètres par maximum de vraisemblance. Cet algorithme porte aussi le nom d'algorithme de Baum-Welch dans le cas de chaînes de Markov cachées (Baum et al., 1970). Selon Archer et Titterington (2002), l'algorithme EM pour chaînes de Markov cachées peut être vu comme un algorithme de "restauration-maximisation". En effet, l'étape E consiste en une restauration probabiliste de toutes les séquences d'états possibles. Il est important de souligner que cette restauration n'est pas explicite. Dans l'étape de maximisation, les paramètres sont obtenus en maximisant l'espérance des données complètes conditionnellement aux données observées.

Log-vraisemblance des données complètes d'une chaîne de Markov cachée

Pour la spécification du problème aux données complètes, nous supposons qu'à la fois la séquence y_1^T et la séquence d'états s_1^T sont observées. À partir de l'équation (2.4), nous pouvons écrire la log-vraisemblance des données complètes

$$\begin{aligned}\log P(S_1^T = s_1^T, Y_1^T = y_1^T; \theta) &= \sum_{j=1}^{J} I\Big(s_1 = j\Big)\log \pi_j + \sum_{i,j=1}^{J}\sum_{t=2}^{T} I\Big(s_t = j, s_{t-1} = i\Big)\log p_{ij} \\ &+ \sum_{j=1}^{J}\sum_{y=1}^{H}\sum_{t=1}^{T} I\Big(s_t = j, y_t = y\Big)\log b_j(y) \qquad (2.8)\end{aligned}$$

où H est le nombre de modalités de la variable discrète observée Y.

Log-vraisemblance des données complètes d'une semi-chaîne de Markov cachée

La prise en compte de la censure à droite du temps de séjour dans l'état final $D_{s_R}(u_R)$ dans la loi jointe d'une semi-chaîne de Markov cachée (équation (2.6)) entraîne des difficultés car cette quantité ne peut être utilisée dans la procédure d'estimation. Guédon (2003) propose de considérer la vraisemblance des données complètes où à la fois la séquence y_1^T et la séquence d'états s_1^{T+1+u} sont observées :

$$\begin{aligned}& P(S_1^{T+1+u} = s_1^{T+u+1}, Y_1^T = y_1^T; \theta) \\ = \;& \mathrm{P}(S_1^T = s_1^T, S_{T+v} = s_T, v = 1, \ldots, u, S_{T+u+1} \neq s_T, Y_1^T = y_1^T; \theta) \\ = \;& \pi_{s_1} d_{s_1}(u_1) \prod_{r=2}^{R} p_{s_{r-1}s_r} d_{s_r}(u_r) I\Big(\sum_{r=1}^{R-1} u_r < T \leq \sum_{r=1}^{R} u_r\Big) \prod_{t=1}^{T} b_{s_t}(y_t) \qquad (2.9)\end{aligned}$$

où θ est l'ensemble des paramètres de la semi-chaîne de Markov cachée (probabilités initiales, probabilités de transition, probabilités d'observation et lois d'occupation). La séquence d'états reste dans le dernier état visité s_T jusqu'au temps $T+u, u = 1, 2, \ldots$. La sortie du dernier état visité a lieu au temps $T+u+1$. Dans cette nouvelle spécification du problème aux données complètes, la séquence d'états est complétée jusqu'à sa sortie par l'état occupé au temps T, état supposé être un état non-absorbant. Nous pouvons déduire de l'équation (2.9) la vraisemblance de la séquence des données observées $\tilde{L}(\theta)$ pour une semi-chaîne de Markov cachée :

$$\tilde{L}(\theta) = P(Y_1^T = y_1^T; \theta) = \sum_{s_1^T} \sum_{u} P(S_1^{T+1+u} = s_1^{T+u+1}, Y_1^T = y_1^T; \theta) \qquad (2.10)$$

où $\sum_{s_1^T}$ signifie "somme sur toutes les séquences d'états possibles de longueur T" (J^T séquences possibles) et $\sum_u$ signifie "somme sur tous les temps supplémentaires depuis le temps $T+1$ passés dans l'état occupé au temps T ".

La log-vraisemblance des données complètes pour une semi-chaîne de Markov cachée s'écrit donc

$$\begin{aligned}
&\log P(S_1^{T+u+1} = s_1^{T+u+1}, Y_1^T = y_1^T; \theta) \\
= &\sum_{j=1}^{J} I\Big(s_1 = j\Big)\log \pi_j + \sum_{i=1}^{J}\sum_{j\neq i}\sum_{t=2}^{T} I\Big(s_t = j, s_{t-1} = i\Big)\log \tilde{p}_{ij} \\
+ &\sum_{j=1}^{J}\Bigg\{\sum_{u}\Big\{\sum_{t=1}^{T-1} I\Big(s_{t+u+1} \neq j, s_{t+u-v} = j, v = 0, \ldots, u-1, s_t \neq j\Big) \\
&\qquad + I\Big(s_{u+1} \neq j, s_{u-v} = j, v = 0, \ldots, u-1\Big)\Big\}\log d_j(u)\Bigg\} I\Big(\tilde{p}_{jj} = 0\Big) \\
+ &\sum_{j=1}^{J}\sum_{y=1}^{H}\sum_{t=1}^{T} I\Big(s_t = j, y_t = y\Big)\log b_j(y) \qquad (2.11)
\end{aligned}$$

2.2.4.2 Algorithme EM avec restauration probabiliste des séquences d'états

L'algorithme EM pour chaînes de Markov cachées et semi-chaînes de Markov cachées repose sur une restauration probabiliste de l'ensemble des séquences d'états possibles à l'étape E. Cette restauration n'est pas explicite.

a) Étape E

On s'intéresse à l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées (cf équation (2.1) dans un cadre général) qui s'écrit

– pour les chaînes de Markov cachées discrètes à partir de l'équation (2.8) :

$$\begin{aligned}
Q(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log P(y_1^T, s_1^T; \theta) \mid Y_1^T = y_1^T; \theta^{(k)}\Big) \\
&= Q_\pi\Big(\{\pi_j\}_{j=1}^J|\theta^{(k)}\Big) + Q_p\Big(\{p_{ij}\}_{j=1}^J|\theta^{(k)}\Big) \\
&+ Q_b\Big(\{b_j(y)\}_{y=1}^H|\theta^{(k)}\Big) \qquad (2.12)
\end{aligned}$$

– pour les semi-chaînes de Markov cachées discrètes à partir de l'équation (2.11) :

$$\begin{aligned}
\tilde{Q}(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log P(y_1^T, s_1^{T+u}; \theta) \mid Y_1^T = y_1^T; \theta^{(k)}\Big) \\
&= Q_\pi\Big(\{\pi_j\}_{j=1}^J|\theta^{(k)}\Big) + \tilde{Q}_{\tilde{p}}\Big(\{\tilde{p}_{ij}\}_{j=1}^J|\theta^{(k)}\Big) \\
&+ \tilde{Q}_d\Big(\{d_j(u)\}|\theta^{(k)}\Big) I\Big(\tilde{p}_{jj} = 0\Big) \\
&+ Q_b\Big(\{b_j(y)\}_{y=1}^H|\theta^{(k)}\Big) \qquad (2.13)
\end{aligned}$$

avec

$$Q_\pi\left(\{\pi_j\}_{j=1}^J|\theta^{(k)}\right) = \sum_j P\left(S_1 = j|Y_1^T = y_1^T; \theta^{(k)}\right)\log \pi_j, \qquad (2.14)$$

$$Q_p\left(\{p_{ij}\}_{j=1}^J|\theta^{(k)}\right) = \sum_{i,j=1}^{J}\sum_{t=2}^{T} P\left(S_t = j, S_{t-1} = i|Y_1^T = y_1^T; \theta^{(k)}\right)\log p_{ij}, \qquad (2.15)$$

$$\tilde{Q}_{\tilde{p}}\left(\{\tilde{p}_{ij}\}_{j=1}^{J}|\theta^{(k)}\right) = \sum_{i=1}^{J}\sum_{j\neq i}\sum_{t=2}^{T} P\left(S_t = j, S_{t-1} = i|Y_1^T = y_1^T;\theta^{(k)}\right)\log\tilde{p}_{ij}, \tag{2.16}$$

$$\begin{aligned}
&\tilde{Q}_d\Big(\{d_j(u)\}|\theta^{(k)}\Big)\\
&= \sum_{j=1}^{J}\sum_{u}\Bigg\{\sum_{t=1}^{T-1} P(S_{t+u+1}\neq j, S_{t+u-v} = j, v = 0,\ldots,u-1, S_t\neq j|Y_1^T = y_1^T;\theta^{(k)})\\
&+ P(S_u\neq j, S_{u-v} = j, v = 1,\ldots,u-1|Y_1^T = y_1^T;\theta^{(k)})\Bigg\}\log d_j(u)
\end{aligned} \tag{2.17}$$

et

$$Q_b\left(\{b_j(y)\}_{y=1}^{H}|\theta^{(k)}\right) = \sum_{j=1}^{J}\sum_{y=1}^{H}\sum_{t=1}^{T} P\left(Y_t = y, S_t = j|Y_1^T = y_1^T;\theta^{(k)}\right)\log b_j(y). \tag{2.18}$$

L'étape E de l'algorithme EM peut être efficacement implémentée par un algorithme dit "avant-arrière" ("forward-backward").

Pour les chaînes de Markov cachées, l'algorithme "avant-arrière" proposé par Devijver (1985) est basé sur la décomposition suivante des probabilités lissées $L_j(t)$:

$$\begin{aligned}
L_j(t) &= P\left(S_t = j|Y_1^T = y_1^T\right)\\
&= \frac{P\left(Y_{t+1}^T = y_{t+1}^T|S_t = j\right)}{P\left(Y_{t+1}^T = y_{t+1}^T|Y_1^t = y_1^t\right)}P\left(S_t = j|Y_1^t = y_1^t\right)\\
&= B_j(t)\,F_j(t)
\end{aligned} \tag{2.19}$$

qui traduit l'indépendance conditionnelle entre le passé et le futur du processus à chaque instant t. Devijver (1985) montre que les quantités $F_j(t)$ peuvent être calculées à l'aide d'une passe avant (c'est-à-dire de 1 à T) tandis que les quantités $B_j(t)$ ou $L_j(t)$ peuvent être calculées par une passe arrière (c'est-à-dire de T à 1). Au sens des modèles à espace d'états, l'algorithme " avant" seul est un algorithme de filtrage alors que l'algorithme "avant-arrière" est un algorithme de lissage.

Pour les semi-chaînes de Markov cachées, l'algorithme " avant-arrière" proposé par Guédon (2003) est basé sur la décomposition suivante :

$$\begin{aligned}
L1_j(t) &= P\left(S_{t+1}\neq j, S_t = j|Y_1^T = y_1^T\right)\\
&= \frac{P\left(Y_{t+1}^T = y_{t+1}^T|S_{t+1}\neq j, S_t = j\right)}{P\left(Y_{t+1}^T = y_{t+1}^T|Y_1^t = y_1^t\right)}P\left(S_{t+1}\neq j, S_t = j|Y_1^t = y_1^t\right)\\
&= B\text{HSMC}_j(t)\,F\text{HSMC}_j(t)
\end{aligned}$$

qui traduit l'indépendance conditionnelle entre le passé et le futur du processus aux instants de changement d'état.

Récurrence avant pour une chaîne de Markov cachée

La récurrence avant implique le calcul des probabilités de filtrage $F_j(t)$ pour chaque état j du temps 1 au temps T.

La récurrence avant (Devijver, 1985) est initialisée pour $t = 1$, par

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

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

$$\begin{aligned} N_1 &= \sum_j P(S_1 = j, Y_1 = y_1) \\ &= \sum_j b_j(y_1) \pi_j. \end{aligned}$$

Dans le cas général, la récurrence avant s'écrit, $t = 2, \ldots, T$:

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

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

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

Pour implémenter efficacement cette récurrence avant, il faut donc dans un premier temps calculer les quantités $P\left(S_t = j,\ Y_t = y_t |\ Y_1^{t-1} = y_1^{t-1}\right)$, puis en déduire par sommation le facteur de normalisation N_t, et enfin extraire les quantités "avant" en effectuant la normalisation $F_j\left(t\right) = P\left(S_t = j,\ Y_t = y_t |\ Y_1^{t-1} = y_1^{t-1}\right) / N_t$.

La récurrence avant peut être utilisée pour calculer la vraisemblance de la séquence observée pour le paramètre θ (voir équation (2.5)). En effet, nous avons la relation suivante :

$$L(\theta) = P(Y_1 = y_1; \theta) \prod_{t=2}^{T} P(Y_t = y_t | Y_1^{t-1} = y_1^{t-1}; \theta) = \prod_{t=1}^{T} N_t.$$

La log-vraisemblance des données observées, directement extraite de la récurrence avant, peut être utilisée pour contrôler la convergence de l'algorithme EM.

Récurrence avant pour une semi-chaîne de Markov cachée

La récurrence avant implique le calcul des probabilités $F_{\text{HSMC}_j}(t)$ pour chaque état j du temps 1 au temps T.

La récurrence avant est donnée par Guédon (2003), $t = 1, \ldots, T-1$:

$$\begin{aligned} F_{\text{HSMC}_j}(t) &= P\left(S_{t+1} \neq j, S_t = j | Y_1^t = y_1^t\right) \\ &= \frac{b_j(y_t)}{\tilde{N}_t}\left[\sum_{u=1}^{t-1}\left\{\prod_{v=1}^{u-1}\frac{b_j(y_{t-v})}{\tilde{N}_{t-v}}\right\} d_j(u)\sum_{i\neq j}\tilde{p}_{ij}F_{\text{HSMC}_i}(t-u)\right. \\ &\qquad \left. + \left\{\prod_{v=1}^{t-1}\frac{b_j(y_{t-v})}{\tilde{N}_{t-v}}\right\} d_j(t+1)\pi_j\right], \end{aligned} \tag{2.20}$$

où $\tilde{N}_t = P(Y_t = y_t | Y_1^{t-1} = y_1^{t-1})$ est un facteur de normalisation.
La censure au temps T du temps de séjour dans le dernier état visité nous conduit à distinguer le cas $t = T$

$$\begin{aligned} F_{\text{HSMC}_j}(T) &= P\left(S_T = j | Y_1^T = y_1^T\right) \\ &= \frac{b_j(y_T)}{\tilde{N}_T}\left[\sum_{u=1}^{T-1}\left\{\prod_{v=1}^{u-1}\frac{b_j(y_{T-v})}{\tilde{N}_{T-v}}\right\} D_j(u)\sum_{i\neq j}\tilde{p}_{ij}F_{\text{HSMC}_i}(T-u)\right. \\ &\qquad \left. + \left\{\prod_{v=1}^{T-1}\frac{b_j(y_{T-v})}{\tilde{N}_{T-v}}\right\} D_j(T+1)\pi_j\right]. \end{aligned} \tag{2.21}$$

Le temps exact passé dans le dernier état visité est inconnu, seulement le temps minimum est connu. Par conséquent, la loi du temps de séjour dans l'état j dans la formule générale de la passe avant (équation (2.20)) est remplacé par la fonction de survie correspondante dans l'équation (2.21).

Le facteur de normalisation $\tilde{N}_t = P(Y_t = y_t | Y_1^{t-1} = y_1^{t-1})$ est obtenu directement durant la recursion avant (Guédon, 2005) et est égal à

$$\tilde{N}_t = \sum_j P(S_t = j, Y_t = y_t | Y_1^{t-1} = y_1^{t-1})$$

où

$$P(S_1 = j, Y_1 = y_1) = b_j(y_1)\pi_j$$

et, pour $t = 2, \ldots, T$:

$$\begin{aligned} & P(S_t = j, Y_t = y_t | Y_1^{t-1} = y_1^{t-1}) \\ = \; & b_j(y_t)\Big\{\sum_{i\neq j} p_{ij}F_i(t-1) - F_j(t-1) + P(S_{t-1} = j | Y_1^{t-1} = y_1^{t-1})\Big\}, \end{aligned}$$

avec $P(S_{t-1} = j|Y_1^{t-1} = y_1^{t-1}) = P(S_{t-1} = j, Y_{t-1} = y_{t-1}|Y_1^{t-2} = y_1^{t-2})/N_{t-1}$.

Par analogie avec les chaînes de Markov cachée, la vraisemblance des données observées (équation (2.10)) peut être directement extraite de la récurrence avant comme étant le produit des constantes de normalisation. Elle peut être utilisée pour contrôler la convergence de l'algorithme EM.

Récurrence arrière pour une chaîne de Markov cachée

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

La récurrence arrière est initialisée pour $t = T$ par :

$$L_j(T) = P(S_T = j|Y_1^T = y_1^T) = F_j(T).$$

La récurrence arrière s'écrit pour $t = T-1, \ldots, 1$ (Devijver, 1985) :

$$\begin{aligned} L_j(t) &= P(S_t = j|Y_1^T = y_1^T) \\ &= \frac{1}{N_{t+1}}\Big\{\sum_k \frac{L_k(t+1)}{F_k(t+1)} b_k(y_{t+1}) p_{jk}\Big\} F_j(t) \\ &= \Big\{\sum_k \frac{L_k(t+1)}{G_k(t+1)} p_{jk}\Big\} F_j(t) \end{aligned}$$

où la quantité $G_k(t+1) = P(S_{t+1} = k|Y_1^t = y_1^t) = \sum_j p_{jk} F_j(t)$ peut être directement extraite et stockée en mémoire durant la récurrence avant. Dans le vocabulaire des modèles à espace d'états, $G_k(t+1)$ est la probabilité prédite.

Récurrence arrière pour une semi-chaîne de Markov cachée

La récurrence arrière consiste à calculer $L_j(t) = P(S_t = j|Y_1^T = y_1^T)$ pour chaque état j en reculant de T à 1.

La récurrence arrière est initialisée pour $t = T$ par :

$$L_j(T) = P(S_T = j|Y_1^T = y_1^T) = F_{\text{HSMC}_j}(T),$$

et est basée sur la décomposition suivante des probabilités $L_j(t)$ pour $t = T-1, \ldots, 1$:

$$\begin{aligned} L_j(t) &= P\left(S_t = j|Y_1^T = y_1^T\right) \\ &= P\left(S_{t+1} \neq j, S_t = j|Y_1^T = y_1^T\right) + P\left(S_{t+1} = j|Y_1^T = y_1^T\right) \\ &\quad - P\left(S_{t+1} = j, S_t \neq j|Y_1^T = y_1^T\right) \\ &= L1_j(t) + L_j(t+1) - P\left(S_{t+1} = j, S_t \neq j|Y_1^T = y_1^T\right). \end{aligned} \tag{2.22}$$

La récurrence arrière est basée sur le calcul de $L1_j(t)$:

$$\begin{aligned} L1_j(t) &= P\left(S_{t+1} \neq j, S_t = j | Y_1^T = y_1^T\right) \\ &= \left[\sum_{k\neq j}\left[\sum_{u=1}^{T-1-t} \frac{L1_k(t+u)}{F\text{HSMC}_k(t+u)}\left\{\prod_{v=0}^{u-1}\frac{b_k(y_{t+u-v})}{\tilde{N}_{t+u-v}}\right\} d_k(u)\right.\right. \\ &+ \left.\left.\left\{\prod_{v=0}^{T-1-t}\frac{b_k(y_{T-v})}{\tilde{N}_{T-v}}\right\} D_k(T-t)\right]\tilde{p}_{jk}\right] F\text{HSMC}_j(t) \\ &= \left\{\sum_{k\neq j} H_k(t+1)\tilde{p}_{jk}\right\} F\text{HSMC}_j(t). \end{aligned}$$

Le troisième terme dans l'équation (2.22) est donné par

$$\begin{aligned} & P\left(S_{t+1} = j, S_t \neq j | Y_1^T = y_1^T\right) \\ = & \left[\sum_{u=1}^{T-1-t} \frac{L1_j(t+u)}{F\text{HSMC}_j(t+u)}\left\{\prod_{v=0}^{u-1}\frac{b_j(y_{t+u-v})}{\tilde{N}_{t+u-v}}\right\} d_j(u)\right. \\ + & \left.\left\{\prod_{v=0}^{T-1-t}\frac{b_j\left(y_{T-v}\right)}{\tilde{N}_{T-v}}\right\} D_j(T-t)\right]\sum_{i\neq j}\tilde{p}_{ij} F\text{HSMC}_i(t) \\ = & H_j(t+1)\sum_{i\neq j}\tilde{p}_{ij} F\text{HSMC}_i(t). \end{aligned}$$

Pour faire ces calculs, nous introduisons la quantité

$$\begin{aligned} H_j(t+1) &= \frac{P(Y_{t+1}^T = y_{t+1}^T | S_{t+1} = j, S_t \neq j)}{P(Y_{t+1}^T = y_{t+1}^T | Y_1^t = y_1^t)} \\ &= \sum_{u=1}^{T-t} H_j(t+1,u), \end{aligned}$$

avec

$$\begin{aligned} H_j\left(t+1,u\right) &= \frac{L1_j\left(t+u\right)}{F\text{HSMC}_j\left(t+u\right)}\left\{\prod_{v=0}^{u-1}\frac{b_j\left(y_{t+u-v}\right)}{\tilde{N}_{t+u-v}}\right\} d_j\left(u\right), \qquad u = 1,\ldots,T-1-t, \\ H_j\left(t+1,T-t\right) &= \left\{\prod_{v=0}^{T-1-t}\frac{b_j\left(y_{T-v}\right)}{\tilde{N}_{T-v}}\right\} D_j\left(T-t\right). \end{aligned}$$

b) Étape M - Ré-estimation des paramètres

Les formules pour ré-estimer les paramètres de ces processus markoviens cachés sont obtenues en maximisant les différents termes de $Q(\theta|\theta^{(k)})$ (voir la décomposition 2.12) et de $\tilde{Q}(\theta|\theta^{(k)})$ (voir la décomposition 2.13), chaque terme dépendant d'un sous ensemble donnée de θ. Dans la suite, nous donnons seulement les formules de ré-estimations qui sont directement déduites en maximisant les équations 2.14, 2.15, 2.16, 2.17 et 2.18.

– probabilités initiales

$$\pi_j^{(k+1)} = P(S_1 = j | Y_1^T = y_1^T; \theta^{(k)}) = L_j^{(k)}(1),$$

– probabilités de transition

$$\begin{aligned} p_{ij}^{(k+1)} &= \frac{\sum_{t=2}^{T} P\left(S_t = j, S_{t-1} = i | Y_1^T = y_1^T; \theta^{(k)}\right)}{\sum_{t=2}^{T} P\left(S_{t-1} = i | Y_1^T = y_1^T; \theta^{(k)}\right)} \\ &= \frac{\sum_{t=2}^{T} L_j^{(k)}(t) p_{ij}^{(k)} F_i^{(k)}(t-1) / G_j^{(k)}(t)}{\sum_{t=2}^{T} L_i^{(k)}(t-1)}; \qquad \text{chaîne de Markov cachée,} \end{aligned}$$

$$\begin{aligned} \tilde{p}_{ij}^{(k+1)} &= \frac{\sum_{t=2}^{T} P\left(S_t = j, S_{t-1} = i | Y_1^T = y_1^T; \theta^{(k)}\right)}{\sum_{t=2}^{T} P\left(S_t \neq i, S_{t-1} = i | Y_1^T = y_1^T; \theta^{(k)}\right)} \\ &= \frac{\sum_{t=2}^{T} H_j^{(k)}(t) \tilde{p}_{ij}^{(k)} F\text{HSMC}_i^{(k)}(t-1)}{\sum_{t=2}^{T} L1_i^{(k)}(t-1)}; \qquad \text{semi-chaîne de Markov cachée,} \end{aligned}$$

– probabilité d'observation

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

– lois du temps de séjour

$$d_j^{(k+1)}(u) = \frac{\eta_{j,u}^{(k)}}{\sum_{t=1}^{T-1} L1_j(t) + L_j(T)},$$

avec $\eta_{j,u}^{(k)} = \sum_{t=1}^{T} P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1, S_t \neq j | Y_1^T = y_1^T; \theta^{(k)}) + P(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u-1 | Y_1^T = y_1^T; \theta^{(k)})$. Le calcul de la quantité $\eta_{j,u}^{(k)}$ est décrit dans Guédon (2003).

2.2.4.3 Algorithme MCEM avec restauration par simulation des séquences d'états

L'algorithme MCEM, dérivé de l'algorithme EM décrit précédemment, est également fréquemment utilisé pour estimer les paramètres des chaînes et des semi-chaînes de Markov cachées. Selon Archer et Titterington (2002), l'algorithme MCEM adapté aux chaînes de Markov cachées peut être vu comme un algorithme de "restauration-maximisation". Dans l'étape de restauration, les séquences d'états sont simulées selon leur distribution sachant les données observées. Dans l'étape de maximisation, les paramètres sont obtenus en maximisant l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées.

La simulation d'une séquence d'états s_1^T à partir de la séquence observée y_1^T est basée sur la factorisation de $S_1^T = s_1^T | Y_1^T = y_1^T$. Pour une chaîne de Markov cachée, la factorisation est la suivante :

$$\begin{aligned} & \mathrm{P}(S_1^T = s_1^T | Y_1^T = y_1^T) \\ = & \Big\{ \prod_{t=1}^{T-1} \mathrm{P}(S_t = s_t | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T) \Big\} \mathrm{P}(S_T = s_T | Y_1^T = y_1^T). \end{aligned}$$

Pour une semi-chaîne de Markov cachée, la factorisation est la suivante (Guédon, 2007) :

$$\begin{aligned} & \mathrm{P}(S_1^T = s_1^T | Y_1^T = y_1^T) \\ = & P(S_v = s_1, v = 1, \ldots, u-2 | S_{u-1} = s_1, S_u \neq s_1, S_u^T = s_u^T, Y_1^T = y_1^T) \\ \times & P(S_{u-1} = s_{u-1} | S_{u-1} \neq s_u, S_u^T = s_u^T, Y_1^T = y_1^T) \\ \times & P(S_{u-1} \neq s_u, S_u^T = s_u^T, Y_1^T = y_1^T). \end{aligned}$$

L'algorithme "avant-arrière" de simulation se décompose en une passe avant identique à la récurrence avant de l'algorithme "avant-arrière" présenté précédemment et en une passe arrière de simulation. Les paragraphes suivants s'appuient sur l'article de Chib (1996) pour les chaînes de Markov cachées et sur l'article de Guédon (2007) pour les semi-chaînes de Markov cachées.

Passe arrière pour une chaîne de Markov cachée

La passe arrière est initialisée pour $t = T$ par :

$$P(S_T = j | Y_1^T = y_1^T) = F_j(T).$$

L'état final s_T est simulé selon les probabilités lissées

$$\Big(P(S_T = j | Y_1^T = y_1^T); j = 1, \ldots, J \Big).$$

La passe arrière s'écrit pour $t = T-1, \ldots, 1$:

$$P(S_t = j | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T) \quad = \quad \frac{p_{j s_{t+1}} F_j(t)}{G_{s_{t+1}}(t+1)},$$

où $F_j(t)$, la probabilité filtrée, et $G_{s_{t+1}}(t+1) = \sum_i p_{i s_{t+1}} F_i(t)$, la probabilité prédite, sont calculées au cours de la récurrence avant.
L'état s_t est, quant à lui, simulé selon la loi conditionnelle

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

Passe arrière pour une semi-chaîne de Markov cachée

La passe arrière est initialisée pour $t = T$ par :

$$P(S_T = j|Y_1^T = y_1^T) = F\text{HSMC}_j(T).$$

L'état final s_T est simulé selon les probabilités lissées

$$\Big(P(S_T = j|Y_1^T = y_1^T); j = 1, \ldots, J\Big).$$

La passe arrière repose pour le changement d'état sur

$$P(S_t = j|S_t \neq s_{t+1}, S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T) \quad = \quad \frac{\tilde{p}_{js_{t+1}} F\text{HSMC}_j(t)}{\sum_{i \neq s_{t+1}} \tilde{p}_{is_{t+1}} F\text{HSMC}_i(t)},$$

où $F\text{HSMC}_j(t)$ la probabilité filtrée est calculée au cours de la récurrence avant.
Le changement d'état s_t est simulé selon la loi conditionnelle

$$\Big(P(S_t = j|S_t \neq s_{t+1}, S_{t+1}^T, s_{t+1}^T, Y_1^T = y_1^T); j = 1, \ldots, J\Big).$$

Il est également nécessaire de calculer le temps passé dans chaque état à l'instant t :

$t = T$:

$$P(S_{T-u} \neq s_T, S_{T-v} = s_T, v = 1, \ldots, u-1|S_T = s_T, Y_1^T = y_1^T)$$

$t < T$:

$$P(S_{t-u} \neq s_t, S_{t-v} = s_t, v = 1, \ldots, u-1|S_t = s_t, S_{t+1} \neq s_t, S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T).$$

Les calculs de ces quantités sont donnés dans Guédon (2007).
Le temps de séjour dans l'état j est simulé selon la loi conditionnelle

$$\Big(P(S_{t-u} \neq j, S_{t-v} = j, v = 1, \ldots, u-1|S_t = j, S_{t+1} \neq j, S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T), u = 1, \ldots, U_j\Big).$$

L'étape M consiste à maximiser en le paramètre θ, la moyenne des log-vraisemblances des données complètes. Comme l'algorithme MCEM pour (semi-)chaîne de Markov cachée est fondé sur une restauration explicite par simulation des séquences d'états, l'étape M est alors basée sur les comptages extraits des séquences d'états simulées : nombre d'états simulés prenant une valeur particulière, nombre de transitions entre états simulés, nombre d'occurrences successives d'un état ou encore nombre de fois où une observation est émise depuis un état simulé prenant une valeur particulière. Les quantités ré-estimées pour chaque paramètre sont quasi identiques à celles obtenues par l'algorithme EM hormis que les probabilités lissées $L_j(t)$ et les probabilités $L1_j(t)$, qui peuvent être vues comme des pseudo-comptages, sont remplacées par les indicatrices des valeurs des états simulés.

2.2.5 Exploration de l'espace des séquences d'états

Dans diverses situations, il est intéressant de connaître la séquence d'états la plus probable, c'est-à-dire celle qui explique au mieux la séquence observée pour un modèle donné. Ceci est réalisé par un algorithme de programmation dynamique appelé algorithme de Viterbi (Forney, 1973). L'algorithme de Viterbi est l'équivalent de l'algorithme "avant" en termes de programmation dynamique. Foreman (1993) a généralisé l'algorithme de Viterbi en introduisant l'algorithme de Viterbi généralisé dont l'objectif est de restaurer les L séquences d'états les plus probables pour une séquence observée donnée. Guédon (2007) a proposé l'algorithme de Viterbi "avant-arrière" qui permet d'explorer l'espace des séquences d'états sous forme de profils d'états ou de profils de changements d'états. L'algorithme de Viterbi "avant-arrière" est l'équivalent de l'algorithme "avant-arrière" en termes de programmation dynamique. Les deux derniers algorithmes sont décrits dans Guédon (2007) et peuvent être qualifiés d'outils de diagnostic.

2.2.6 Propriétés asymptotiques

Une chaîne de Markov cachée ergodique n'est jamais identifiable, dans le sens où nous pouvons toujours permuter les états sans changer la vraisemblance (Leroux, 1992; MacDonald et Zucchini, 1997). Cependant, il est possible de déterminer des conditions suffisantes pour satisfaire la propriété d'identifiabilité. Bickel et al. (1998) expliquent que si la chaîne de Markov cachée a une paramétrisation usuelle, si les mélanges finis de densités conditionnelles des observations sont identifiables et si les paramètres de ces densités sont distincts, alors cette chaîne est identifiable. La famille pour laquelle les mélanges finis sont identifiables sont les distributions gaussiennes, de Poisson ou exponentielles.

Dans le cas d'une chaîne de Markov cachée où le processus d'état est à valeur dans un espace d'états fini, Leroux (1992) et Bickel et al. (1998) ont montré, sous certaines conditions, la consistence et la normalité asymptotique de l'estimateur du maximum de vraisemblance. En plus de supposer l'identifiabilité du modèle, ces auteurs imposent des conditions sur les probabilités de transition et sur la distribution du processus d'observation. La chaîne de Markov sous-jacente doit notamment être ergodique, c'est-à-dire irréductible et apériodique. Bickel et al. (1998) ont aussi prouvé que la matrice d'information des observations est un estimateur consistent de la matrice d'information de Fisher ; la matrice d'information converge en probabilité vers la matrice d'information de Fisher sous la condition d'ergodicité de la chaîne de Markov sous-jacente. Cappé et al. (2005) ont présenté des résultats similaires pour les chaînes de Markov cachées où le processus d'état est à valeur dans un espace d'états continu.

Actuellement, seul l'article de Barbu et Limnios (2006) traite des propriétés asymptotiques des semi-chaînes de Markov cachées ; c'est-à-dire la consistence et la normalité asymptotique des estimateurs non paramétriques.

2.2.7 Remarque

Dans le cadre de la modélisation de la croissance des plantes, Guédon et al. (2007) ont montré que la composante ontogénique pouvait être modélisée par un modèle semi-markovien de type "gauche-droite" où les états sont ordonnés et où chaque état ne peut être visité au maximum qu'une fois. Comme la dernière année d'observation est arbitraire au regard du développement d'un arbre, la longueur de la dernière phase de croissance est supposée être systématiquement censurée à droite et ne peut donc pas être modélisée par une loi de temps de séjour. Le dernière phase de croissance sera donc modélisée par un état final absorbant. Par suite, la structure sous-jacente étant constituée d'une succession d'états transitoires avec un état final absorbant, la propriété d'ergodicité ne sera alors pas respectée. Il en va par conséquent de même pour les propriétés asymptotiques.

2.3 Modèles linéaires généralisés (GLM)

Au début du $19^{ème}$ siècle, en analysant des données issues du domaine de l'astronomie (principalement des mesures de quantités continues), Legendre et Gauss ont introduit la notion de modèles linéaires. Ils ont développé la méthode des moindres carrés ordinaires en modélisant les erreurs de mesure par une loi dite "gaussienne". Ce n'est que bien plus tard que Gauss s'est rendu compte que cette méthode d'estimation était plus justifiée par l'hypothèse de données indépendantes et de variance constante que par l'hypothèse de normalité. Depuis et principalement au début du $20^{ème}$ siècle, de nouvelles distributions sont venues s'ajouter à la distribution gaussienne, sous l'impulsion de Fisher. En effet, la nature des données à exploiter se diversifiant au cours du temps, l'analyse de données discrètes sous forme binaire ou de comptage s'est développée. Les distributions binomiales ou de Poisson ont alors fait leur apparition. Ces différentes lois ont été regroupées au sein d'une même famille, appelée famille exponentielle. Une nouvelle classe de modèles est née : la classe des modèles linéaires généralisés dont la terminologie a été introduite par Nelder et Wedderburn (1972). Elle généralise les modèles linéaires classiques en termes de loi de probabilité et de lien à la linéarité.

Les modèles linéaires généralisés permettent donc la modélisation de variables réponses dont la loi appartient à la famille exponentielle. Ces variables peuvent être de différents types : binaires (présence/absence de sexualité, mortalité ou non de l'apex), ordinales (pas de rameau/rameau avorté/rameau développé), de comptage (nombre d'entrenoeuds, nombre de rameaux portés par pousse annuelle), ou exponentielles (durée de vie des rameaux portés) par exemple. Les modèles linéaires généralisés ont pris une place importante dans la modélisation statistique, trouvant leur intérêt dans de nombreux domaines d'application. Les ouvrages de McCullagh et Nelder (1989) et de Agresti (2002) fournissent une revue complète sur les modèles linéaires généralisés, leurs méthodes d'estimation et leurs domaines d'application.

2.3.1 Définition

On note y le vecteur de taille N des observations, réalisation du vecteur aléatoire Y, variable à expliquer. Un **modèle linéaire généralisé (GLM, Generalized Linear Model)** est caractérisé par trois hypothèses :

- Les composantes $Y_a,\ a = 1, \ldots, N$ sont supposées indépendantes et identiquement distribuées selon une loi appartenant à la famille exponentielle dont la fonction de densité s'écrit :

$$f(y_a; \theta, \phi) = \exp\Big(\frac{y_a\theta_a - k(\theta_a)}{h_a(\phi)} + c(y_a; \phi)\Big)$$

  où θ_a est un paramètre canonique et ϕ est un paramètre de dispersion. Les fonctions k et c sont spécifiques à chaque distribution et la fonction h_a s'écrit $h_a(\phi) = \phi/\omega_a$ où ω_a est un poids connu associé à la réalisation y_a.

  Sous conditions que la fonction k soit doublement dérivable et de dérivés continues, l'espérance et la variance de la variable associée s'exprime à l'aide des fonctions h_a et k et de leurs dérivées :

$$\mathrm{E}(Y_a) = k'(\theta_a),$$

$$\mathrm{var}(Y_a) = h_a(\phi)k''(\theta_a).$$

  On pose $\mathrm{E}(Y_a) = \mu_a$. Il existe alors une relation directe entre l'espérance et la variance de Y_a :

$$\mathrm{var}(Y_a) = h_a(\phi)k''(k'^{-1}(\mu_a)) = h_a(\phi)v(\mu_a)$$

  où $v = k'' \circ k'^{-1}$ est appelée fonction de variance.

- Comme dans les modèles linéaires classiques, les covariables interviennent linéairement dans la modélisation. On définit ainsi le prédicteur linéaire :

$$\eta = X\beta$$

  où β est un vecteur de paramètres inconnus de taille Q et X est la matrice des covariables de dimension $N \times Q$.

- Le lien à la linéarité (par suite, le lien entre la variable à expliquer et les covariables).

  Le lien entre l'espérance de Y_a et la $a^{ème}$ composante du prédicteur linéaire défini précédemment est établi à l'aide de la fonction monotone et différentiable g :

$$\eta_a = g(\mu_a).$$

  La fonction g est appelée fonction de lien.

Remarques :

1. *Lorsqu'une fonction de lien g permet d'obtenir la relation suivante (égalité du prédicteur linéaire η et du paramètre canonique θ) :*

$$\eta = \theta = X\beta$$

alors ce lien est appelé ***lien canonique****. On montre simplement que la fonction de lien canonique est $g = k'^{-1}$.*

2. *Dans le cadre des modèles linéaires classiques, la fonction de lien canonique associée à la loi gaussienne est l'identité.*

Nous nous intéressons maintenant à l'estimation des paramètres β d'un modèle linéaire généralisé. La partie suivante résume de manière non exhaustive les méthodes d'estimation.

2.3.2 Méthodes d'estimation

Sous l'hypothèse i.i.d. des composantes de Y, la log-vraisemblance du vecteur des paramètres canoniques θ pour les données observées y s'écrit :

$$\log f(y;\theta) = \sum_{a=1}^{N} \log f(y_a;\theta_a) = \sum_{a=1}^{N} \Big[\frac{y_a\theta_a - k(\theta_a)}{\phi/\omega_a} + c(y_a,\phi) \Big]$$

2.3.2.1 Par maximum de vraisemblance

Pour obtenir les équations du maximum de vraisemblance pour l'estimation de β, il faut dériver la log-vraisemblance des données observées par rapport aux différentes composantes du vecteur des paramètres canoniques θ. L'écriture des équations de vraisemblance d'un modèle linéaire généralisé amène, dans le cas général, à des équations non-linéaires en les paramètres β. Une résolution itérative est envisagée par des méthodes générales permettant de résoudre des équations non linéaires et de déterminer le maximum d'une fonction de vraisemblance (Agresti, 2002).

Algorithme de Newton-Raphson

L'algorithme de Newton-Raphson est une méthode itérative pour résoudre des équations non-linéaires. Il repose sur le principe suivant : on se donne une valeur initiale puis on obtient une seconde valeur en approchant la fonction à maximiser dans le voisinage de la valeur initiale par un polynôme du second degré et en trouvant la valeur maximisant ce polynôme. Cela fait appel à la matrice hessienne, matrice des dérivées secondes de la log-vraisemblance (Lange, 2004). Puis on réitère le même procédé en approchant la fonction à maximiser dans le voisinage de la seconde valeur obtenue et ainsi de suite. La méthode génère un ensemble de valeurs.

Algorithme des scores de Fisher

L'algorithme des scores de Fisher est une méthode itérative pour résoudre des équations de vraisemblance. Elle ressemble à l'algorithme de Newton-Raphson, la différence provenant de la matrice hessienne. L'algorithme des scores de Fisher utilise l'espérance de cette matrice, appelée information espérée, tandis que celui de Newton-Raphson utilise la matrice même, appelée information observée.

Dans le cas du GLM, l'algorithme usuel des scores de Fisher itère :

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - \left[\mathrm{E}\left(\frac{\partial^2 \log f(y;\theta)}{\partial\beta\partial\beta'}\right)^{(k)}\right]^{-1} \frac{\partial \log f(y;\theta)}{\partial\beta} \\ &= \beta^{(k)} + (X'W^{-1}_{\beta^{(k)}}X)^{-1}X'W^{-1}_{\beta^{(k)}}\frac{d\eta}{d\mu}^{(k)}(y-\mu^{(k)})\end{aligned}$$

où

$$W_\beta = \mathrm{Diag}\Big\{\mathrm{var}(Y_a)g'(\mu_a)^2\Big\}_{a=1,\ldots,N} = \mathrm{Diag}\Big\{\frac{\phi}{\omega_a}v(\mu_a)g'(\mu_a)^2\Big\}_{a=1,\ldots,N}$$

et

$$\frac{d\eta}{d\mu} = \mathrm{Diag}\Big\{\frac{d\eta_a}{d\mu_a}\Big\}_{a=1,\ldots,N} = \mathrm{Diag}\Big\{g'(\mu_a)\Big\}_{a=1,\ldots,N}.$$

Notons que le lien entre β et θ est décrit par la relation $X\beta = g(k'(\theta))$.

Remarques :

Dans le cas d'un lien canonique,

1. *l'algorithme des scores de Fisher est identique à l'algorithme de Newton-Raphson car $H = -\mathcal{I}$ où H est la matrice hessienne et $\mathcal{I}$ est la matrice d'information de Fisher.*
2. *comme $v(\mu_a) = \dfrac{1}{g'(\mu_a)}$, nous avons*

$$W_\beta = Diag\Big\{\frac{\phi}{\omega_a}g'(\mu_a)\Big\}_{a=1,\ldots,N}.$$

2.3.2.2 Par quasi-vraisemblance

McCullagh et Nelder (1989) consacrent un chapitre de leur livre à l'approche par quasi-vraisemblance. Sous les conditions données par McCullagh et Nelder (1989), on définit le logarithme de la fonction de quasi-vraisemblance, la log-quasi-vraisemblance, pour N individus i.i.d., par :

$$\mathcal{Q}(\mu;y) = \sum_{a=1}^{N}\int_{y_a}^{\mu_a}\frac{y_a - t}{h_a(\phi)v(t)}dt$$

Les équations de quasi-vraisemblance pour estimer les paramètres de régression β sont obtenues en annulant les dérivées de $\mathcal{Q}$. L'ensemble des dérivées est appelé fonction de

quasi-score. Ces équations n'étant pas linéaires, les estimateurs de la quasi-vraisemblance peuvent êtres obtenus par des méthodes itératives telles que l'algorithme des scores de Fisher ou celui de Newton-Raphson. Cette approche permet de contourner l'écriture explicite de la densité de Y en ne s'appuyant que sur les deux premiers moments : l'espérance et la variance.

2.3.3 Propriétés asymptotiques

Dans le cadre général des modèles linéaires généralisés, Fahrmeir et Kaufmann (1985) démontrent différents résultats dont, en particulier, le théorème sur la normalité asymptotique de $\hat{\beta}_N$, solution des équations du maximum de vraisemblance pour un jeu de données de taille N. Ce théorème repose principalement sur des hypothèses concernant les matrices hessiennes et d'information de Fisher. Sous les conditions émises par Fahrmeir et Kaufmann (1985), l'estimateur du maximum de vraisemblance $\hat{\beta}_N$ est asymptotiquement gaussien : $\mathcal{N}(\beta_0, (X'W_{\beta_0}^{-1}X)^{-1})$ où β_0 est la vrai valeur inconnue du paramètre.

2.4 Modèles linéaires mixtes (LMM)

Dans toute expérience, les données présentent une certaine variabilité dont on aimerait pouvoir déterminer la nature, l'importance, les origines ou sources. Lors de ses travaux sur l'analyse de variance et sur le coefficient de corrélation inter-classe, Ronald Fisher a tenté de séparer les différentes sources de variation, et de répondre notamment à des questions sur la significativité de différences observées entre les moyennes de sous-groupes de données.

Les modèles à effets aléatoires constituent un outil puissant pour étudier la variabilité dans une expérience. Dans cette modélisation, les effets aléatoires sont introduits en complément aux effets fixes. Il est important de bien distinguer la notion d'effet fixe et d'effet aléatoire. On distingue en effet deux types de facteurs :

- ▷ le facteur à effets fixes. Tous les niveaux de ce facteur sont observés et on veut mesurer l'effet de chaque niveau sur la variable à expliquer.

- ▷ le facteur à effets aléatoires. Tous les niveaux de ce facteur ne sont pas observés. Seulement un échantillon de ces niveaux est alors représenté. Ici, ce n'est pas l'influence de chacun des niveaux sur la variable à expliquer qui nous intéresse mais l'effet global de ce facteur.

L'introduction d'effets aléatoires permet d'être plus précis sur l'origine de la variabilité totale par rapport à la modélisation statistique sans effets aléatoires. En effet, cette variabilité se divise en deux parties : le variabilité due aux effets aléatoires et la variabilité due aux erreurs. On parle de composantes de la variance. Searle et al. (1992) consacrent leur

ouvrage à la décomposition de la variabilité. La suite de cette partie s'appuiera également sur les ouvrages très pédagogiques de Verbeke et Molenberghs (2000), Diggle et al. (2002) et Fitzmaurice et al. (2004) traitant de l'analyse de données longitudinales.

2.4.1 Exemple introductif

Un exemple inspiré des exemples fournis par Diggle et al. (2002) permet de rendre les notions d'effet fixe et d'effet aléatoire plus intuitives.

Le lait des vaches béarnaises

Supposons que l'on s'intéresse à la teneur en protéines du lait. On dispose pour cela d'un échantillon de 78 vaches béarnaises dont le lait est collecté et analysé toutes les semaines. Ces animaux sont répartis de manière égale et aléatoire en 3 groupes qui subiront respectivement les régimes alimentaires suivants : que de l'orge, un mélange d'orge et de lupin ou que du lupin. On relève toutes les semaines pendant 19 semaines la teneur en protéines du lait. Nous avons donc au final un jeu de 78*19 = 1482 données. On parle également de données longitudinales, composées de 78 séries de longueur 19. Rappelons que l'objectif de cette étude est de déterminer comment le régime alimentaire affecte la teneur en protéines du lait. Ainsi, chaque niveau du facteur "régime alimentaire" parait important et on aimerait en mesurer l'effet. Ce facteur à niveau fini (3) est alors considéré comme facteur à effet fixe. Cependant, un autre facteur peut avoir de l'influence sur la teneur en protéines du lait : la vache concernée. Les 78 vaches, choisies de manière aléatoire, ne sont qu'un échantillon de l'ensemble des vaches béarnaises. Ce ne sera donc pas l'effet de chaque vache qui aura de l'importance mais la variabilité des données induite par ces individus. Le facteur "vache béarnaise" à 78 niveaux est alors considéré comme effet aléatoire et représentera une des composantes de le variabilité totale des données. Il existe une dépendance entre les 19 mesures faites pour chaque vache.

Dans cet exemple, la nature de chacun des facteurs semble évidente. Mais, dans la pratique, la modélisation des différents effets ainsi que la distinction entre effets fixes et effets aléatoires peut être plus compliquée.

2.4.2 Définition

Les modèles linéaires mixtes (LMM, Linear Mixed Model), introduits en 1959 pour l'analyse de données de génétique animale par Henderson, sont en fait une extension des modèles linéaires classiques où viennent s'ajouter aux effets fixes des effets aléatoires. Actuellement, le modèle linéaire mixte a des applications dans de nombreux domaines, notamment, en économie, en biologie, en agronomie et en médecine (Verbeke et Molenberghs, 2000; McCulloch et al., 2008). Les modèles linéaires mixtes se formalisent de la

manière suivante :

$$Y = \underbrace{X\beta}_{\text{partie effets fixes}} + \underbrace{U\xi}_{\text{partie effets aléatoires}} + \underbrace{\epsilon}_{\text{partie résiduelle}}$$

où

- Y est le vecteur aléatoire à expliquer de taille N,
- β est le vecteur des paramètres inconnus des effets fixes de taille Q,
- X, de dimension $N \times Q$ est la matrice d'incidence de β. Elle est supposée fixe et connue et est usuellement appelée matrice des covariables.
- ξ est le vecteur d'effets aléatoires de taille q. En toute généralité, ce vecteur se décompose en K parties $\xi = (\xi_1', \ldots, \xi_K')'$ où K est le nombre d'effets aléatoires considérés dans le modèle. Chaque composante ξ_j est un vecteur aléatoire de dimension q_j. Il est constitué des q_j réalisations du $j^{ème}$ effet aléatoire, observées au sein des données ($\sum_{j=1}^{K} q_j = q$).
  Dans l'exemple introductif de la section précédente, nous n'avons introduit qu'un seul effet aléatoire ($K = 1$) : effet "vache béarnaise". Il sera modélisé par un vecteur de taille 78 où on affectera une réalisation de l'effet à chaque vache. Cela introduit une dépendance entre les 19 mesures relevées sur chaque vache.

  On suppose en général une distribution gaussienne centrée des effets aléatoires, c'est-à-dire : $\forall j \in \{1, \ldots, K\}, \quad \xi_j \sim \mathcal{N}_{q_j}(0, \tau_j^2 A_j)$ avec A_j matrice de dimension $q_j \times q_j$ supposée connue. D'autre part, $\forall i, j \in \{1, \ldots, K\}^2, i \neq j$, ξ_i et ξ_j sont indépendants. Donc $\xi \sim \mathcal{N}_N(0, D)$ où D est une matrice diagonale par blocs $D = \text{Diag}\{\tau_j^2 A_j\}_{j=1,\ldots,K}$.
- U, de dimension $N \times q$, est la matrice d'incidence associé à ξ. Elle est connue et formée des matrices d'incidence U_j de dimension $N \times q_j$ de chaque effet aléatoire : $U = [U_1 \vdots \cdots \vdots U_K]$. Elle se compose le plus souvent de 0 et de 1.
- ϵ est le vecteur aléatoire d'erreurs de taille N. Puisque le modèle considéré est linéaire, la distribution de ϵ est $\epsilon \sim \mathcal{N}_N(0, \sigma^2 V_0)$. On notera aussi $R = \sigma^2 V_0$. On suppose que $\forall j \in \{1, \ldots, K\}$, ϵ et ξ_j sont indépendants.

Sous ces différentes hypothèses, on a :

- l'espérance et la variance de Y conditionnellement aux effets aléatoires ξ

$$\begin{aligned} \text{E}(Y|\xi) &= X\beta + U\xi, \\ \text{var}(Y|\xi) &= R = \sigma^2 V_0. \end{aligned}$$

- l'espérance et la variance de Y (loi marginale)

$$\text{E}(Y) = X\beta,$$

$$\begin{aligned} \mathrm{var}(Y) &= R + UDU' \\ &= \sigma^2 V_0 + \sum_{j=1}^{K} \tau_j^2 U_j A_j U_j' \\ &= \Gamma. \end{aligned}$$

À partir des propriétés de conditionnement de la loi gaussienne, nous pouvons déduire les distributions suivantes :

- $Y \sim \mathcal{N}_N(X\beta, \Gamma)$,
- $Y|\xi \sim \mathcal{N}_N(X\beta + U\xi, R)$,
- $\begin{pmatrix} Y \\ \xi \end{pmatrix} \sim \mathcal{N}_{N+q}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} R + UDU' & UD \\ DU' & D \end{pmatrix} \right)$,
- $\xi|Y \sim \mathcal{N}_q(DU'\Gamma^{-1}(y - X\beta), D - DU'\Gamma^{-1}UD)$.

Dans les modèles linéaires mixtes, le vecteur des effets fixes β ainsi que le vecteur des $K+1$ paramètres de variance $(\sigma^2, \tau_1^2, \ldots, \tau_K^2)'$ sont des paramètres inconnus à estimer. Nous nous intéressons également aux effets aléatoires ξ dont les réalisations sont indirectement observées dans les données.

2.4.3 Méthodes d'estimation

Sous l'hypothèse i.i.d. des composantes de Y, la log-vraisemblance du vecteur des paramètres θ pour les données observées s'écrit :

$$\log f(y;\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\Gamma| - \frac{1}{2}(y - X\beta)'\Gamma^{-1}(y - X\beta),$$

où $|\Gamma|$ représente le déterminant de la matrice Γ.

2.4.3.1 Par maximum de vraisemblance (ML)

En cherchant les valeurs qui annulent la dérivée de la log-vraisemblance des données Y, on obtient un système d'équations non-linéaires en les paramètres de variance. Ces équations sont résolues à l'aide d'un algorithme itératif. À partir de valeurs initiales de τ^2 et σ^2, on itére la résolution des $K+1$ équations relatives aux composantes de la variance jusqu'à convergence, en résolvant le système linéaire à chaque étape. Puis à l'aide des valeurs alors obtenues, on résout l'équation en les effets fixes pour trouver l'estimation de β.

2.4.3.2 Par maximum de vraisemblance restreint (REML)

La méthode proposée par Patterson et Thompson (1971) est proche de l'approche par maximum de vraisemblance. En effet, dans un premier temps, on supprime provisoirement les effets fixes en projetant le modèle sur l'orthogonal du sous-espace vectoriel engendré

par les colonnes de X. Puis, on ne maximise que la partie de la vraisemblance concernant les composantes de la variance. L'estimation par maximum de vraisemblance restreint est en fait l'estimation par maximum de vraisemblance dans le modèle projeté. Après une estimation itérative des composantes de la variance, on estime directement les effets fixes β. 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.

2.4.3.3 Par la méthode de Henderson

Proposée par Henderson et al. (1959), cette approche diffère des méthodes présentées précédemment. En effet, les réalisations des effets aléatoires ξ sont prédites contrairement à précédemment où on ne s'intéressait qu'à leur loi. Un exemple de génétique animale (Trottier, 1998) permet de justifier l'intérêt de ces prédictions. D'autres exemples sont envisageables. Lors d'un processus de sélection d'arbre fruitier, par exemple, on cherche à déterminer, au vu du rendement fruitier de ses descendants, le géniteur idéal pour la prochaine descendance. Pour cela, on est amené à prédire des réalisations non observées de l'effet aléatoire (caractérisant l'effet des parents) à l'intérieur d'un modèle mixte.
Le système d'équations de Henderson est obtenue après maximisation de la log-vraisemblance de la loi jointe de Y et ξ. Pour obtenir les estimateurs ML et REML, Harville (1977) a proposé un schéma itératif alternant, pour des valeurs de τ_k^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.

2.4.4 Algorithme EM pour les LMM

L'algorithme EM (section 2.1) est principalement utilisé pour traiter des problèmes aux données incomplètes. Le vecteur des effets aléatoires étant non observé (on parle de structure cachée), nous pouvons supposer que notre problématique est en adéquation avec le type de problème que permet de résoudre l'algorithme EM. Le vecteur des données complètes devient alors (Y, ξ). Nous présentons ici rapidement l'algorithme EM pour les LMM. Cette présentation s'appuie sur la thèse de Trottier (1998) et sur le tutoriel de Foulley (2002).

La log-vraisemblance des données complètes s'écrit :

$$\log f(\xi, y; \theta) \quad = \quad -\frac{1}{2}(\sum_{j=0}^{K} q_j) \log 2\pi - \frac{1}{2}\sum_{j=0}^{K}(q_j \log \tau_j^2 + \log |A_j|) - \frac{1}{2}\sum_{j=0}^{K} \frac{\xi_j' A_j^{-1} \xi_j}{\tau_j^2}$$

avec $\tau_0^2 = \sigma^2$, $q_0 = N$, $A_0 = V_0$ et $\xi_0 = \epsilon = y - X\beta - U\xi$.

L'étape E de l'algorithme consiste à calculer l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées. Dans le cas des LMM, l'étape E consiste à remplacer les statistiques exhaustives $\xi_j' A_j^{-1} \xi_j$ et $y - U\xi$ par leurs

espérances conditionnelles sachant y. On utilise pour cela le conditionnement des variables aléatoires gaussiennes. Pour $j = 0, \ldots, K$, on a :

$$\begin{aligned}
\mathrm{E}(\xi_j|Y=y) &= \tau_j^2 A_j U_j' \Gamma^{-1}(y - X\beta), \\
\mathrm{E}(\xi_j' A j^{-1} \xi_j|Y=y) &= \tau_j^4 (y-X\beta)'\Gamma^{-1} V_j \Gamma^{-1}(y-X\beta) + \mathrm{tr}(\tau_j^2 Id_{q_j} - \tau_j^4 U_j' \Gamma^{-1} U_j A_j) \\
&= \tau_j^4 (y-X\beta)'\Gamma^{-1} V_j \Gamma^{-1}(y-X\beta) + q_j \tau_j^2 - \tau_j^4 \mathrm{tr}(\Gamma^{-1} V_j), \\
\mathrm{E}(Y - U\xi|Y=y) &= y - \sum_{j=1}^{K} \tau_j^2 V_j \Gamma^{-1}(y - X\beta) \\
&= y - (\Gamma - \sigma^2 V_0)\Gamma^{-1}(y - X\beta) \\
&= X\beta + \sigma^2 V_0 \Gamma^{-1}(y - X\beta).
\end{aligned}$$

Ainsi, à l'étape M de l'algorithme EM, les nouvelles valeurs des paramètres à l'itération k sont données par :

$$\begin{aligned}
q_j \tau_j^{2(k+1)} &= \tau_j^{4(k)} (y - X\beta^{(k)})' \Gamma^{-1(k)} V_j \Gamma^{-1(k)} (y - X\beta^{(k)}) + q_j \tau_j^{2(k)} - \tau_j^{4(k)} \mathrm{tr}(\Gamma^{-1(k)} V_j), \\
X\beta^{(k+1)} &= X(X' V_0^{-1} X)^{-1} X' V_0^{-1} (X\beta^{(k)} + \sigma^{2(k)} V_0 \Gamma^{-1(k)} (y - X\beta^{(k)})).
\end{aligned}$$

2.4.5 Propriétés asymptotiques

Rao et Kleffe (1988) présentent les différentes conditions nécessaires pour les propriétés asymptotiques des modèles linéaires mixtes. Sous condition de régularité concernant la log-vraisemblance des données observées en le vecteur de paramètre θ, les propriétés de convergence presque sûre et de normalité asymptotique de l'estimateur du maximum de vraisemblance peuvent être établies. Elles reposent sur des conditions de régularité de la fonction de vraisemblance. Searle et al. (1992) donnent la matrice d'information de Fisher et la matrice de variances asymptotiques des estimateurs.

Chapitre 3

Combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés (MS-GLM/SMS-GLM)

Ce chapitre est consacré à la modélisation statistique de données de type séquence ou série chronologique présentant les caractéristiques suivantes :

- plusieurs individus sont étudiés simultanément,
- les données observées sont structurées en phases successives, asynchrones entre individus,
- les données observées sont influencées par des covariables pouvant varier dans le temps et pouvant être communes aux individus.

Ce type de données est notamment illustré par la croissance de plantes. Cette croissance est constituée d'une succession de phases liées à la morphogénèse : phase d'établissement, phase adulte et phase de sénescence par exemple (Véra, 2004). Le changement de phase de croissance est spécifique à chaque plante, d'où l'asynchronisme. Par ailleurs, la croissance d'une plante est affectée par des facteurs environnementaux, notamment climatiques, pouvant avoir une influence plus ou moins forte selon le stade de développement de la plante. Des applications dans le domaine médical permettent également d'illustrer ce type de données.

Dans un premier temps, nous présenterons la famille des combinaisons markoviennes de modèles linéaires généralisés et des combinaisons semi-markoviennes de modèles linéaires généralisés. Dans une deuxième partie, nous discuterons des méthodes d'estimation des paramètres de ces modèles. Dans une troisième partie, nous traiterons le cas de ces familles pour données observées poissonniennes. Des simulations et une application médicale illustreront l'intérêt de ces modèles. Dans une dernière partie, nous présenterons ces familles dans le cadre de données observées binaires pour lequel nous donnerons des résultats de simulations et une application en botanique.

3.1 Définitions

Nous avons vu dans la section 2.2.2 qu'une chaîne de Markov cachée gaussienne pouvait être définie comme une paire de processus stochastiques $\{S_t, Y_t\}$ où le processus d'observation $\{Y_t\}$ est lié au processus caché $\{S_t\}$ par la distribution gaussienne $Y_t|_{S_t=s_t} \sim \mathcal{N}(\mu_{s_t}, \sigma^2_{s_t})$. Lindgren (1978) a introduit la famille des combinaisons markoviennes de modèles linéaires (MS-LM, Markov Switching Linear Model) qui étend la famille des chaînes de Markov cachées gaussiennes en incorporant l'influence de covariables comme effets fixes dans le processus d'observation. Les combinaisons markoviennes de modèles linéaires ont depuis été utilisées dans de nombreux domaines tels qu'en économie ou pour l'analyse de réseaux de gènes en biologie (Gupta et al., 2007). L'ouvrage de Frühwirth-Schnatter (2006), ouvrage de référence sur les mélanges finis et les combinaisons markoviennes, fournit de nombreux exemples d'application de ces modèles. Turner et al. (1998) et Wang et Puterman (1999) ont introduit, à la fin des années 1990, des modèles de type markovien caché permettant de traiter des données dont la distribution du processus d'observation est une loi de Poisson. L'influence de covariables est également prise en compte dans le processus d'observation. Cependant, le modèle proposé par Wang et Puterman (1999) ne modélise qu'un seul individu. Si le jeu de données comporte plusieurs individus, un modèle différent est associé à chaque individu et ils sont modélisés indépendamment les uns des autres. Turner et al. (1998) ont appliqué leur modèle sur des données des comptages de coliformes[1] fécaux dans l'eau de mer sur différents sites et pour différentes profondeurs. Wang et Puterman (1999) ont appliqué leur modèle pour analyser les effets de la gammaglobuline[2] administrée par voie intraveineuse sur le nombre quotidien de crises d'épilepsie.

Russell (1993) a introduit les combinaisons semi-markoviennes de modèles linéaires (SMS-LM, Semi-Markov Switching Linear Model) qui étend la famille des combinaisons markoviennes de modèles linéaires au cas semi-markovien. Il a appliqué ces modèles pour la reconnaissance de la parole. Nous n'avons trouvé dans la littérature aucun travail consacré aux combinaisons semi-markoviennes de modèles linéaires généralisés.

Nous allons ici formaliser les combinaisons markoviennes de modèles linéaires généralisés. Ces modèles généralisent les modèles introduits par Wang et Puterman (1999) à plusieurs individus dont la distribution du processus d'observation appartient à la famille exponentielle et prend en compte l'influence de covariables. Afin d'illustrer ces idées, nous pouvons revenir sur l'exemple de la modélisation de la croissance des plantes. La chaîne de Markov sous-jacente représente la succession de phases de croissance tandis que le modèle linéaire généralisé associé à chaque état modélise dans la phase de croissance correspondante l'influence des covariables sur les données observées.

[1] Le coliforme est une entérobactérie fermentant le lactose à 30°C avec production de gaz.
[2] Substance protéique du sang contenant des anticorps.

Notations :

Par la suite :

- a désigne l'individu et N est le nombre d'individus,
- t désigne le temps,
- T_a est la longueur de la séquence observée relative à l'individu a et $\sum_{a=1}^{N} T_a = T$,
- y est le vecteur de taille T des observations, réalisation du vecteur aléatoire Y,
- s est le vecteur de taille T des états non-observables, réalisation du vecteur aléatoire S,
- y_{at} est l'observation relative à l'individu a au temps t,
- s_{at} est l'état relatif à l'individu a au temps t,
- $Y_{a1}^{T_a} = y_{a1}^{T_a}$ est la suite des variables observées et de leurs réalisations $Y_{a1} = y_{a1}, Y_{a2} = y_{a2}, \ldots, Y_{aT_a} = y_{aT_a}$ relatives à l'individu a,
- $S_{a1}^{T_a} = s_{a1}^{T_a}$ est la suite des variables cachées et de leurs réalisations $S_{a1} = s_{a1}, S_{a2} = s_{a2}, \ldots, S_{aT_a} = s_{aT_a}$ relatives à l'individu a.

Définition 3.1 *Une **combinaison markovienne de modèles linéaires généralisés** (MS-GLM, Markov Switching Generalized Linear Model) se caractérise par un couple de processus stochastiques $\{S_{at}, Y_{at}; a = 1, \ldots, N,\ t = 1, \ldots, T_a\}$ combinant :*

- *une chaîne de Markov sous-jacente $\{S_{at}, t = 1, \ldots, T_a\}$ d'ordre1, homogène dans le temps et à valeurs dans l'espace d'états fini $\{1, \ldots, J\}$,*
- *un processus d'observation $\{Y_{at}, t = 1, \ldots, T_a\}$ pour chaque individu a. Chaque observation y_{at} est liée au processus d'état S_{at} par un modèle linéaire généralisé.*

Les combinaisons markoviennes de modèles linéaires généralisés peuvent être vues comme des mélanges finis de modèles linéaires généralisés avec dépendances markoviennes. Comme l'influence des covariables est uniquement prise en compte dans le processus d'observation, nous pouvons étendre les MS-GLM au cas semi-markovien. Ces modèles sont alors appelés combinaisons semi-markoviennes de modèles linéaires généralisés. Les combinaisons semi-markoviennes de modèles linéaires généralisés peuvent être vues comme des mélanges finis de modèles linéaires généralisés avec dépendances semi-markoviennes.

Définition 3.2 *Une **combinaison semi-markovienne de modèles linéaires généralisés** (SMS-GLM, Semi-Markov Switching Generalized Linear Model) se caractérise par un couple de processus stochastiques $\{S_{at}, Y_{at}; a = 1, \ldots, N,\ t = 1, \ldots, T_a\}$ combinant :*

- *une semi-chaîne de Markov sous-jacente $\{S_{at}, t = 1, \ldots, T_a\}$ homogène dans le temps et à valeurs dans l'espace d'états fini $\{1, \ldots, J\}$,*
- *un processus d'observation $\{Y_{at}, t = 1, \ldots, T_a\}$ pour chaque individu a. Chaque observation y_{at} est liée au processus d'état S_{at} par un modèle linéaire généralisé.*

Nous définissons à présent les modèles linéaires généralisés qui lient les processus d'observation et les processus d'état. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation Y_{at} d'un individu a au temps t est modélisée par le modèle linéaire généralisé suivant (section 2.3) :

$$b_{s_{at}}(y_{at}) = f(y_{at}|S_{at} = s_{at}; \theta_{ats_{at}}, \phi) = \exp\left\{\frac{y_{at}\theta_{ats_{at}} - k(\theta_{ats_{at}})}{h_{at}(\phi)} + c(y_{at}; \phi)\right\}$$

où ϕ est un paramètre de dispersion et $\theta_{ats_{at}}$ est un paramètre canonique pour l'individu a dans l'état $S_{at} = s_{at}$ au temps t.

Conditionnellement à l'état $S_{at} = s_{at}$, l'espérance et la variance de la variable associée s'écrit à l'aide des fonctions h_{at} et k :

$$\begin{aligned} \mathrm{E}(Y_{at}|S_{at} = s_{at}) &= \mu_{ats_{at}} = k'(\theta_{ats_{at}}), \\ \mathrm{var}(Y_{at}|S_{at} = s_{at}) &= k''(\theta_{ats_{at}})h_{at}(\phi). \end{aligned}$$

Le lien entre la variable à expliquer Y_{at} et les covariables X_{at} est établi à l'aide de la fonction de lien g :

$$\eta_{ats_{at}} = g(\mu_{ats_{at}}) = g(k'(\theta_{ats_{at}})) = X_{at}\beta_{s_{at}}$$

où sur l'état $S_{at} = s_{at}$, $\beta_{s_{at}}$ est un vecteur de paramètres inconnus de taille Q et X_{at} est le vecteur ligne des covariables de taille Q pour l'individu a au temps t.

Les MS-GLM et les SMS-GLM ne diffèrent que dans le type du processus caché sous-jacent : une chaîne de Markov pour le premier et une semi-chaîne de Markov pour le second. Aussi, nous traiterons de manière détaillé le cas des combinaisons markoviennes de modèles linéaires généralisés. La transposition au SMS-GLM est directe ; nous la présenterons brièvement par la suite.

Rappelons quelques-unes des hypothèses faites sur les combinaisons markoviennes de modèles linéaires généralisés :

1. Les individus et par suite les séquences d'états sous-jacentes sont indépendantes : $\forall a \neq a';\ \mathrm{cov}(Y_{a1}^{T_a}, Y_{a'1}^{T_{a'}}) = 0_{T_a \times T_{a'}}$ et $\mathrm{cov}(S_{a1}^{T_a}, S_{a'1}^{T_{a'}}) = 0_{T_a \times T_{a'}}$.
2. Conditionnellement aux états, les observations pour un même individu a sont indépendantes : $\forall t \neq t';\ \mathrm{cov}(Y_{at}, Y_{at'}|S_{at} = s_{at}, S_{at'} = s_{at'}) = 0$.

3.2 Méthodes d'estimation

Les paramètres du MS-GLM peuvent être scindés en deux catégories : les paramètres $(\pi_j; j = 1, \ldots, J)$ et $(p_{ij}; i, j = 1, \ldots, J)$ de la chaîne de Markov sous-jacente et les

paramètres $(\beta_j; j = 1, \ldots, J)$ des J modèles linéaires généralisés. Nous dénotons par $\theta = (\pi, P, \beta)$, l'ensemble des paramètres à estimer.

Nous avons vu dans la section 2.2.4 qu'il existait divers algorithmes pour estimer les paramètres d'une chaîne de Markov cachée classique où le lien entre processus d'observation et processus d'état est paramétré par de simples lois d'observation qui ne prennent pas en compte d'éventuelles covariables. Nous nous intéressons ici, comme pour les chaînes de Markov cachées classiques, aux approches basées sur l'algorithme EM.

Lindgren (1978) et Cosslett et Lee (1985) estiment les paramètres des combinaisons markoviennes de modèles linéaires à l'aide de l'algorithme EM pour chaînes de Markov cachées classiques. L'étape E de l'algorithme EM est implémentée par l'algorithme "avant-arrière" dont les sorties sont les probabilités *a posteriori* résumant l'action de tel ou tel paramètre du modèle sur l'ensemble des séquences d'états possibles sachant la séquence observée. Archer et Titterington (2002) parle de restauration probabiliste des séquences d'états possibles. L'étape M repose sur la maximisation directe de l'espérance de la log-vraisemblance des données complètes sachant les données observées. Du fait de la linéarité des processus d'observation sur chaque état, cette maximisation est facile. Vous pouvez trouver le détail des calculs dans la thèse de Véra (2004). Nous pouvons cependant noter que d'autres approches sont envisageables. Chopin et Pelgrin (2004) proposent par exemple d'utiliser une approche bayésienne pour estimer les paramètres des combinaisons markoviennes de modèles linéaires.

Afin d'estimer les paramètres des combinaisons markoviennes de modèles linéaires généralisés, Turner et al. (1998) proposent une méthode basée sur l'algorithme EM sous la condition d'équilibre de la chaîne de Markov sous-jacente. Le processus d'observation est vu comme un modèle linéaire généralisé pondéré par la probabilité d'être dans un état à un temps donné. Ils supposent que l'influence des covariables est la même sur chacun des états et introduisent l'effet de l'état comme étant un effet fixe. L'estimation des paramètres liés aux processus d'observation se fait à l'aide de la procédure glm() de S-PLUS[3]. Les données doivent être structurées comme suit : chaque vecteur d'observation est répété J fois, une fois pour chaque état. Puis la probabilité d'être dans l'état j à un temps donné et le facteur d'état j sont ajoutés à la $j^{ème}$ répétition. Cette approche peut s'avérer vite lourde et coûteuse si le nombre d'états est élevé, si l'influence des covariables est supposée différente d'une phase à l'autre ou si la chaîne de Markov sous-jacente n'est pas en équilibre. De plus, l'utilisation de la procédure glm() de S-PLUS ne permet pas de traiter tous les cas et notamment celui des données catégorielles.

Wang et Puterman (1999) proposent d'estimer les paramètres du MS-GLM en utilisant une approche basée sur l'algorithme EM. L'étape E de l'algorithme EM est implémentée par l'algorithme "avant-arrière" (Baum et al., 1970). La maximisation à l'étape

[3]http ://www.insightful.com/products/splus/default.asp

M se fait à l'aide de méthodes de quasi-Newton (Lange, 2004). Les travaux de Wang et Puterman (1999) reposent sur l'hypothèse d'une unique séquence observée (i.e. d'un seul individu). Dans le cas de plusieurs individus observés, Wang et Puterman (1999) les traitent indépendamment avec un processus caché propre à chaque individu ; c'est-à-dire que $\forall a \neq a',\ P(S_{a1} = j) \neq P(S_{a'1} = j)$ et $P(S_{at} = j|S_{a,t-1} = i) \neq P(S_{a't} = j|S_{a',t-1} = i)$. Il y a donc une MS-GLM associée à chaque individu.

L'objectif de ce travail est de proposer des algorithmes d'inférence sous la contrainte qu'ils tiennent compte de la paramétrisation choisie, qu'ils ne soient pas contraints par le type de structure sous-jacente (ergodique ou pas, équilibre ou pas, Markov ou semi-Markov), par le type des covariables et par le type des variables réponses et qu'ils soient stables numériquement.

3.2.1 Formalisme de l'algorithme du gradient EM pour MS-GLM

Dans le but d'estimer les paramètres des combinaisons markoviennes de modèles linéaires généralisés, nous proposons un algorithme itératif, fondé sur l'algorithme du gradient EM (section 2.1.3.1), en deux étapes que nous détaillerons par la suite :

- **Étape E :**
  - une restauration probabiliste de toutes les séquences d'états à partir de l'algorithme "avant-arrière",
- **Étape M :**
  - une ré-estimation des paramètres de la chaîne de Markov sous jacente (probabilités initiales et probabilités de transition) par maximisation directe de l'espérance de la log-vraisemblance des données complètes sachant les données observées,
  - une ré-estimation itérative des paramètres des modèles linéaires généralisés par un algorithme des scores de Fisher.

Considérons la log-vraisemblance des données complètes où à la fois les observations y et les états s des chaînes de Markov sous-jacentes sont observées :

$$\begin{aligned}
\log f(s,y;\theta) &= \sum_{a=1}^{N} \log f(s_{a1}^{T_a}, y_{a1}^{T_a}; \theta) \\
&= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1} = j\Big)\log \pi_j + \sum_{a=1}^{N}\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} I\Big(s_{at} = j, s_{a,t-1} = i\Big)\log p_{ij} \\
&+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\Big(s_{at} = j\Big)\log b_j(y_{at}) \\
&= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1} = j\Big)\log \pi_j + \sum_{a=1}^{N}\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} I\Big(s_{at} = j, s_{a,t-1} = i\Big)\log p_{ij} \\
&+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\Big(s_{at} = j\Big)\Big(\frac{y_{at}\theta_{atj} - k(\theta_{atj})}{h_{at}(\phi)} + c(y_{at};\phi)\Big). \qquad (3.1)
\end{aligned}$$

Étape E :

On s'intéresse à l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées (cf équation (2.12) pour les chaînes de Markov cachées classiques) qui vaut dans le cas du MS-GLM :

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log f(s,y;\theta) \mid Y = y;\theta^{(k)}\Big) \\ Q(\theta|\theta^{(k)}) &= \sum_{a=1}^{N}\sum_{j=1}^{J} L_{aj}(1)\log\pi_j + \sum_{a=1}^{N}\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} \mathrm{P}(S_{at}=j, S_{a,t-1}=i|Y_{a1}^{T}=y_{a1}^{T};\theta^{(k)})\log p_{ij} \\ &+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} L_{aj}(t)\Big(\frac{y_{at}\theta_{atj}-k(\theta_{atj})}{h_{at}(\phi)} + c(y_{at};\phi)\Big) \end{aligned} \tag{3.2}$$

avec les probabilités lissées $L_{aj}(t) = \mathrm{P}(S_{at}=j|Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)})$.

L'étape E consiste en une restauration probabiliste de toutes les séquences d'états possibles pour chaque individu a à partir des probabilités lissées $L_{aj}(t)$. Ces probabilités sont calculées récursivement par l'algorithme "avant-arrière" décrit à la section 2.2.4.2 pour les chaînes de Markov cachées.

Étape M :

Estimation des paramètres liés aux chaînes de Markov cachées π_j et p_{ij}

On obtient après maximisation de $Q(\theta|\theta^{(k)})$ (équation (3.2)) les estimations des paramètres associés à la chaîne de Markov sous-jacente. Ces estimations généralisent à N individus indépendants les formules de ré-estimations obtenues pour les chaînes de Markov cachées classiques. Ceci nous conduit aux formules de ré-estimation à l'itération k :

– probabilités initiales

$$\pi_j^{(k+1)} = \frac{\sum_a L_{aj}^{(k)}(1)}{N},$$

– probabilités de transition

$$p_{ij} = \frac{\sum_a \sum_{t=1}^{T_a-1} L_{aj}^{(k)}(t+1) p_{ij}^{(k)} F_{ai}^{(k)}(t)/G_{aj}^{(k)}(t+1)}{\sum_a \sum_{t=1}^{T_a-1} L_{ai}^{(k)}(t)},$$

où $F_{ai}^{(k)}(t) = P(S_{at}=i|Y_{a1}^{t}=y_{a1}^{t};\theta^{(k)})$ est la probabilité filtrée et $G_{aj}^{(k)}(t+1) = P(S_{a,t+1}=j|Y_{a1}^{t}=y_{a1}^{t};\theta^{(k)})$ est la probabilité prédite pour l'individu a au temps t à l'itération k.

Estimation des paramètres liés aux modèles linéaires généralisés β_j

Avant de calculer la dérivée première de $Q(\theta|\theta^{(k)})$ par rapport à β_j, nous calculons :

$$\begin{aligned} \frac{\partial \frac{y_{at}\theta_{atj}-k(\theta_{atj})}{h_{at}(\phi)}}{\partial\beta_j} &= \frac{\partial\eta_{atj}}{\partial\beta_j}\frac{\partial\mu_{atj}}{\partial\eta_{atj}}\frac{\partial\theta_{atj}}{\partial\mu_{atj}}\frac{\partial \frac{y_{at}\theta_{atj}-k(\theta_{atj})}{h_{at}(\phi)}}{\partial\theta_{atj}} \\ &= X'_{at}\frac{1}{g'(\mu_{atj})}\frac{1}{k''(\mu_{atj})}\frac{y_{at}-k'(\theta_{atj})}{h_{at}(\phi)} \end{aligned}$$

$$
\begin{aligned}
&= X'_{at}\frac{1}{g'(\mu_{atj})}\frac{h_{at}(\phi)}{\mathrm{var}(Y_{at}|S_{at}=j)}\frac{y_{at}-\mu_{atj}}{h_{at}(\phi)} \\
&= X'_{at}\frac{1}{g'(\mu_{atj})^2\mathrm{var}(Y_{at}|S_{at}=j)}g'(\mu_{atj})\ (y_{at}-\mu_{atj}).
\end{aligned}
$$

Par suite, la dérivée première de $Q(\theta|\theta^{(k)})$ par rapport à β_j est :

$$
\begin{aligned}
\frac{\partial Q(\theta|\theta^{(k)})}{\partial\beta_j} &= \sum_{a=1}^{N}\sum_{t=1}^{T_a} L_{aj}(t)\frac{\partial\frac{y_{at}\theta_{atj}-k(\theta_{atj})}{h_{at}(\phi)}}{\partial\beta_j} \\
&= \sum_{a=1}^{N}\sum_{t=1}^{T_a} L_{aj}(t)X'_{at}\frac{1}{g'(\mu_{atj})^2\mathrm{var}(Y_{at}|S_{at}=j)}g'(\mu_{atj})\ (y_{at}-\mu_{atj}) \qquad (3.3)
\end{aligned}
$$

Posons

$$
\begin{aligned}
W_{aj} &= \mathrm{Diag}\Big\{\mathrm{var}(Y_{at}|S_{at}=j)g'(\mu_{atj})^2\Big\}_{t=1,\ldots,T_a}, \\
\frac{\partial\eta_{aj}}{\partial\mu_{aj}} &= \mathrm{Diag}\Big\{\frac{\partial\eta_{atj}}{\partial\mu_{atj}}\Big\}_{t=1,\ldots,T_a} = \mathrm{Diag}\Big\{g'(\mu_{atj})\Big\}_{t=1,\ldots,T_a}, \\
L_{aj} &= \mathrm{Diag}\Big\{L_{aj}(t)\Big\}_{t=1,\ldots,T_a}, \\
\mu_{aj} &= \big(\mu_{atj}\big)_{t=1,\ldots,T_a}.
\end{aligned}
$$

A partir de l'équation (3.3), nous pouvons écrire sous forme matricielle les équations du maximum de vraisemblance pour β_j :

$$
\sum_{a=1}^{N} X'_a L_{aj} W_{aj}^{-1}\frac{\partial\eta_{aj}}{\partial\mu_{aj}}(y_{a1}^{T_a}-\mu_{aj}) = 0
$$

où X_a est la matrice des covariables de dimension $T_a \times Q$.

Ce système d'équations n'est pas linéaire en β_j, qui intervient à la fois dans les matrices W_{aj}, $\frac{\partial\eta_{aj}}{\partial\mu_{aj}}$ et dans le vecteur μ_{aj}. C'est pourquoi, il est envisagé une résolution itérative. Comme expliqué par Titterington (1984) et Foulley et al. (2000), il est possible d'appliquer l'algorithme des scores de Fisher à $Q(\theta|\theta^{(k)})$ dont l'itération k s'écrit :

$$
\begin{aligned}
\beta_j^{(k+1)} &= \beta_j^{(k)} - \left[\mathrm{E}\Big(\frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\beta_j\partial\beta'_j}\Big)\right]^{-1}\frac{\partial Q(\theta|\theta^{(k)})}{\partial\beta_j} \\
\beta_j^{(k+1)} &= \beta_j^{(k)} + \Big(\sum_{a=1}^{N} X'_a L_{aj}^{(k)} W_{aj}^{-1(k)} X_a\Big)^{-1}\Big(\sum_{a=1}^{N} X'_a L_{aj}^{(k)} W_{aj}^{-1(k)}\frac{\partial\eta_{aj}}{\partial\mu_{aj}}^{(k)}[y_{a1}^{T_a}-\mu_{aj}^{(k)}]\Big). \qquad (3.4)
\end{aligned}
$$

Si l'on introduit le vecteur dépendant défini par :

$$z_{aj} = \eta_{aj} + \frac{\partial \eta_{aj}}{\partial \mu_{aj}}[y_{a1}^{T_a} - \mu_{aj}] = X_a\beta_j + \frac{\partial \eta_{aj}}{\partial \mu_{aj}}[y_{a1}^{T_a} - \mu_{aj}].$$

Les équations (3.3) deviennent alors :

$$\sum_{a=1}^{N} X_a' L_{aj} W_{aj}^{-1}[z_{aj} - X_a\beta_j] = 0,$$

ou encore

$$X' L_j W_j^{-1}[z_j - X\beta_j] = 0, \tag{3.5}$$

où

- $X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$ est la matrice de dimension $T \times Q$ des covariables,
- $L_j W_j^{-1} = \text{Diag}\Big\{L_{aj} W_{aj}^{-1}\Big\}_{a=1,\ldots,N}$ est une matrice diagonale par blocs de dimension $T \times T$,
- $z_j - X\beta_j = \begin{pmatrix} z_{1j} - X_1\beta_j \\ \vdots \\ z_{Nj} - X_N\beta_j \end{pmatrix}$ est un vecteur colonne de taille T.

Nous pouvons ainsi conclure que résoudre itérativement les équations (3.3) est équivalent à résoudre itérativement les équations (3.5) comme des équations normales (propres aux modèles linéaires). À chaque itération, la valeur courante de β_j permet le calcul de la matrice des poids $(L_j W_j^{-1})^{-1}$ et du vecteur dépendant z_j, et alors l'obtention, par résolution du système linéarisé (3.5) d'une nouvelle valeur de β_j. Cette réécriture permet une interprétation de type linéaire (McCullagh et Nelder, 1989).

Pour β_j fixé, en considérant z comme un nouveau vecteur de données et $(L_j W_j^{-1})^{-1}$ comme une matrice de poids connue et fixée, les équations (3.5) peuvent être vues comme les équations classiques des moindres carrés généralisés associés au modèle :

$$Z_j = X\beta_j + e_j \qquad \text{où} \qquad \text{E}(e_j) = 0; \qquad \text{var}(e_j) = (L_j W_j^{-1})^{-1}.$$

La matrice W_j^{-1} est pondérée par la matrice L_j, matrice diagonale des probabilités d'appartenir à un état sachant les données observées.

Simplification dans le cas des liens canoniques

Rappelons qu'un lien canonique se définit de la manière suivante pour un état j :

$$\eta_{atj} = \theta_{atj} = g(\mu_{atj}) = X_{at}\beta_j.$$

Dans le cas d'un lien canonique, nous avons :

$$\frac{\partial \mu_{atj}}{\partial \eta_{atj}} = \frac{\partial \mu_{atj}}{\partial \theta_{atj}} = \frac{\partial k'(\theta_{atj})}{\partial \theta_{atj}} = k''(\theta_{atj}) = \frac{\text{var}(Y_{at}|S_{at} = j)}{h_{at}(\phi)}.$$

Nous obtenons ainsi

$$W_{aj}^{-1} = \text{Diag}\Big\{\frac{\text{var}(Y_{at}|S_{at}=j)}{h_{at}(\phi)^2}\Big\}_{t=1,\ldots,T_a},$$

$$W_{aj}^{-1}\frac{\partial \eta_{aj}}{\partial \mu_{aj}} = \text{Diag}\Big\{\frac{1}{h_{at}(\phi)}\Big\}_{t=1,\ldots,T_a}.$$

Remarque :

L'inverse de la matrice d'information de Fisher $\left[E\Big(\frac{\partial^2 Q(\theta|\theta)}{\partial\beta_j\partial\beta_j'}\Big)\right]^{-1} = \Big(\sum_{a=1}^{N} X_a' L_{aj} W_{aj}^{-1} X_a\Big)^{-1}$ *(équation (3.4)) peut être utilisée comme mesure de précision des estimations des paramètres de régression conditionnellement aux états non-observés. En effet, il fournit dans chaque état un estimateur de la matrice de variance-covariance des paramètres de régression estimés.*

3.2.2 Formalisme de l'algorithme du gradient MCEM pour MS-GLM

Nous avons présenté le formalisme du gradient EM pour les combinaisons markoviennes de modèles linéaires généralisés. Dans cette approche, les séquences d'états sont restaurées de manière probabiliste dans l'étape E. Nous pouvons transposer l'algorithme proposé au cas où les séquences d'états sont restaurées de manière explicite par simulation (cf section 2.2.4.3 pour les chaînes de Markov cachées). Dans ce cas, nous proposons un algorithme itératif, fondé sur l'algorithme du gradient MCEM (section 2.1.3.1), en deux étapes :

- **Étape E :**
  - une restauration par simulation de séquences d'états à partir de l'algorithme "avant-arrière" de simulation décrit à la section 2.2.4.3 pour les chaînes de Markov cachées classiques (Chib, 1996),
- **Étape M :**
  - une ré-estimation des paramètres de la chaîne de Markov sous jacente (probabilités initiales et probabilités de transition) par maximisation directe de l'approximation de la log-vraisemblance des données complètes (équation (3.1)),
  - une ré-estimation itérative des paramètres des modèles linéaires généralisés par un algorithme des scores de Fisher où la matrice de poids ne prend plus en compte les pseudo-comptages $L_{aj}(t)$ mais les comptages $I(S_{at}=j)$, les séquences d'états étant maintenant restaurées explicitement.

Étape M :

Estimation des paramètres liés aux chaînes de Markov cachées π_j et p_{ij}

On obtient après maximisation de l'approximation de $\log f(y,s;\theta)$ (équation (3.1)) les estimations des paramètres associés à la chaîne de Markov sous-jacente. Ceci nous conduit aux formules de ré-estimation à l'itération k :

– probabilités initiales

$$\pi_j^{(k+1)} = \frac{\sum_a \sum_{m=1}^{M_k} I\Big(s_{a1}^{(k)}(m) = j\Big)}{N M_k},$$

– probabilités de transition

$$p_{ij}^{(k+1)} = \frac{\sum_a \sum_{m=1}^{M_k} \sum_{t=1}^{T_a-1} I\Big(s_{at}^{(k)}(m) = i, s_{a,t+1}^{(k)}(m) = j\Big)}{\sum_a \sum_{m=1}^{M_k} \sum_{t=1}^{T_a-1} I\Big(s_{at}^{(k)}(m) = i\Big)},$$

où

- $s_{at}^{(k)}(m)$ est l'état dans lequel se trouve l'individu a au temps t pour la $m^{ème}$ séquence d'états simulée à l'itération k,
- M_k est le nombre de séquences d'états simulées pour chaque individu à l'itération k.

Dans le cas de la restauration par simulation des séquences d'états (cf section 2.2.4.3 pour les chaînes de Markov cachées), comme nous estimons nos paramètres directement à partir de la log-vraisemblance des données complètes, nous pouvons utiliser l'algorithme des scores de Fisher qui se traduit pour les combinaisons markoviennes de modèles linéaires généralisés par :

$$\begin{aligned}
\beta_j^{(k+1)} &= \beta_j^{(k)} - \left[\mathrm{E}\Big(\frac{\partial^2 \log(s,y;\theta)}{\partial\beta_j \partial\beta_j'}\Big)\right]^{-1} \frac{\partial \log(s,y;\theta)}{\partial\beta_j} \qquad (3.6)\\
\beta_j^{(k+1)} &= \beta_j^{(k)} + \Big(\sum_{a=1}^{N}\sum_{m=1}^{M_k} X_a' R_{aj}^{(k)}(m) W_{aj}^{-1(k)} X_a\Big)^{-1} \Big(\sum_{a=1}^{N}\sum_{m=1}^{M_k} X_a' R_{aj}^{(k)}(m) W_{aj}^{-1(k)} \frac{\partial \eta_{aj}}{\partial \mu_{aj}}^{(k)} [y_{a1}^{T_a} - \mu_{aj}^{(k)}]\Big)
\end{aligned}$$

où

$$R_{aj}(m) = \mathrm{Diag}\Big\{ I\Big(s_{at}(m) = j\Big)\Big\}_{t=1,\ldots,T_a}.$$

Nous pouvons remarquer que l'étape de maximisation pour les paramètres β_j de l'algorithme du gradient EM avec restauration probabiliste des séquences d'états se transpose naturellement à l'étape de maximisation pour les paramètres β_j de l'algorithme du gradient MCEM avec restauration par simulation des séquences d'états. En effet, la matrice L_{aj} est simplement remplacée par la matrice des indicatrices $R_{aj}(m)$ pour chaque séquence d'états simulée pour un individu a.

3.2.3 Extension au SMS-GLM

Les combinaisons semi-markoviennes de modèles linéaires généralisés (définition 3.2) différent uniquement des combinaisons markoviennes de modèles linéaires généralisés (définition 3.1) dans la structure du processus sous-jacent. Par suite, comme l'influence des

covariables n'est prise en compte que dans le lien entre le processus d'observation et le processus d'état, les algorithmes proposés pour estimer les paramètres du MS-GLM se transposent directement au SMS-GLM.

Par conséquent, dans le but d'estimer les paramètres des combinaisons semi-markoviennes de modèles linéaires généralisés, nous proposons les deux types d'algorithmes itératifs :

Algorithme du gradient EM avec restauration probabiliste

- **Étape E :**
  - une restauration probabiliste de toutes les séquences d'états à partir de l'algorithme "avant-arrière" propre aux semi-chaînes de Markov cachées,

- **Étape M :**
  - une ré-estimation des paramètres de la semi-chaîne de Markov sous jacente (probabilités initiales, probabilités de transition et lois d'occupation) par maximisation directe de l'espérance de la log-vraisemblance des données complètes sachant les données observées,
  - une ré-estimation itérative des paramètres des modèles linéaires généralisés par un algorithme des scores de Fisher identique à l'équation (3.4).

Algorithme du gradient MCEM avec restauration par simulation

- **Étape E :**
  - une restauration par simulation de séquences d'états à partir de l'algorithme "avant-arrière" de simulation propre aux semi-chaînes de Markov cachées,

- **Étape M :**
  - une ré-estimation des paramètres de la semi-chaîne de Markov sous jacente (probabilités initiales, probabilités de transition et lois d'occupation) par maximisation directe de l'approximation de la log-vraisemblance des données complètes,
  - une ré-estimation itérative des paramètres des modèles linéaires généralisés par un algorithme des scores de Fisher identique à l'équation (3.6).

3.2.4 Convergence des algorithmes proposés

La convergence des algorithmes proposés est contrôlée numériquement par le calcul de la vraisemblance des données observées. En effet, comme pour les chaînes de Markov cachées et les semi-chaînes de Markov cachées, cette quantité peut être directement obtenue lors de l'étape E. Cette quantité est directement déduite de l'algorithme "avant-arrière" pour chaînes de Markov cachées ou semi-chaînes de Markov cachées comme étant le produit des constantes de normalisation de la récurrence "avant".

Nous nous intéressons par la suite à deux cas particuliers : les données poissonniennes et les données binaires. La présentation de ces cas particuliers est accompagnée de simulations et d'exemples sur données réelles.

3.3 Données de comptage

3.3.1 Algorithme du gradient EM pour données de comptage

Dans le cadre des combinaisons markoviennes de modèles linéaires généralisés pour données de comptage, les données observées appartiennent au domaine $\mathbb{N}$, ensemble des entiers naturels positifs. Ces données peuvent être par exemple, pour des données de croissance d'arbres, le nombre de rameaux portés par pousse annuelle ou le nombre d'entrenoeuds par pousse annuelle. La chaîne de Markov sous-jacente représente alors la succession de phases de croissance tandis que le modèle linéaire généralisé associé à chaque état modélise par exemple dans la phase de croissance correspondante l'influence du climat sur le nombre d'entrenoeuds.

Nous allons dans un premier temps définir le lien entre le processus d'observation et le processus d'état. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation Y_{at} d'un individu a au temps t est modélisée par la fonction de masse de la loi de Poisson de paramètre $\mu_{ats_{at}}$:

$$b_{s_{at}}(y_{at}) = \mathrm{P}(Y_{at} = y_{at}|S_{at} = s_{at}) = \frac{[\mu_{ats_{at}}]^{y_{at}}}{y_{at}!} \exp\Big(-\mu_{ats_{at}}\Big).$$

Dans le cas des données poissonniennes, la fonction $h_{at}(\phi)$ est égale à 1.

Le lien canonique pour les données poissonniennes est le lien logarithme qui conduit à la relation suivante :

$$\log(\mu_{s_{at}}) = X_{at}\beta_{s_{at}}.$$

La log-vraisemblance des données complètes (cf équation (3.1)) est donnée par :

$$\begin{aligned}\log f(s,y;\theta) &= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1}=j\Big)\log\pi_j + \sum_{a=1}^{N}\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} I\Big(s_{at}=j, s_{a,t-1}=i\Big)\log p_{ij} \\ &+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\Big(s_{at}=j\Big)\Big(-\log(y_{at}!) - \mu_{atj} + y_{at}\log(\mu_{atj})\Big),\end{aligned}$$

qui peut se réécrire de la façon suivante :

$$\begin{aligned}\log f(s,y;\theta) &= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1}=j\Big)\log\pi_j + \sum_{a=1}^{N}\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} I\Big(s_{at}=j, s_{a,t-1}=i\Big)\log p_{ij} \\ &+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\Big(s_{at}=j\Big)\Big(-\log(y_{at}!) - \exp(X_{at}\beta_j) + y_{at}X_{at}\beta_j\Big).\end{aligned}$$

L'écriture du modèle, à partir des résultats obtenus dans un cadre général, nous amène à estimer, à l'étape de maximisation, le vecteur β_j des paramètres du modèle linéaire généralisé associé à l'état j par les équations des scores de Fisher :

$$\beta_j^{(k+1)} = \beta_j^{(k)} + \Big(\sum_{a=1}^{N} X_a' L_{aj}^{(k)} W_{aj}^{-1(k)} X_a\Big)^{-1} \Big(\sum_{a=1}^{N} X_a' L_{aj}^{(k)} (y_{a1}^{T_a} - \mu_{aj}^{(k)})\Big)$$

avec

$$W_{aj}^{-1} = \mathrm{Diag}\Big\{\exp(X_{at}\beta_j)\Big\}_{t=1,\ldots,T_a}$$

et

$$\mu_{aj} = \Big(\exp(X_{at}\beta_j)\Big)_{t=1,\ldots,T_a}.$$

3.3.2 Simulations

Nous présentons les résultats de simulations dans le cas de données poissonniennes. Nous considérons une combinaison markovienne de modèles linéaires généralisés avec une chaîne de Markov sous-jacente ergodique à 2 états, définie par le vecteur π des probabilités initiales et la matrice P des probabilités de transition suivants :

$$\pi = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix},$$

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}.$$

Les paramètres d'effets fixes des modèles linéaires généralisés sont pour l'état 1, $\beta_1 = (-1, 0.5)$ et pour l'état 2, $\beta_2 = (0.1, 0.7)$. La matrice d'incidence X associée à ces effets fixes est définie par une première colonne de 1 et une seconde colonne générée à partir de la loi $\mathcal{N}(5, 1)$. Nous avons simulé 50 individus ($N = 50$) sur 30 temps ($\forall a, T_a = 30$). Les moyennes et les écart-types des estimations calculées sur 100 jeux de données simulés sont résumés dans le tableau 3.1.

					état 1		état 2	
paramètre		π_1	p_{11}	p_{21}	β_{11}	β_{12}	β_{21}	β_{22}
valeur simulée		0.9	0.8	0.3	-1	0.5	0.1	0.7
estimation	moyenne	0.901	0.799	0.304	-1.002	0.500	0.098	0.700
	écart-type	0.040	0.014	0.021	0.076	0.013	0.032	0.006

TAB. 3.1 – *Résultats d'estimation des paramètres d'une combinaison markovienne à 2 états de modèles linéaires généralisés pour données poissonniennes par l'algorithme du gradient EM avec restauration probabiliste sur 100 simulations : cas Poisson - lien log - composantes bien séparées.*

Nous constatons un bon comportement de notre approche basée sur l'algorithme du gradient EM, que ce soit pour les paramètres β des probabilités d'observation ou pour l'estimation des paramètres π et P de la chaîne de Markov sous-jacente. Nous notons une certaine stabilité de l'algorithme pour données poissonniennes avec lien logarithme ; voir les faibles écart-types des estimations dans le tableau 3.1.

Afin d'évaluer la qualité des estimations obtenues, nous nous sommes également intéressés :

- à l'influence du degré de séparabilité entre les états,
- à l'influence du nombre d'observations.

Afin d'évaluer l'influence du degré de séparabilité entre les états (i.e. si les états sont bien distincts ou pas), nous avons effectué des simulations similaires où seules les probabilités d'observation (par suite, les effets fixes) sont modifiées. Nous avons considéré de nouvelles valeurs pour $\beta_1 = (0.2, 0.6)$ sur l'état 1 et pour $\beta_2 = (0.1, 0.7)$ sur l'état 2. Le tableau 3.2 donne les moyennes et les écart-types des estimations obtenues sur 100 jeux de données simulés.

					état 1		état 2	
paramètre		π_1	p_{11}	p_{21}	β_{11}	β_{12}	β_{21}	β_{22}
valeur simulée		0.9	0.8	0.3	0.2	0.6	0.1	0.7
estimation	moyenne	0.907	0.798	0.304	0.203	0.600	0.102	0.700
	écart-type	0.051	0.017	0.026	0.035	0.006	0.043	0.007

TAB. 3.2 – *Résultats d'estimation des paramètres d'une combinaison markovienne à 2 états de modèles linéaires généralisés pour données poissonniennes par l'algorithme du gradient EM avec restauration probabiliste sur 100 simulations : cas Poisson - lien log - composantes moins bien séparées.*

Nous pouvons constater que dans le cas des données poissonniennes, le degré de séparabilité entre les états de la chaîne de Markov sous-jacente a peu d'influence sur l'estimation des paramètres en moyenne. Cette influence est cependant plus prononcée sur les écart-types des estimations pour les paramètres de la chaîne de Markov sous-jacente. En effet, plus les états sont séparés, plus les écart-types des estimations diminuent.

Nous avons également calculé la matrice de corrélation des paramètres de régression estimés $(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})$ dans le cas de composantes bien séparées :

$$\begin{pmatrix} 1 & -0.99 & 0.04 & -0.06 \\ -0.99 & 1 & -0.03 & 0.05 \\ 0.04 & -0.03 & 1 & -0.99 \\ -0.06 & 0.05 & -0.99 & 1 \end{pmatrix}$$

et dans le cas de composantes moins bien séparées :

$$\begin{pmatrix} 1 & -0.99 & -0.22 & 0.23 \\ -0.99 & 1 & 0.26 & -0.28 \\ -0.22 & 0.26 & 1 & -0.99 \\ 0.23 & -0.28 & -0.99 & 1 \end{pmatrix}.$$

Nous pouvons constater que les estimations des paramètres de régression dans chaque état sont fortement corrélées négativement. Les corrélations entre les estimations des paramètres de régression du premier état et les estimations des paramètres de régression du deuxième état sont faibles et d'autant plus faibles lorsque les états sont bien séparés. Le degré de séparabilité entre les états influence fortement ces corrélations. Nous avons observé le même phénomène pour les estimations des probabilités de transition.

Nous avons étudié l'influence du nombre d'observations sur l'estimation des paramètres des MS-GLM pour données poissonniennes avec lien logarithme dans le cas où les composantes étaient moins bien séparées. Nous avons fixé la longueur des séquences observées $\forall a, T_a = 30$ et avons simulé 100 jeux de données avec N=5, 10, 20 ou 50 observations. Les comportements des estimations pour chaque paramètre du MS-GLM sont résumés dans les boîtes à moustaches de la figure 3.1. Nous avons de plus fixé le nombre d'individus observés à $N = 50$ et avons simulé 100 jeux de données avec $\forall a, T_a = 5, 10, 20$ ou 30 temps. Les comportements des estimations pour chaque paramètre du MS-GLM sont résumés dans les boîtes à moustaches de la figure 3.2. Les résultats obtenus montrent que le nombre d'observations ($\sum_a T_a$) a une forte influence sur l'estimation des paramètres. Plus son nombre est élevé, meilleure est l'estimation. Il semblerait que dans le cas des données poissonniennes, au moins 600 observations (nombre d'observations supérieur à 20 séquences de longueur 30 ou supérieur à 50 séquences de longueur 10) soient nécessaires pour obtenir de bonnes estimations.

3.3.3 Application aux données d'IRM

Les IRM (images par résonance magnétique) de patients atteints de sclérose en plaques sont une source de données pouvant être convenablement modélisées par des combinaisons markoviennes de modèles linéaires généralisés pour données poissonniennes (Altman, 2007). En effet, les patients touchés par la sclérose en plaques alternent des périodes de rémission et des périodes de rechute. La principale caractéristique permettant de déterminer l'état d'un patient à une date donnée est de compter à l'aide d'IRM le nombre de lésions dans le cerveau. Si ce nombre est élevé, le patient est en phase de rechute. Par analogie, si ce nombre est faible, le patient est en phase de rémission. Des IRM de patients atteints de sclérose en plaques ont été initialement analysées par Albert (1991). Les données issues d'une étude à Vancouver PRISMS Study Group (1998) ont été traitées par Li et Paty (1999) puis par Altman et Petkau (2005) et Altman (2007). Ces données

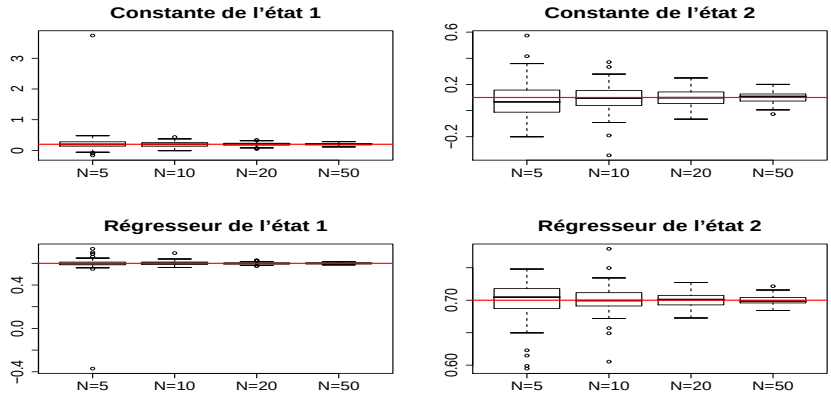

FIG. 3.1 – *Boîtes à moustaches des estimations des paramètres des modèles linéaires généralisés pour N=5,10, 20 et 50 individus et T=30 temps pour chaque individu sur 100 simulations : cas Poisson - lien log - composantes moins bien séparées. Le trait rouge représente la vraie valeur pour chaque paramètre.*

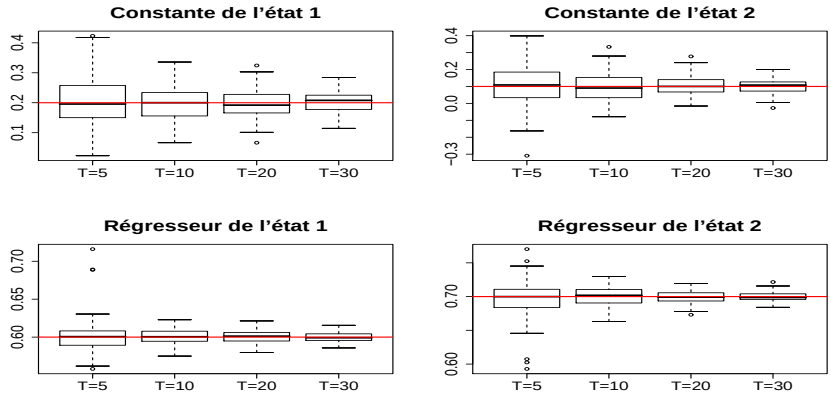

FIG. 3.2 – *Boîtes à moustaches des estimations des paramètres des modèles linéaires généralisés pour N=50 individus et T=5, 10, 20 ou 30 temps pour chaque individu sur 100 simulations : cas Poisson - lien log - composantes moins bien séparées. Le trait rouge représente la vraie valeur pour chaque paramètre.*

nous ont été gracieusement fournies par le Dr. Rachel Altman en accord avec le Dr. Paul Albert. Les données de Vancouver sont constituées de 13 patients placebo. Les séquences observées comptent au minimum 2 observations et au maximum 26 observations (figure 3.3).

Une combinaison markovienne de modèles linéaires généralisés à 2 états (état 1 : rémission ; état 2 : rechute) a été estimée sur la base du nombre de lésions dans le cerveau. Soit Y_{at} le nombre de lésions du patient a au temps t et S_{at} l'état caché associé. Altman (2003) suppose que conditionnellement à $S_{at} = s_{at}$, $Y_{at}|S_{at} = s_{at}$ suit une loi de Poisson de paramètre $\mu_{ats_{at}}$ telle que

$$\log(\mu_{ats_{at}}) = \beta_{s_{at}}$$

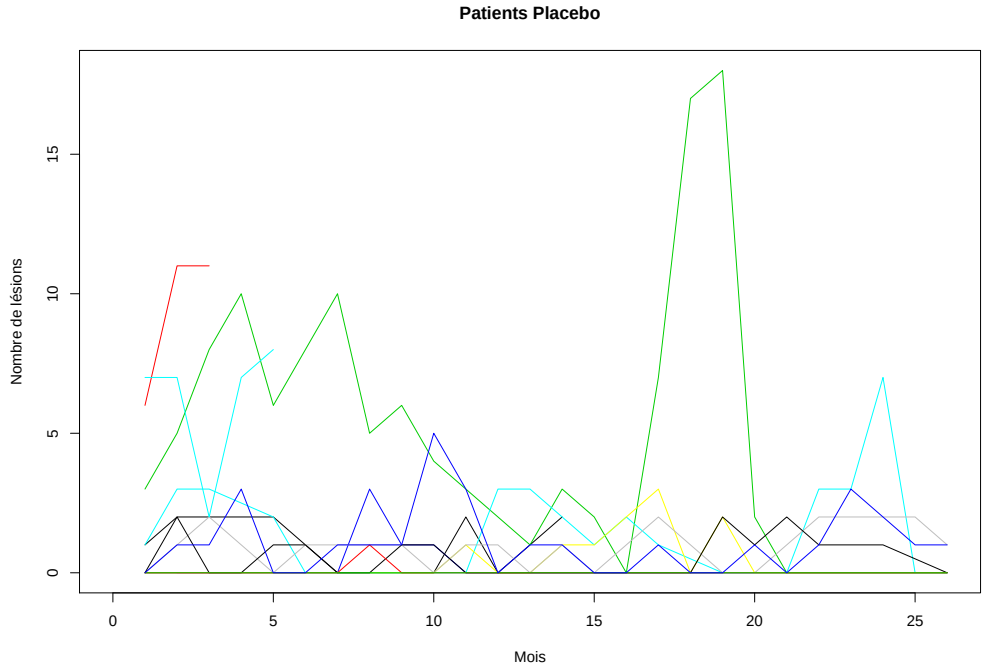

FIG. 3.3 – *Données de PRISMS Vancouver : nombre de lésions dans le cerveau en fonction du mois de mesure pour les 13 patients placebo.*

où $\beta_{s_{at}}$ est le paramètre d'effet fixe associé à l'état $S_{at} = s_{at}$. Le tableau 3.3 résume les estimations obtenues par notre modèle et les résultats obtenus par l'approche de Altman (2003) qui maximise directement la vraisemblance des données observées par des approches numériques de type quasi-Newton.

Altman suppose que la chaîne de Markov sous-jacente est en équilibre. Sous cette hypothèse, il n'est pas nécessaire d'estimer les probabilités initiales. En effet, la distribution stationnaire est :

$$\begin{aligned}\begin{pmatrix}\pi_1 & \pi_2\end{pmatrix} &= \begin{pmatrix}\frac{p_{21}}{p_{12}+p_{21}} & \frac{p_{12}}{p_{12}+p_{21}}\end{pmatrix} \\ &= \begin{pmatrix}\frac{0.184}{0.024+0.184} & \frac{0.024}{0.024+0.184}\end{pmatrix} \\ &= \begin{pmatrix}0.885 & 0.115\end{pmatrix}.\end{aligned}$$

Cette hypothèse permet d'expliquer la légère différence entre les log-vraisemblances. En effet, les valeurs des probabilités initiales obtenues par les deux approches semblent significativement différentes. Ces dernières ont un impact sur les restaurations probabilistes des séquences d'états pour chaque individu. Nous constatons également que les deux matrices de transition estimées sont relativement proches ; voir le tableau 3.3. Les probabilités d'observations conditionnellement aux états sont proches ; les différences pouvant être expliquées par l'hypothèse d'équilibre de la chaîne de Markov sous-jacente et par le faible

nombre d'observations (13 patients observés sur au minimum 2 mois et au maximum 26 mois). Selon notre estimation, nous pouvons considérer que l'état de rémission est caractérisé par un nombre moyen de lésions égal à $\exp(\beta_1) = \exp(-0.614) = 0.54$ et que l'état de rechute est caractérisé par un nombre moyen de lésions égal à $\exp(\beta_2) = \exp(2.001) = 7.40$.

paramètre	Altman (2003)	Gradient EM
π_1	NA	0.836
p_{11}	0.976	0.986
p_{21}	0.184	0.198
β_1	-0.907	-0.614 (0.09)
β_2	1.657	2.001 (0.08)
log-vraisemblance	-326.34	-319.41

TAB. 3.3 – *Comparaison des paramètres des combinaisons markoviennes de modèles linéaires généralisés à 2 états, estimés par l'approche de Altman (2003) et par l'algorithme du gradient EM proposé sur la base des 13 patients placebo des données de PRISMS Vancouver. La précision (écart-type) des estimations des paramètres de régression est donnée entre parenthèses.*

Nous avons restauré les séquences d'états les plus probables pour chaque patient à partir d'une adaptation de l'algorithme de Viterbi pour chaînes de Markov cachées (Foreman, 1993) aux MS-GLM. Nous nous intéressons en particulier au patient 3 et avons restauré les 10 séquences d'états les plus probables :

```
3 5 8 10 6 8 10 5 6 4 3 2 1 3 2 0 7 17 18 2 0 0 0 0 0 0 (nombre de lésions)

2 2 2  2 2 2  2 2 2 2 2 1 1 1 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.5018  0.5018)
2 2 2  2 2 2  2 2 2 2 1 1 1 1 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.1961  0.6979)
2 2 2  2 2 2  2 2 2 2 2 1 1 1 1 1 1 2  2  2 2 1 1 1 1 1 1 (0.0996  0.7975)
2 2 2  2 2 2  2 2 2 2 2 2 1 1 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.0995  0.8970)
2 2 2  2 2 2  2 2 2 2 2 1 1 1 1 1 1 2  2  2 2 1 1 1 1 1 1 (0.0389  0.9359)
2 2 2  2 2 2  2 2 2 2 2 2 1 1 1 1 1 2  2  2 2 1 1 1 1 1 1 (0.0198  0.9557)
1 2 2  2 2 2  2 2 2 2 2 1 1 1 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.0089  0.9646)
2 2 2  2 2 2  2 2 2 1 1 1 1 1 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.0059  0.9705)
2 2 2  2 2 2  2 2 2 2 2 2 2 2 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.0039  0.9744)
2 2 2  2 2 2  2 2 2 2 2 1 1 2 1 1 1 2  2  2 1 1 1 1 1 1 1 (0.0038  0.9782)
```

où 1 représente l'état de rémission, 2 représente l'état de rechute, le premier indicateur entre parenthèses est la probabilité *a posteriori* de la séquence d'états et le deuxième indicateur est la probabilité *a posteriori* cumulée des n séquences d'états les plus probables. La séquence d'états la plus probable a un poids fort (0.5018) par rapport aux autres séquences d'états. La différence entre ces séquences d'états réside principalement dans le premier passage de la phase de rechute à la phase de rémission (entre les mois 11 et 12 pour la première séquence, entre les mois 10 et 11 pour la seconde). À la vue de ces

séquences d'états, nous pouvons conclure de manière unanime qu'il y a un changement d'état pour ce patient entre les mois 16 et 17.

3.4 Données binaires

3.4.1 Algorithme du gradient EM pour données binaires

Dans le cadre des combinaisons markoviennes de modèles linéaires généralisés pour données binaires, les données observées appartiennent au domaine $\{0;1\}$. Dans le cadre de la croissance d'arbres forestiers, ces données peuvent par exemple caractériser l'absence (0) ou la présence (1) d'inflorescence, la mortalité (0) ou non (1) de l'apex. La chaîne de Markov sous-jacente représente alors la succession de phases de croissance tandis que le modèle linéaire généralisé associé à chaque état modélise par exemple dans la phase de croissance correspondante l'influence du climat sur la présence/absence d'inflorescence.

Posons

$$\mu_{ats_{at}} = \mathrm{P}(Y_{at} = 1 | S_{at} = s_{at}) = 1 - \mathrm{P}(Y_{at} = 0 | S_{at} = s_{at}).$$

Dans un cas concret, $\mu_{ats_{at}}$ peut représenter la probabilité que l'apex meure sachant les covariables climatiques et la phase de croissance dans laquelle l'individu a se trouve au temps t.

Nous allons dans un premier temps définir le lien entre le processus d'observation et le processus d'état. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation Y_{at} d'un individu a au temps t est modélisée par la fonction de masse de la loi de Bernoulli de paramètre $\mu_{ats_{at}}$:

$$b_{s_{at}}(y_{at}) = \mathrm{P}(Y_{at} = y_{at} | S_{at} = s_{at}) = (\mu_{ats_{at}})^{y_{at}} (1 - \mu_{ats_{at}})^{1-y_{at}}.$$

Dans le cas des données binaires, la fonction $h_{at}(\phi)$ est égale à 1.

Le lien canonique pour les données binaires est le lien logit qui nous donne la relation suivante :

$$\mathrm{logit}(\mu_{ats_{at}}) = \log \frac{\mu_{ats_{at}}}{1 - \mu_{ats_{at}}} = X_{at} \beta_{s_{at}}.$$

La log-vraisemblance des données complètes (cf équation (3.1)) est donnée par :

$$\begin{aligned} \log f(s, y; \theta) &= \sum_{a=1}^{N} \sum_{j=1}^{J} I\Big(s_{a1} = j\Big) \log \pi_j + \sum_{a=1}^{N} \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} I\Big(s_{at} = j, s_{a,t-1} = i\Big) \log p_{ij} \\ &+ \sum_{a=1}^{N} \sum_{j=1}^{J} \sum_{t=1}^{T_a} I\Big(s_{at} = j\Big) \left(y_{at} \log\Big(\frac{\mu_{atj}}{1 - \mu_{atj}}\Big) + \log(1 - \mu_{atj}) \right), \end{aligned}$$

qui peut se réécrire de la façon suivante :

$$\log f(s, y; \theta) = \sum_{a=1}^{N} \sum_{j=1}^{J} I\Big(s_{a1} = j\Big) \log \pi_j + \sum_{a=1}^{N} \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} I\Big(s_{at} = j, s_{a,t-1} = i\Big) \log p_{ij}$$

$$+ \sum_{a=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\Big(s_{at}=j\Big)\Big(y_{at}X_{at}\beta_j - \log(1+\exp(X_{at}\beta_j))\Big).$$

L'écriture du modèle, à partir des résultats obtenus dans un cadre général, nous amène à estimer, à l'étape de maximisation, le vecteur β_j des paramètres, associé à l'état j par les équations des scores de Fisher :

$$\beta_j^{(k+1)} = \beta_j^{(k)} + \Big(\sum_{a=1}^{N} X_a' L_{aj}^{(k)} W_{aj}^{-1(k)} X_a\Big)^{-1}\Big(\sum_{a=1}^{N} X_a' L_{aj}^{(k)}(y_{a1}^{T_a} - \mu_{aj}^{(k)})\Big)$$

avec

$$W_{aj}^{-1} = \mathrm{Diag}\Big\{\mu_{atj}(1-\mu_{atj})\Big\}_{t=1,\ldots,T_a} = \mathrm{Diag}\Big\{\frac{\exp(X_{at}\beta_j)}{\big(1+\exp(X_{at}\beta_j)\big)^2}\Big\}_{t=1,\ldots,T_a}$$

et

$$\mu_{aj} = \Big(\frac{\exp(X_{at}\beta_j)}{1+\exp(X_{at}\beta_j)}\Big)_{t=1,\ldots,T_a}.$$

3.4.2 Simulations

Dans cette partie, nous présentons les résultats de simulation dans le cas de données binaires avec lien logit. Nous considérons une combinaison markovienne de modèles linéaires généralisés avec une chaîne de Markov sous-jacente ergodique à 2 états, définie par le vecteur π des probabilités initiales et la matrice P des probabilités de transition suivants :

$$\pi = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix},$$

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}.$$

Les paramètres d'effets fixes sont pour l'état 1, $\beta_1 = (-5, 0.4)$ et pour l'état 2, $\beta_2 = (5, -0.4)$. La matrice d'incidence X associée à ces effets fixes est définie par une première colonne de 1 et une seconde colonne générée à partir de la loi $\mathcal{N}(5,1)$. Nous avons simulé 50 individus ($N = 50$) sur 30 temps ($\forall a, T_a = 30$). Les moyennes et les écart-types des estimations calculées sur 100 jeux de données simulés sont résumés dans le tableau 3.4.

Excepté pour l'estimation des constantes, nous constatons un bon comportement de l'approche par gradient EM pour l'estimation des paramètres de la chaîne de Markov sous-jacente et des paramètres des modèles linéaires généralisés. En effet, les écart-types associés aux constantes sur chaque état β_{11} et β_{21} sont relativement élevés.

Comme pour les simulations de données poissonniennes, nous avons étudié la qualité des estimations selon le degré de séparabilité entre les états (c'est-à-dire si les états sont

					état 1		état 2	
paramètre		π_1	p_{11}	p_{21}	β_{11}	β_{12}	β_{21}	β_{22}
valeur simulée		0.9	0.8	0.3	-5	0.4	5	-0.4
estimation	moyenne	0.911	0.80	0.296	-5.415	0.457	4.909	-0.353
	écart-type	0.050	0.023	0.031	2.766	0.436	4.069	0.647

TAB. 3.4 – *Résultats d'estimation des paramètres d'une combinaison markovienne à 2 états de modèles linéaires généralisés pour données binaires par l'algorithme du gradient EM avec restauration probabiliste sur 100 simulations : cas binaire - lien logit - composantes bien séparées.*

					état 1		état 2	
paramètre		π_1	p_{11}	p_{21}	β_{11}	β_{12}	β_{21}	β_{22}
valeur simulée		0.9	0.8	0.3	-3	0.4	3	-0.4
estimation	moyenne	0.848	0.827	0.186	-3.043	0.413	2.017	-0.279
	écart-type	0.140	0.072	0.069	1.548	0.252	1.961	0.319

TAB. 3.5 – *Résultats d'estimation des paramètres d'une combinaison markovienne à 2 états de modèles linéaires généralisés pour données binaires par l'algorithme du gradient EM avec restauration probabiliste sur 100 simulations : cas binaire - lien logit - composantes moins bien séparées.*

bien distincts ou pas). Le tableau 3.4, que nous venons de commenter, résume les résultats dans le cas où les probabilités d'observation conditionnellement à l'état d'appartenance sont éloignées. Afin de se placer dans un cas où ces probabilités sont proches, nous avons effectué des simulations identiques où seules les probabilités d'observation sont modifiées. Nous avons considéré de nouvelles valeurs $\beta_1 = (-3, 0.4)$ et $\beta_2 = (3, -0.4)$. Le tableau 3.5 donne les moyennes et les écarts-types des estimations obtenues sur 100 jeux de données simulées. Nous pouvons noter que l'approche par gradient EM a plus de difficultés à estimer les paramètres de la chaîne de Markov sous-jacente lorsque les états sont moins bien séparés. Nous constatons également pour ces paramètres, une augmentation des écart-types lorsque la séparabilité entre les états diminue. Nous pouvons remarquer que contrairement au premier état, l'algorithme a des difficultés à estimer correctement les paramètres associés au deuxième état. Martinez (2006) et Lavergne et al. (2007) ont constaté les mêmes phénomènes dans les cas de mélanges de modèles linéaires mixtes pour le premier et de mélanges pour des données répétées de loi exponentielle pour les seconds. De plus, comme dans le cas des simulations pour données poissonniennes, le degré de séparabilité entre les états influence fortement les corrélations entre les estimations des paramètres de régression du premier état et les estimations des paramètres de régression du deuxième état.

Nous ne présentons pas ici l'influence du nombre d'observations sur les estimations des paramètres des MS-GLM. Nous avons cependant constaté le même phénomène que pour les données poissonniennes : plus il y a d'observations, meilleures sont les estimations.

3.4.3 Application aux données de croissance de pins Laricio

Les combinaisons semi-markoviennes de modèles linéaires généralisés trouvent notamment leurs intérêts dans la modélisation de données de croissance d'arbres forestiers en fonction de covariables climatiques. Nous nous intéressons en particulier dans cet exemple aux pins Laricio de Corse (cf section 1.3.1). Les données sont constituées des unités de croissance par pousse annuelles des 30 pins Laricios de 18 ans (la première année n'a pas été mesurée). Bien que les pins laricios soient supposés monocycliques (une unité de croissance (UC) par pousse annuelle), les mesures faites sur ces 30 pins Laricios ont montré un phénomène rare de polycyclisme avec 2 unités de croissance par pousse annuelle (Meredieu, 1998). Un arbre pouvant donc avoir 1 ou 2 UC, nous supposons que la variable "nombre d'UCs par pousse annuelle" peut être modélisée sous forme d'une variable binaire avec pour modalité 0 : une unité de croissance et 1 : 2 unités de croissance. Dans ce cas, conditionnellement à l'état $S_{at} = s_{at}$ pour l'individu a au temps t,

$$\begin{aligned} \mu_{ats_{at}} &= \mathrm{P}(Y_{at} = 1 | S_{at} = s_{at}) = \mathrm{P}(2 \text{ UC } | S_{at} = s_{at}) \\ &= 1 - \mathrm{P}(Y_{at} = 0 | S_{at} = s_{at}) = 1 - \mathrm{P}(1 \text{ UC } | S_{at} = s_{at}). \end{aligned}$$

Le diagramme en bâton 3.4 nous fournit une représentation de la fréquence annuelle du nombre d'UC. Nous pouvons constater au vu des données observées que le phénomène de bicyclisme (2 UC) se produit principalement au début de la vie de l'arbre (de 1981 à 1984).

Une combinaison semi-markovienne de modèles linéaires généralisés, de type "gauche-droite" à 2 états (représentant chacun une phase de croissance) composée d'un état transitoire suivi d'un état final absorbant, a été estimée sur la base des 30 pins Laricio de 17 ans (cf remarque 2.2.7). Le changement d'état semble se produire en 1985 (figure 3.4). Nous sommes dans un cas que l'on peut qualifier d'"extrême". En effet, la probabilité d'observer 2 unités de croissance par pousse annuelle sur la première phase de croissance est faible (≈ 0.15) et est quasiment nulle sur la deuxième phase de croissance.

Dans les régions tempérées, les précipitations peuvent avoir soit un effet retard d'une année (sur le nombre d'éléments), soit un effet immédiat (sur l'allongement des pousses) selon qu'elles surviennent au cours de l'organogénèse (définition 1.1) ou de l'allongement (définition 1.2). Les pins Laricio étant à croissance rythmique, la période d'organogénèse se situe l'année précédant la période d'allongement. Afin de couvrir chaque année la période d'organogénèse et la période d'allongement d'une pousse annuelle, nous avons choisi comme covariable climatique les précipitations cumulées (en mm) du début du mois de juin de l'année précédente jusqu'à la fin du mois de juin de l'année courante.

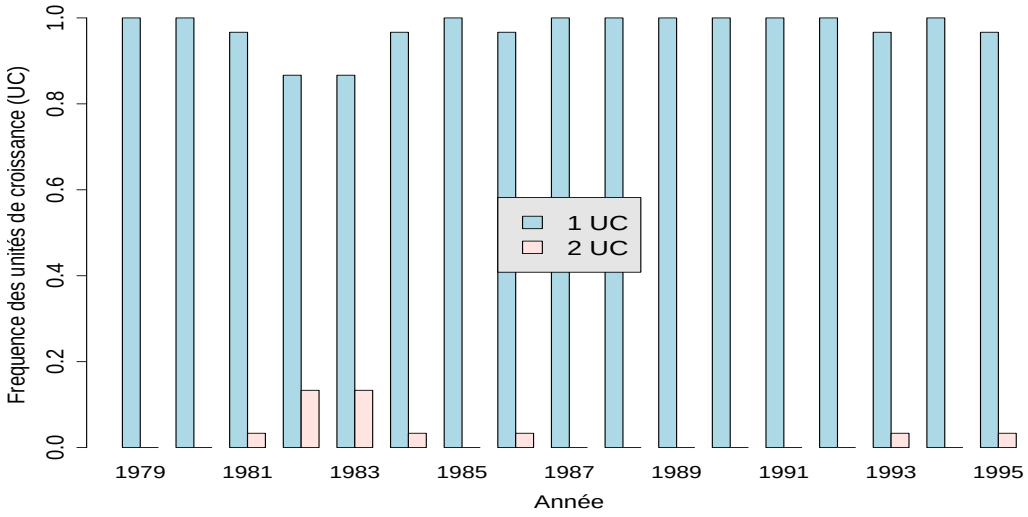

FIG. 3.4 – *Fréquence du nombre d'unités de croissance des 30 pins Laricios de 17 ans en fonction des années.*

Nous allons dans un premier temps définir le lien entre le processus d'observation et le processus d'état. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation Y_{at} d'un arbre a au temps t est modélisée par la fonction de masse de la loi de Bernoulli de paramètre $\mu_{ats_{at}}$:

$$\mathrm{P}(Y_{at} = y_{at}|S_{at} = s_{at}) = (\mu_{ats_{at}})^{y_{at}}(1 - \mu_{ats_{at}})^{1-y_{at}},$$

avec

$$\mathrm{logit}(\mu_{ats_{at}}) = \log \frac{\mu_{ats_{at}}}{1 - \mu_{ats_{at}}} = \beta_{s_{at}1} + X_t\beta_{s_{at}2}.$$

où $\beta_{s_{at}1}$ représente la constante, X_t représente la pluie cumulée de la pousse annuelle de l'année t, $\beta_{s_{at}2}$ représente le paramètre de régression et Y_{at} est à valeur dans le domaine $\{0; 1\}$. Comme ces arbres proviennent de la même forêt, la covariable climatique X_t est supposée commune à tous les individus chaque année.

Les paramètres estimés de la semi-chaîne de Markov sous-jacente, obtenus à partir de l'algorithme du gradient EM proposé, sont résumés dans le tableau 3.6. La plupart des pins Laricios ne possèdent qu'une unique phase de croissance ; la probabilité initiale de la seconde phase (état absorbant) est relativement élevée (≈ 0.837). Les arbres caractérisés par deux phases de croissance successives passent en moyenne 5.97 années dans la première phase avant d'atteindre la seconde phase. La valeur des paramètres de régression associés à la covariable climatique semble indiquer que les précipitations cumulées ont plus d'influence sur le nombre d'UC par pousse annuelle sur la première phase de croissance que sur la seconde phase de croissance.

À partir des estimations des paramètres, la séquence d'états la plus probable est calculée pour chaque séquence observée à partir de l'algorithme de Viterbi (cf section 2.2.5

		état 1	état 2
semi-chaîne de Markov sous-jacente	probabilité initiale π_j	0.163	0.837
	loi d'occupation $d_j(u)$	B(2, 9, 0.57)	
	moyenne, écart-type	5.97, 1.31	
modèle linéaire généralisé	constante β_{j1}	−13.015 (5.20)	−6.875 (2.45)
	paramètre de regression β_{j2}	0.017 (0.007)	0.003 (0.003)
probabilité moyenne d'avoir 2 UC	$\frac{1}{N}\sum_a \frac{1}{T_a}\sum_{t=1}^{T_a} P(2 \text{ UC}\|S_{at}=j)$	0.302	0.010

Tab. 3.6 – *Paramètres estimés de la combinaison semi-markovienne de modèles linéaires généralisé : probabilités initiales, loi d'occupation et modèles linéaires généralisés. Pour chaque modèle linéaire généralisé, la probabilité moyenne d'avoir 2 UC par pousse annuelle est donnée. La précision (écart-type) des estimations des paramètres de régression est fournie entre parenthèses.*

et Guédon (2007)) adapté aux combinaisons semi-markoviennes de modèles linéaires généralisés. La séquence d'états la plus probable restaurée peut être vue comme la segmentation qui explique au mieux la séquence observée pour un modèle donné. À partir de ces séquences d'états, nous avons calculé empiriquement les probabilités d'observation $P(2 \text{ UC}|S_{at} = j)$ pour chaque arbre a au temps t. Les deux phases de croissance se distinguent principalement par la probabilité du nombre d'UC par pousse annuelle. En effet, la probabilité moyenne d'avoir 2 UC sur la première phase de croissance (état 1) est plus forte que la probabilité moyenne d'avoir 2 UC sur la seconde phase de croissance (état 2) ; voir le tableau 3.6.

Les fréquences observées du nombre d'UC par pousse annuelle sont comparées aux fréquences du nombre d'UC par pousse annuelle prédites à partir de la séquence d'états la plus probable pour chaque arbre ; voir la figure 3.5. Du fait du caractère "extrême" de la variable "nombre d'UC par pousse annuelle", les résultats obtenus sont encourageants et montrent bien que les événements de 1981 et 1983 sont liées aux conditions pluviométriques durant une période recouvrant l'organogénèse et l'allongement de la pousse. Nous pouvons noter une certaine robustesse de notre approche basée sur l'algorithme du gradient EM.

3.5 Conclusion et discussion

Nous avons proposé des algorithmes de type gradient EM ou MCEM pour estimer les paramètres des MS-GLM et des SMS-GLM. Ces algorithmes peuvent être qualifiés d'algorithmes de restauration-maximisation (Archer et Titterington, 2002). L'étape de res-

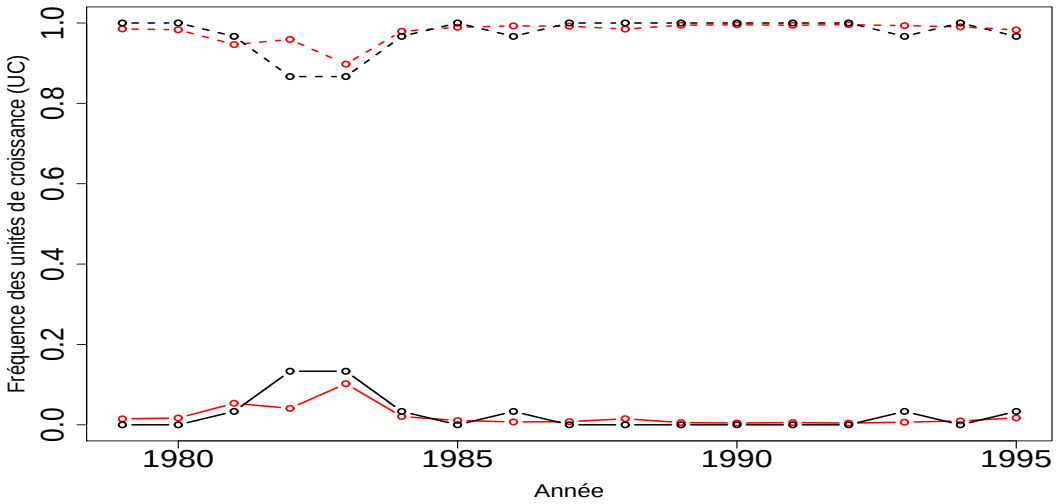

FIG. 3.5 – *Fréquences observées et fréquences prédites du nombre d'unités de croissance en fonction des années. Les pointillés en noir et les pointillés en rouge correspondent respectivement à la fréquence prédite et à la fréquence observée de 1 unité de croissance. Le trait noir et le trait rouge correspondent respectivement à la fréquence observée et à la fréquence calculée de 2 unités de croissance.*

tauration consiste soit en une restauration probabiliste de toutes les séquences d'états possibles par l'algorithme " avant-arrière", soit en une restauration par simulation de séquences d'états par l'algorithme "avant-arrière" de simulation. Pour les paramètres du processus (semi-)markovien sous-jacent, l'étape de maximisation repose soit sur les probabilités lissées soit sur les comptages extraits des séquences d'états simulées. Les paramètres des modèles linéaires généralisés sont obtenus par les équations des scores de Fisher. Ces équations sont quasi identiques à celles des modèles linéaires généralisés classiques hormis que les matrices hessiennes et les gradients sont pondérés soit par les probabilités lissées, soit par les comptages extraits des séquences d'états simulées.

La maximisation des paramètres des GLM se fait par les équations des scores de Fisher ou par les équations de Newton-Raphson. Cependant, d'autres approches numériques peuvent être envisagées telles que les méthodes de quasi-Newton ou du gradient conjugué (Lange, 2004).

Nous avons présenté le formalisme dans le cas des données binaires avec le lien logit et dans le cas des données poissonniennes avec le lien log. La transposition au cas de données catégorielles, ordinales ou exponentielles est naturelle. En effet, la matrice de poids W_j pour l'état j est identique à la matrice de poids d'un modèle linéaire généralisé classique.

Chapitre 4

Combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes (MS-LMM/SMS-LMM)

Ce chapitre est consacré à la famille des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes. Cette famille de modèles permet de modéliser des données de type séquence ou série chronologique présentant les caractéristiques suivantes :

- les données observées sont structurées en phases successives, asynchrones entre individus,
- les données observées sont influencées par des covariables pouvant varier dans le temps et pouvant être communes aux individus,
- les données observées présentent une hétérogénéité inter-individuelle pouvant être modulée sur les différentes phases.

Deux modélisations sont envisagées :

- un modèle linéaire mixte est associé à chaque état de la (semi-)chaîne de Markov sous-jacente avec des effets fixes modélisant des covariables pouvant varier dans le temps et un effet aléatoire "individuel" traduisant l'hétérogénéité inter-individuelle,
- un modèle linéaire mixte est associé à chaque état de la (semi-)chaîne de Markov sous-jacente avec un effet aléatoire "temporel" traduisant l'environnement commun à tous les individus à chaque date.

La deuxième modélisation trouve son intérêt lorsque les covariables variant dans le temps et communes à tous les individus ne sont pas disponibles. C'est souvent le cas pour les jeux de données de croissance de plantes. Comme les mesures se font rétrospectivement et sur des périodes pouvant aller jusqu'à 100 ans, les données concernant l'environnement commun à tous les arbres chaque année (pluviométrie, température, évapotranspiration)

sont rarement disponibles sur de longues périodes (section 1.1, paragraphe protocole de mesure).

Dans une première partie, nous présentons de manière générale la famille des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes. Les méthodes d'estimation existantes dans la littérature sont présentées dans une deuxième partie. Dans une troisième partie, nous traitons du cas d'effets aléatoires modélisant l'hétérogénéité inter-individuelle et proposons des méthodes d'estimation basées sur l'algorithme MCEM. La dernière partie est consacrée au cas d'effets aléatoires modélisant l'environnement commun et aux méthodes d'estimation proposées. Des simulations et des applications sur données de croissance d'arbres complètent ces différentes parties.

4.1 Définitions

Dans le chapitre précédent, nous avons pris en compte l'influence de covariables climatiques comme effet fixes dans les processus d'observation. Une extension possible de tels modèles est l'introduction d'effets aléatoires (cf section 2.4) afin de distinguer et de caractériser les différentes sources de variation. Dans la littérature, les chaînes de Markov cachées avec des effets aléatoires dans le processus d'observation sont étudiées depuis peu d'années. Véra (2004) a introduit les combinaisons markoviennes de modèles linéaires mixtes pour l'analyse des composantes de la croissance des arbres. Ces modèles combinent les modèles linéaires mixtes de manière markovienne et étendent la famille des combinaisons markoviennes de modèles linéaires en incorporant, dans le processus d'observation, des effets aléatoires "individuels" modélisant l'hétérogénéité inter-individuelle. Ces modèles peuvent être vus comme des mélanges finis de modèles linéaires mixtes avec dépendances markoviennes. Altman (2007) a introduit les combinaisons markoviennes de modèles linéaires généralisés mixtes (MS-GLMM, Markov Switching Generalized Linear Mixed Model) où le processus d'observation est supposé, conditionnellement aux effets aléatoires, appartenir à la famille exponentielle, et a appliqué ces modèles aux données de comptage de lésions dans le cerveau pour les patients atteints de scléroses en plaques. Les comptages de lésions sont supposées suivre une loi de Poisson dont la moyenne est supposée dépendante de l'état non-observée du patient. Les MS-GLMM ont été également utilisés pour analyser le lien entre le nombre de tumeurs primaires ou métastatiques dans le cerveau et certains symptômes en prenant en compte l'hétérogénéité inter-individuelle (Rijmen et al., 2008).

Kim et Smyth (2006) ont étendu les combinaisons markoviennes de modèles linéaires mixtes au cas semi-markovien en introduisant les combinaisons semi-markoviennes de modèles linéaires mixtes de type "gauche-droite" avec un effet aléatoire associé à chaque état et modélisant l'hétérogénéité inter-individuelle. Ils ont utilisé ces modèles en traitement du signal pour analyser les formes d'ondes.

Nous allons dans cette section donner une définition générale des combinaisons markoviennes de modèles linéaires mixtes et des combinaisons semi-markoviennes de modèles linéaires mixtes. Ce chapitre s'appuie sur les notations introduites dans la section 3.1.

Définition 4.1 *Une* ***combinaison markovienne de modèles linéaires mixtes*** *(MS-LMM, Markov Switching Linear Mixed Model) se caractérise par un couple de processus stochastiques* $\{S_{at}, Y_{at}; a = 1, \ldots, N,\ t = 1, \ldots, T_a\}$ *combinant :*

- *une chaîne de Markov sous-jacente* $\{S_{at}, t = 1, \ldots, T_a\}$ *d'ordre 1, homogène dans le temps et à valeurs dans l'espace d'états fini* $\{1, \ldots, J\}$,
- *un processus d'observation* $\{Y_{at}, t = 1, \ldots, T_a\}$ *où chaque observation* y_{at} *est liée au processus d'état* S_{at} *par un modèle linéaire mixte.*

Comme l'influence des covariables et les effets aléatoires sont uniquement pris en compte dans le processus d'observation, la généralisation aux combinaisons semi-markoviennes de modèles linéaires mixtes est directe. Une combinaison semi-markovienne de modèles linéaires mixtes peut être vue comme un mélange fini de modèles linéaires mixtes avec dépendances semi-markoviennes.

Définition 4.2 *Une* ***combinaison semi-markovienne de modèles linéaires mixtes*** *(SMS-LMM, Semi-Markov Switching Linear Mixed Model) se caractérise par un couple de processus stochastiques* $\{S_{at}, Y_{at}; a = 1, \ldots, N,\ t = 1, \ldots, T_a\}$ *combinant :*

- *une semi-chaîne de Markov sous-jacente* $\{S_{at}, t = 1, \ldots, T_a\}$ *homogène dans le temps et à valeurs dans l'espace d'états fini* $\{1, \ldots, J\}$,
- *un processus d'observation* $\{Y_{at}, t = 1, \ldots, T_a\}$ *où chaque observation* y_{at} *est liée au processus d'état* S_{at} *par un modèle linéaire mixte.*

Nous pouvons prendre un exemple pour éclaircir les idées. En analyse de croissance d'arbres forestiers, la chaîne de Markov sous-jacente représente la succession des phases tandis le modèle linéaire mixte associé à chaque état de la chaîne de Markov sous-jacente représente, dans la phase correspondante, l'influence de covariables climatiques et l'hétérogénéité inter-individuelle (Véra, 2004).

Dans la suite de ce chapitre, nous allons étudier deux cas particuliers de MS-LMM et de SMS-LMM où l'effet aléatoire modélise soit l'hétérogénéité inter-individuelle, soit l'environnement commun à tous les individus à chaque date. Nous verrons que bien que ces modèles soient proches, les difficultés d'estimation sont différentes. Cependant, avant de discuter de ces cas particuliers, nous allons présenter les méthodes d'estimation proposées dans la littérature et les hypothèses faites dans chaque article.

4.2 Méthodes d'estimation proposées dans la littérature

L'estimation des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes est un problème difficile dans la mesure où cela nécessite de prendre en compte, non plus une unique structure cachée mais deux types de structures cachées : les états du processus markovien sous-jacent d'une part, et les effets aléatoires des modèles linéaires mixtes d'autre part.

Véra (2004) a étudié un algorithme itératif de type " restauration-maximisation-prédiction" pour estimer les paramètres des MS-LMM avec des effets aléatoires "individuels" pouvant être associés à chaque état. L'étape de restauration des séquences d'états sachant les effets aléatoires repose sur une restauration déterministe par l'algorithme de Viterbi (Forney, 1973). L'étape de prédiction des effets aléatoires sachant les séquences d'états repose pour chaque individu sur la séquence d'états la plus probable restaurée de manière déterministe. L'approche déterministe n'est adaptée que dans le cadre de chaîne de Markov sous-jacente de type "gauche-droite" où la séquence d'états la plus probable concentre une bonne partie de la vraisemblance de toutes les séquences d'états possibles.

Les quelques travaux sur les MS-GLMM (Altman, 2007; Rijmen et al., 2008) reposent sur l'hypothèse que les effets aléatoires "individuels" sont indépendants des états ; c'est-à-dire qu'un individu est supposé avoir toujours le même comportement par rapport à l'individu moyen (meilleur ou moins bon). Altman (2007) a proposé une méthode déterministe et une méthode stochastique pour estimer les paramètres des MS-GLMM. L'approche déterministe repose sur une combinaison de méthodes d'intégration numérique de type quadrature de Gauss et de méthodes de quasi-Newton sous l'hypothèse que la vraisemblance d'une chaîne de Markov cachée peut s'écrire sous la forme d'un produit de matrices (cf section 2.2.4.1). Cependant, comme la vraisemblance d'une semi-chaîne de Markov cachée ne peut pas être écrite sous la forme d'un produit de matrices, cette méthode déterministe ne peut pas être transposée au cas semi-markovien. L'approche stochastique est fondée sur l'algorithme MCEM (cf section 2.1.3.2) où l'étape de maximisation se fait à l'aide de méthodes de quasi-Newton. Altman a souligné les limites des deux méthodes proposées : une sensibilité aux valeurs initiales, une convergence lente et un fort coût en calculs. Comme les relations d'indépendance conditionnelle d'une combinaison markovienne de modèles linéaires généralisés mixtes peuvent être représentées sous forme de graphe orienté acyclique, Rijmen et al. (2008) ont proposé de réaliser l'étape E de l'algorithme EM par l'algorithme d'arbre de jonction (Smyth et al., 1997). L'algorithme d'arbre de jonction proposée repose sur deux étapes : transformation du graphe orienté acyclique représentant le MS-GLMM en arbre de jonction (Cowell et al., 1999), application d'un algorithme de propagation pour obtenir les probabilités *a posteriori* des variables latentes ; cf Rijmen et al. (2007) pour une description technique détaillée et pour

des références. L'étape de maximisation nécessite l'utilisation d'algorithmes des scores de Fisher et de méthodes d'intégration numérique de type quadrature de Gauss. L'algorithme d'arbre de jonction possède quelques inconvénients. D'une part, les calculs ne sont pas effectués à partir de probabilités mais à partir de potentiels de cliques, ce qui a déjà l'inconvénient de ne pas prendre en compte les paramètres naturels du modèle (à savoir les probabilités de transition dans le cas d'une structure orienté acyclique). Du coup, il est difficile d'interpréter les calculs et les quantités intermédiaires de l'algorithme et de réutiliser les calculs déjà effectués. Il est également extrêmement délicat d'utiliser ces formules pour faire du calcul analytique afin de déterminer, par exemple, l'espérance de variables aléatoires en fonction de paramètres. D'autre part, le principe de l'algorithme d'arbre de jonction décompose la loi jointe des variables observées, ceci cause des instabilités numériques dès que le nombre de variables aléatoires du modèle est modérément grand. De plus, comme les relations d'indépendance conditionnelle d'une combinaison semi-markovienne de modèles linéaires mixtes ne peuvent pas être efficacement représentées par un graphe orienté acyclique, cette méthode ne peut pas être transposée au cas semi-markovien.

Kim et Smyth (2006) ont proposé une méthode pour estimer les paramètres des SMS-LMM de type "gauche-droite" avec un effet aléatoire modélisant l'hétérogénéité inter-individuelle différent pour chaque état ; c'est-à-dire qu'un individu peut avoir un comportement différent par rapport à l'individu moyen d'un état à un autre. La structure de type "gauche-droite" entraîne un ordre sur les états et chaque état ne peut être visité au maximum qu'une fois. La méthode proposée, qui est en fait une application de l'algorithme EM avec une étape E basée sur l'algorithme "avant-arrière", repose fortement sur les deux hypothèses spécifiques au modèle (état visité au maximum une fois et un effet aléatoire différent pour chaque état). La complexité de l'algorithme est cubique en les longueurs des séquences car l'approche proposée nécessite le calcul dans la récurrence arrière et dans la récurrence avant, de la distribution marginale des observations pour chaque individu et pour chaque temps de séjour possible dans chaque état.

Aucun travail dans la littérature n'est pour l'instant consacré aux MS-LMM et aux SMS-LMM avec un effet aléatoire "temporel". Nous pouvons cependant évoquer les travaux de Picard et al. (2007) qui s'intéressent à la modélisation de cet effet aléatoire en détection de ruptures multiples où un modèle linéaire mixte est associé à chaque segment des séquences observées ; les modèles de détection de ruptures pouvant être vus comme des semi-chaînes de Markov cachées (Fearnhead, 2008).

Un des objectifs de cette thèse est de développer des algorithmes d'inférence sous la contrainte qu'ils tiennent compte de la paramétrisation choisie, qu'ils se transposent facilement d'un processus markovien sous-jacent à un autre, qu'ils soient stables numériquement et qu'ils soient relativement peu coûteux en calculs.

4.3 Effet aléatoire "individuel"

Le caractère répétitif des données mesurées pour un même individu nécessite d'introduire dans la modélisation deux niveaux de lecture du comportement des individus : un niveau global traduit par les effets fixes et un niveau individuel traduit par les effets aléatoires. Cela vient principalement du fait que l'ensemble des observations relatives à un individu est généralement corrélé.

4.3.1 Modèle d'observation

Nous allons dans un premier temps définir les modèles linéaires mixtes qui lient le processus d'observation et le processus d'état. Nous proposons deux familles de MS-LMM et SMS-LMM qui différent par l'hypothèse faite sur l'hétérogénéité inter-individuelle dans le processus d'observation :

- un unique effet aléatoire pour toute la séquence observée. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation y_{at} d'un individu a au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=s_{at}} = X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_a + \epsilon_{at}, \tag{4.1}$$

$$\xi_a \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=s_{at}} \sim \mathcal{N}(0, \sigma^2_{s_{at}}).$$

  Le statut d'un individu (par rapport à l'individu moyen) à l'intérieur de la population est commun à tous les états non-observables. Cette modélisation correspond par exemple à un arbre dont la croissance est meilleure sur toute sa période de croissance mais modulée sur les différents phases de croissance.

- un effet aléatoire différent dans chaque état. Conditionnellement à l'état $S_{at} = s_{at}$, l'observation y_{at} d'un individu a au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=s_{at}} = X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_{as_{at}} + \epsilon_{at}, \tag{4.2}$$

$$\xi_{as_{at}} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=s_{at}} \sim \mathcal{N}(0, \sigma^2_{s_{at}}).$$

  Le statut d'un individu peut être différent pour chaque état non-observable. Nous pouvons prendre pour exemple un arbre dont le comportement est totalement différent d'une phase de croissance à une autre.

Dans ces définitions, X_{at} est le vecteur ligne de taille Q des covariables pour l'individu a au temps t, et ξ_a est l'effet aléatoire pour l'individu a. Sachant l'état $S_{at} = s_{at}$, $\beta_{s_{at}}$ est le vecteur colonne de taille Q des paramètres d'effets fixes, $\xi_{as_{at}}$ est l'effet aléatoire pour l'individu a, $\tau_{s_{at}}$ est l'écart-type de l'effet aléatoire et $\sigma^2_{s_{at}}$ est la variance résiduelle. Par commodité, les effets aléatoires sont supposés suivre la loi gaussienne $\mathcal{N}(0,1)$.

Les individus sont supposés indépendants. Dans le modèle linéaire mixte (4.2) avec un effet aléatoire différent dans chaque état, les effets aléatoires pour un individu a sont supposés indépendants entre les états non-observables ($\text{cov}(\xi_{aj}, \xi_{aj'}) = 0; j \neq j'$). Dans le cas des combinaisons markoviennes de modèles linéaires mixtes, l'introduction d'effets aléatoires modélisant l'hétérogénéité inter-individuelle invalide l'hypothèse selon laquelle les observations successives pour un individu sont conditionnellement indépendantes sachant les états non-observables. Les observations successives pour un individu sont dans notre cas indépendantes sachant les états non-observables et les effets aléatoires. Le raisonnement est similaire dans le cas des combinaisons semi-markoviennes de modèles linéaires mixtes. L'introduction d'un effet aléatoire permet de décomposer dans l'état j la variabilité totale γ_j^2 d'un individu à un temps donné en deux termes : la variabilité liée à l'hétérogénéité inter-individuelle τ_j^2 et la variabilité résiduelle σ_j^2.

Dans la suite de cette section, on se place dans le cadre de la famille des combinaisons markoviennes de modèles linéaires mixtes. La vraisemblance et les méthodes d'estimation sont présentées dans le cas des combinaisons markoviennes de modèles linéaires mixtes avec un effet aléatoire différent pour chaque état (modèle 4.2). Un paragraphe sera consacré à la transposition aux MS-LMM avec un unique effet aléatoire pour toute la séquence observée (modèle 4.1). Comme les MS-LMM et SMS-LMM ne différent que dans le type du processus sous-jacent, la transposition au cas des combinaisons semi-markoviennes de modèles linéaires mixtes est directe.

4.3.2 Vraisemblance du MS-LMM

Les paramètres du MS-LMM peuvent être scindés en deux sous-ensembles : les paramètres $(\pi_j; j = 1, \ldots, J)$ et $(p_{ij}; i, j = 1, \ldots, J)$ de la chaîne de Markov sous-jacente et les paramètres $(\beta_j; j = 1, \ldots, J)$, $(\tau_j; j = 1, \ldots, J)$ et $(\sigma_j^2; j = 1, \ldots, J)$ des J modèles linéaires mixtes. Nous notons par $\theta = (\pi, P, \beta, \tau, \sigma^2)$ l'ensemble des paramètres à estimer.

Soit $\xi_{a1}^J = (\xi_{aj}; j = 1, \ldots, J)$ le vecteur de taille J des effets aléatoires pour l'individu a. La vraisemblance du vecteur de paramètres θ pour les données observées dans le modèle (4.2) s'écrit :

$$\begin{aligned} L(\theta) &= \prod_{a=1}^{N} \int_{\xi_{a1}^J} \left\{ \sum_{s_{a1}^{T_a}} f(s_{a1}^{T_a}, \xi_{a1}^J, y_{a1}^{T_a}; \theta) \right\} d\xi_{a1}^J \\ &= \prod_{a=1}^{N} \int_{\xi_{a1}^J} \left\{ \sum_{s_{a1}^{T_a}} f(s_{a1}^{T_a}; \theta) f(\xi_{a1}^J; \theta) f(y_{a1}^{T_a} | s_{a1}^{T_a}, \xi_{a1}^J; \theta) \right\} d\xi_{a1}^J, \end{aligned} \tag{4.3}$$

où $\sum_{s_{a1}^{T_a}}$ signifie “somme sur toutes les séquences d'états possibles de longueur T_a pour l'individu a”.

La maximisation directe de cette vraisemblance (équation (4.3)) est difficile. Aussi, comme à la fois les états de la chaîne de Markov sous-jacente et les effets aléatoires

ne sont pas observés, l'algorithme EM semble être le meilleur candidat pour estimer les paramètres des combinaisons markoviennes de modèles linéaires mixtes.

4.3.3 Algorithme EM pour MS-LMM, difficultés et propositions

Considérons la log-vraisemblance des données complètes où à la fois les observations y, les effets aléatoires ξ et les états s de la chaîne de Markov sous-jacente sont observés :

$$\begin{aligned}
\log f(s,\xi,y;\theta) &= \sum_{a=1}^{N} \log f(s_{a1}^{T_a}, \xi_{a1}^{J}, y_{a1}^{T_a};\theta) \\
&= \sum_{a=1}^{N} \Big\{ \log f(s_{a1}^{T_a};\theta) + \log f(\xi_{a1}^{J};\theta) + \log f(y_{a1}^{T_a}|s_{a1}^{T_a},\xi_{a1}^{J};\theta) \Big\} \\
&= \sum_{a=1}^{N} \Big\{ \log\big(\pi_{s_{a1}} \prod_{t=2}^{T_a} p_{s_{a,t-1}s_{at}}\big) + \log\big(\prod_{j=1}^{J} \phi(\xi_{aj};0,1)\big) \\
&\qquad + \log\big(\prod_{t=1}^{T_a} \phi(y_{at}; X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_{as_{at}}, \sigma^2_{s_{at}})\big) \Big\} \\
&= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1}=j\Big)\log\pi_j + \sum_{a=1}^{N}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} I\Big(s_{at}=j, s_{a,t-1}=i\Big)\log p_{ij} \\
&+ \sum_{a=1}^{N}\sum_{j=1}^{J} \log\phi(\xi_{aj};0,1) \\
&+ \sum_{a=1}^{N}\sum_{t=1}^{T_a}\sum_{j=1}^{J} I\Big(s_{at}=j\Big)\log\phi(y_{at}; X_{at}\beta_j + \tau_j\xi_{aj}, \sigma_j^2),
\end{aligned}$$

où $\phi(y;\mu,\sigma^2)$ est la densité de la distribution gaussienne de moyenne μ et de variance σ^2.

L'étape E de l'algorithme EM repose sur le calcul de l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées qui est dans le cas du MS-LMM :

$$\begin{aligned}
Q(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log f(s,\xi,y;\theta)|Y=y;\theta^{(k)}\Big) \\
Q(\theta|\theta^{(k)}) &= \sum_{a=1}^{N}\sum_{j=1}^{J} L_{aj}(1)\log\pi_j + \sum_{a=1}^{N}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} P\Big(S_{at}=j, S_{a,t-1}=i|Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)}\Big)\log p_{ij} \\
&- \sum_{a=1}^{N}\sum_{j=1}^{J} \frac{\mathrm{E}(\xi_{aj}^2|Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)})}{2} - \sum_{a=1}^{N}\frac{J+T_a}{2}\log 2\pi \qquad (4.4) \\
&- \sum_{a=1}^{N}\sum_{t=1}^{T_a}\sum_{j=1}^{J} L_{aj}(t)\Big(\log\sigma_j + \frac{\mathrm{E}\big((y_{at}-X_{at}\beta_j-\tau_j\xi_{aj})^2|S_{at}=j, Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)}\big)}{2\sigma_j^2}\Big),
\end{aligned}$$

avec les probabilités lissées $L_{aj}(t) = P(S_{at}=j|Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)})$.

Plusieurs difficultés apparaissent pour le calcul de la quantité (4.4) à l'étape E. La non-indépendance des données successives observées pour chaque individu sachant les séquences d'états empêche la décomposition des probabilités lissées $L_{aj}(t)$ dans l'algorithme “ avant-arrière” usuellement utilisé dans le cas des chaînes de Markov cachées classiques (équation (2.19)). De plus, l'algorithme EM pour mélanges finis de modèles linéaires mixtes (Celeux et al., 2005) ne peut pas être transposé car les quantités $\mathrm{E}(\xi_{aj}^2|Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)})$, $\mathrm{E}(\xi_{aj}|S_{at} = j, Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)})$ et $\mathrm{E}(\xi_{aj}^2|S_{at} = j, Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)})$ ne peuvent pas être calculées analytiquement pour chaque individu a au temps t (Véra, 2004).

Nous proposons d'utiliser l'algorithme MCEM, où les quantités calculées dans l'étape déterministe E sont approchées en utilisant des méthodes de Monte Carlo. Dans la suite de cette partie, nous adoptons le cadre des algorithmes de restauration-maximisation proposé par Archer et Titterington (2002), l'étape E pouvant être vue comme une étape de restauration des données non-observées. L'étape E de l'algorithme MCEM repose sur la restauration des séquences d'états et des effets aléatoires à partir de leur loi jointe sachant les données observées. Cependant, cette loi jointe est difficile à évaluer. C'est pourquoi, par analogie avec les travaux de Shi et Lee (2000) et selon le principe de l'échantillonneur de Gibbs (Robert et Casella, 2004), nous choisissons d'effectuer deux étapes de restauration conditionnelle dans l'étape E, une pour les effets aléatoires sachant les séquences d'états (et les données observées) et une pour les séquences d'états sachant les effets aléatoires (et les données observées). Cet algorithme de restauration-maximisation peut se décomposer à chaque itération k :

- Étape de restauration conditionnelle des séquences d'états : pour chaque individu a, on restaure $S_{a1}^{T_a}$ à partir de $P(S_{a1}^{T_a} = s_{a1}^{T_a}|\xi_{a1}^J, Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)})$,
- Étape de restauration conditionnelle des effets aléatoires : pour chaque individu a, on restaure ξ_{a1}^J à partir de $P(\xi_{a1}^J|S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)})$,
- Étape de maximisation : on maximise l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées (équation (4.4)).

La restauration des séquences d'états sachant les effets aléatoires peut se faire soit de manière probabiliste, soit par simulation. Sachant les séquences d'états, les effets aléatoires peuvent être restaurés soit de manière déterministe, soit par simulation. Nous pouvons envisager par conséquent quatre associations :

- Association 1 : restauration conditionnelle probabiliste des séquences d'états et restauration conditionnelle déterministe des effets aléatoires,
- Association 2 : restauration conditionnelle probabiliste des séquences d'états et restauration conditionnelle des effets aléatoires par simulation,
- Association 3 : restauration conditionnelle des séquences d'états par simulation et restauration conditionnelle déterministe des effets aléatoires,

- Association 4 : restauration conditionnelle des séquences d'états par simulation et restauration conditionnelle des effets aléatoires par simulation.

La restauration probabiliste des séquences d'états sachant les effets aléatoires consiste en une restauration implicite de l'ensemble des séquences d'états possibles sous forme de probabilités. Par opposition, sachant les effets aléatoires, des séquences d'états sont restaurées explicitement par simulation. Sachant les séquences d'états, les effets aléatoires peuvent être restaurés soit de manière déterministe, soit par simulation. La restauration déterministe repose sur le calcul de l'espérance *a posteriori* des effets aléatoires sachant les séquences d'états et les données observées. Cette restauration nécessite de connaître explicitement les séquences d'états. Par conséquent, toutes les associations de restauration conditionnelle ne donnent pas des solutions opératoires. En effet, contrairement aux trois autres associations, l'association 1 ne peut être envisagée.

Nous pouvons distinguer deux types de restauration par simulation des effets aléatoires sachant les séquences d'états selon le type de la restauration conditionnelle des séquences d'états. La connaissance explicite des séquences d'états restaurées par simulation permet de connaître explicitement les distributions des effets aléatoires sachant les séquences d'états et les données observées. Par suite, les effets aléatoires peuvent être directement simulés selon ces distributions. Par contre, lorsque les séquences d'états sont restaurées de manière probabiliste, ces distributions ne peuvent plus s'écrire explicitement. C'est pourquoi, dans ce cas, la restauration par simulation des effets aléatoires sachant les séquences d'états repose sur l'algorithme de Metropolis-Hastings (Gelman et al., 2004; Robert et Casella, 2004). Cette méthode MCMC (Monte Carlo par Chaînes de Markov) permet de résoudre les cas où les distributions *a posteriori* ne sont pas connues. Cet algorithme repose sur une méthode d'acceptation-rejet. Son avantage est qu'une connaissance limitée de la distribution à simuler est suffisante.

Nous proposons par conséquent trois algorithmes itératifs qui se différencient dans les méthodes de restauration des séquences d'états et des effets aléatoires à l'étape E de l'algorithme MCEM :

- **Étape E de simulation-prédiction** (Association 3)
  - une restauration conditionnelle par simulation des séquences d'états par une transposition directe de l'algorithme "avant-arrière" de simulation proposé par Chib (1996),
  - une restauration conditionnelle déterministe des effets aléatoires par le calcul des meilleures prédictions *a posteriori*.

- **Étape E de simulation-simulation** (Association 4)
  - une restauration conditionnelle par simulation des séquences d'états par une transposition directe de l'algorithme "avant-arrière" de simulation proposé par Chib (1996),

  - une restauration conditionnelle par simulation des effets aléatoires selon leurs distributions conditionnelles.

- **Étape E de restauration probabiliste-simulation** (Association 2)
  - une restauration conditionnelle probabiliste des séquences d'états par une transposition directe de l'algorithme “avant-arrière”,
  - une restauration conditionnelle par simulation des effets aléatoires par l'algorithme de Metropolis-Hastings (Gelman et al., 2004; Robert et Casella, 2004).

Afin de ne pas alourdir les écritures, nous omettrons $\theta^{(k)}$ dans les différents calculs.

4.3.4 Algorithme MCEM avec une étape E de simulation-prédiction

L'étape E de cet algorithme MCEM consiste en une approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées (équation (4.4)). Nous détaillons ici les deux étapes de restauration conditionnelle et l'étape de maximisation.

4.3.4.1 Algorithme “ avant-arrière” pour simuler des séquences d'états sachant les effets aléatoires

Pour chaque individu a, les séquences d'états sont simulées selon la distribution conditionnelle $P(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J})$.
Pour une combinaison markovienne de modèles linéaires mixtes, comme

$$\begin{aligned} P\big(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}\big) &= \Big\{ \prod_{t=1}^{T_a-1} P\big(S_{at} = s_{at} | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}\big) \Big\} \\ &\quad \times P\big(S_{aT_a} = s_{aT_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}\big), \end{aligned}$$

les distributions conditionnelles suivantes sont utilisées pour simuler les séquences d'états :

- état final (initialisation)

$$P\big(S_{aT_a} = s_{aT_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}\big),$$

- état précédent

$$P\big(S_{at} = s_{at} | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}\big).$$

L'algorithme avant-arrière pour simuler des séquences d'états sachant les effets aléatoires peut être décomposé en deux passes, une récurrence avant similaire à celle de l'algorithme avant-arrière pour chaîne de Markov cachée et une passe arrière pour simuler des séquences d'états (Chib, 1996).

Récurrence avant

Dans le cas du MS-LMM, la récurrence avant est initialisée pour $t = 1$ par :

$$\begin{aligned} F_{aj}(1) &= P\big(S_{a1} = j | Y_{a1} = y_{a1}, \xi_{a1}^J\big) \\ &= \frac{\phi(y_{a1}; X_{a1}\beta_j + \tau_j \xi_{aj}, \sigma_j^2)}{N_{a1}} \ \pi_j, \end{aligned}$$

où $N_{a1} = P(Y_{a1} = y_{a1} | \xi_{a1}^J)$ est le facteur de normalisation tel que :

$$N_{a1} = \sum_{j=1}^{J} P(S_{a1} = j, Y_{a1} = y_{a1} | \xi_{a1}^J) = \sum_{j=1}^{J} \phi(y_{a1}; X_{a1}\beta_j + \tau_j \xi_{aj}, \sigma_j^2)\pi_j.$$

Pour $t = 2, \ldots, T_a$, la récurrence avant est ensuite donnée par :

$$\begin{aligned} F_{aj}(t) &= P\big(S_{at} = j | Y_{a1}^t = y_{a1}^t, \xi_{a1}^J\big) \\ &= \frac{\phi(y_{at}; X_{at}\beta_j + \tau_j \xi_{aj}, \sigma_j^2)}{N_{at}} \sum_{i=1}^{J} p_{ij} F_{ai}(t-1). \end{aligned}$$

Le facteur de normalisation $N_{at} = P\big(Y_{at} = y_{at} | Y_{a1}^{t-1} = y_{a1}^{t-1}, \xi_{a1}^J\big)$ est obtenu directement durant la récurrence avant :

$$N_{at} = \sum_{j=1}^{J} P(S_{at} = j, Y_{at} = y_{at} | Y_{a1}^{t-1} = y_{a1}^{t-1}, \xi_{a1}^J) = \sum_{j=1}^{J} \phi(y_{at}; X_{at}\beta_j + \tau_j \xi_{aj}, \sigma_j^2) \sum_{i=1}^{J} p_{ij} F_{ai}(t-1).$$

La récurrence avant peut être utilisée pour calculer la log-vraisemblance des données observées sachant les effets aléatoires :

$$\begin{aligned} \log P(Y = y | \xi; \theta) &= \sum_{a=1}^{N} \left(\log P(Y_{a1} = y_{a1} | \xi_{a1}^J; \theta) + \sum_{t=2}^{T_a} \log P(Y_{at} = y_{at} | Y_{a1}^{t-1} = y_{a1}^{t-1}, \xi_{a1}^J) \right) \\ &= \sum_{a=1}^{N} \sum_{t=1}^{T_a} \log N_{at}. \end{aligned} \tag{4.5}$$

Passe arrière

La passe arrière est initialisée pour $t = T_a$ par :

$$P\big(S_{aT_a} = j | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) = F_{aj}(T_a).$$

L'état final s_{aT_a} est simulé selon les probabilités lissées

$$\Big(P\big(S_{aT_a} = j | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big); j = 1, \ldots, J \Big).$$

Pour $t = T_a - 1, \ldots, 1$, la passe arrière est donnée par :

$$P\big(S_{at} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) = \frac{p_{j s_{a,t+1}} F_{aj}(t)}{\sum_{i=1}^{J} p_{i s_{a,t+1}} F_{ai}(t)},$$

où les quantités $F_{aj}(t)$ sont directement extraites de la récurrence avant.
L'état s_{at} est simulé selon la distribution conditionnelle

$$\Big(P\big(S_{at} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big); j = 1, \ldots, J \Big).$$

4.3.4.2 Prédiction des effets aléatoires sachant les séquences d'états

Pour chaque séquence d'états simulée $s_{a1}^{T_a}$ pour l'individu a, le vecteur des prédictions des effets aléatoires attachés à l'individu a est donné par :

$$\xi_{a1}^{J} \;=\; \mathrm{E}\Big(\xi_{a1}^{J} | S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}\Big).$$

Comme

$$\xi_{a1}^{J} \sim \mathcal{N}_J(0_J,\ \mathrm{Id}_J)$$

et

$$Y_{a1}^{T_a}|_{S_{a1}^{T_a}=s_{a1}^{T_a}} \sim \mathcal{N}_{T_a}\Big(\sum_{j=1}^{J} I_{aj} X_a \beta_j,\ U_a \Omega^2 {U_a}' + \mathrm{Diag}\{U_a\sigma^2\}\Big),$$

la loi jointe des effets aléatoires et des données observées est :

$$\begin{pmatrix} \xi_{a1}^{J} \\ \\ Y_{a1}^{T_a} \end{pmatrix} |_{S_{a1}^{T_a}=s_{a1}^{T_a}} \sim \mathcal{N}_{T_a+J}\left(\begin{pmatrix} 0_J \\ \\ \sum_{j=1}^{J} I_{aj} X_a \beta_j \end{pmatrix}, \begin{pmatrix} \mathrm{Id}_J & \Omega {U_a}' \\ \\ U_a\Omega & U_a\Omega^2{U_a}' + \mathrm{Diag}\{U_a\sigma^2\} \end{pmatrix} \right),$$

où

- Id_J est la matrice identité de dimension $J \times J$,
- $\Omega = \mathrm{Diag}\{\tau_j; j = 1, \ldots, J\}$ est la matrice de dimension $J \times J$ des écart-types des effets aléatoires,
- U_a est la matrice de design de dimension $T_a \times J$ associée à la séquence d'états $s_{a1}^{T_a}$, composée de 1 et de 0 avec $\sum_j U_a(t,j) = 1$ et $\sum_t \sum_j U_a(t,j) = T_a$,
- $u_{at} = \Big(I(s_{at} = 1) \cdots I(s_{at} = J)\Big)$ est la $t^{ème}$ ligne de la matrice de design U_a,
- $\sigma^2 = (\sigma_1^2 \cdots \sigma_J^2)'$ est le vecteur de taille J des variances résiduelles,
- $\mathrm{Diag}\{U_a\sigma^2\}$ est la matrice diagonale de dimension $T_a \times T_a$ avec $\{u_{at}\sigma^2; t = 1, \ldots, T_a\}$ sur sa diagonale,
- $I_{aj} = \mathrm{Diag}\Big\{I(s_{at} = j),\ t = 1, \cdots, T_a\Big\}$ est une matrice diagonale de dimension $T_a \times T_a$,
- X_a est la matrice de dimension $T_a \times Q$ des covariables.

Par suite, on obtient :

$$\xi_{a1}^{J} = \mathrm{E}\Big(\xi_{a1}^{J} | S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}\Big) = \Omega {U_a}' \Big(U_a\Omega^2{U_a}' + \mathrm{Diag}\{U_a\sigma^2\}\Big)^{-1} \Big(y_{a1}^{T_a} - \sum_{j=1}^{J} I_{aj} X_a \beta_j\Big).$$

4.3.4.3 Étape de maximisation

Dans l'algorithme MCEM avec une étape E de simulation-prédiction, l'espérance de la log-vraisemblance des données complètes sachant les données observées (équation (4.4)) est approchée à l'itération k par :

$$\mathrm{E}\Big(\log f(s, \xi, y; \theta) | Y = y; \theta^{(k)}\Big) = \sum_{a=1}^{N} \mathrm{E}\Big(\log f(s_{a1}^{T_a}, \xi_{a1}^{J}, y_{a1}^{T_a}; \theta) | Y_{a1}^{T_a} = y_{a1}^{T_a}; \theta^{(k)}\Big)$$

$$
\begin{aligned}
&\approx \frac{1}{M_k}\sum_{a=1}^{N}\sum_{m=1}^{M_k}\log f(s_{a1}^{T_a(k)}(m),\xi_{a1}^{J(k)}(m),y_{a1}^{T_a};\theta^{(k)})\\
&\approx \frac{1}{M_k}\sum_{a=1}^{N}\sum_{m=1}^{M_k}\sum_{j=1}^{J} I\Big(s_{a1}^{(k)}(m)=j\Big)\log\pi_j^{(k)}\\
&+ \frac{1}{M_k}\sum_{a=1}^{N}\sum_{m=1}^{M_k}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} I\Big(s_{at}^{(k)}(m)=j, s_{a,t-1}^{(k)}(m)=i\Big)\log p_{ij}^{(k)}\\
&+ \frac{1}{M_k}\sum_{a=1}^{N}\sum_{m=1}^{M_k}\sum_{j=1}^{J}\log\phi(\xi_{aj}^{(k)}(m);0,1) \qquad (4.6)\\
&+ \frac{1}{M_k}\sum_{a=1}^{N}\sum_{m=1}^{M_k}\sum_{t=1}^{T_a}\sum_{j=1}^{J} I\Big(s_{at}^{(k)}(m)=j\Big)\log\phi(y_{at};X_{at}\beta_j^{(k)}+\tau_j^{(k)}\xi_{aj}^{(k)}(m),\sigma_j^{2(k)}),
\end{aligned}
$$

où M_k est le nombre de séquences d'états simulées à l'itération k, $s_{a1}^{T_a}(m)$ est la m^{ème} séquence d'états simulée pour l'individu a et $\xi_{a1}^{J}(m)$ sont les effets aléatoires associés.

À l'itération k, les ré-estimations des paramètres du MS-LMM sont obtenues en maximisant les différents termes de l'équation (4.6), chaque terme dépendant d'un sous-ensemble de θ.

Pour les paramètres de la chaîne de Markov cachée sous-jacente, nous obtenons :

– probabilités initiales

$$\pi_j^{(k+1)}=\frac{\sum_a\sum_m I\Big(s_{a1}^{(k)}(m)=j\Big)}{NM_k},$$

– probabilités de transition

$$p_{ij}^{(k+1)}=\frac{\sum_a\sum_m\sum_{t=2}^{T_a} I\Big(s_{at}^{(k)}(m)=j,\ s_{a,t-1}^{(k)}(m)=i\Big)}{\sum_a\sum_m\sum_{t=2}^{T_a} I\Big(s_{a,t-1}^{(k)}(m)=i\Big)}.$$

Pour les paramètres des J modèles linéaires mixtes, nous obtenons :

– paramètres d'effet fixe

$$\beta_j^{(k+1)}=\Big(\sum_a\sum_m X_a' I_{aj}^{(k)}(m)X_a\Big)^{-1}\Big(\sum_a\sum_m X_a' I_{aj}^{(k)}(m)\Big(y_{a1}^{T_a}-\tau_j^{(k)}\xi_{aj}^{(k)}(m)\Big)\Big),$$

– variances résiduelles

$$\sigma_j^{2(k+1)}=\frac{\sum_a\sum_m\Big(y_{a1}^{T_a}-X_a\beta_j^{(k)}-\tau_j^{(k)}\xi_{aj}^{(k)}(m)\Big)' I_{aj}^{(k)}(m)\Big(y_{a1}^{T_a}-X_a\beta_j^{(k)}-\tau_j^{(k)}\xi_{aj}^{(k)}(m)\Big)}{\sum_a\sum_m \mathrm{tr}\{I_{aj}^{(k)}(m)\}},$$

– écart-type des effets aléatoires

$$\tau_j^{(k+1)}=\frac{\sum_a\sum_m\sum_t I\Big(s_{at}^{(k)}(m)=j\Big)\xi_{aj}^{(k)}(m)\Big(y_{at}-X_{at}\beta_j^{(k)}\Big)}{\sum_a\sum_m\sum_t I\Big(s_{at}^{(k)}(m)=j\Big)\xi_{aj}^{2(k)}(m)},$$

où $I_{aj}(m) = \text{Diag}\Big\{I(s_{at}(m) = j),\ t = 1, \cdots, T_a\Big\}$ est une matrice diagonale de dimension $T_a \times T_a$.

La ré-estimation des paramètres des modèles linéaires mixtes est similaire à l'estimation des paramètres des modèles linéaires par la méthode des moindres carrés ordinaires. La différence vient du fait que les paramètres des modèles linéaires mixtes sont pondérés par le nombre de séquences d'états simulées et par le nombre d'occurrences de chaque état.

4.3.4.4 Transposition au cas d'un unique effet aléatoire pour toute la séquence observée

La transposition de cette méthode d'estimation aux combinaisons markoviennes de modèles linéaires mixtes avec un unique effet aléatoire pour toute la séquence observée (équation (4.1)) est directe. Comme l'effet aléatoire modélisant l'hétérogénéité inter-individuelle est intégré dans le processus d'observation, la principale différence concerne l'étape de prédiction des effets aléatoires sachant les séquences d'états. Dans l'étape de maximisation (section 4.3.4.3), la récurrence avant et la passe arrière de simulation (section 4.3.4.1), les J effets aléatoires ξ_{a1}^J sont remplacés par un unique effet aléatoire ξ_a. Avec les mêmes notations que dans la section 4.3.4.2, la prédiction de l'effet aléatoire pour chaque individu est donnée par :

$$\xi_a \ = \ \text{E}\Big(\xi_a | S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}\Big)$$

où $s_{a1}^{T_a}$ est la séquence d'états simulée pour l'individu a.

Comme

$$\xi_a \sim \mathcal{N}(0,\ 1)$$

et

$$Y_{a1}^{T_a}|_{S_{a1}^{T_a} = s_{a1}^{T_a}} \sim \mathcal{N}_{T_a}\Big(\sum_{j=1}^{J} I_{aj} X_a \beta_j,\ U_a \tau \tau' {U_a}' + \text{Diag}\{U_a \sigma^2\}\Big),$$

la loi jointe de l'effet aléatoire et des données observées est :

$$\begin{pmatrix} \xi_a \\ \\ Y_{a1}^{T_a} \end{pmatrix} |_{S_{a1}^{T_a} = s_{a1}^{T_a}} \sim \mathcal{N}_{T_a+1}\left(\begin{pmatrix} 0 \\ \\ \sum_{j=1}^{J} I_{aj} X_a \beta_j \end{pmatrix}, \begin{pmatrix} 1 & \tau' {U_a}' \\ \\ U_a \tau & U_a \tau \tau' {U_a}' + \text{Diag}\{U_a \sigma^2\} \end{pmatrix} \right),$$

où $\tau = (\tau_1 \dots \tau_J)'$ est le vecteur de taille J des écart-types de l'effet aléatoire.

Par suite, on obtient :

$$\xi_a = \text{E}\Big(\xi_a | S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}\Big) = \tau' {U_a}' \Big(U_a \tau \tau' {U_a}' + \text{Diag}\{U_a \sigma^2\}\Big)^{-1} \Big(y_{a1}^{T_a} - \sum_{j=1}^{J} I_{aj} X_a \beta_j\Big).$$

4.3.4.5 Remarques

Choix du nombre de séquences d'états à simuler

L'étape de restauration conditionnelle repose sur la restauration de plusieurs couples $(s_{a1}^{T_a}, \xi_{a1}^{J})$ pour chaque individu a. Dans le cadre de l'algorithme MCEM, il est recommandé d'augmenter ce nombre au fil des itérations ; voir dans la section 2.1.3.2 pour une discussion plus complète. Afin de contrôler le nombre de couples à chaque itération, nous proposons d'introduire une étape supplémentaire à l'algorithme MCEM à l'itération k :

- on simule M_k séquences d'états sachant les effets aléatoires pour chaque individu,
- on prédit les M_k effets aléatoires associées aux séquences d'états simulées pour chaque individu,
- on maximise l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées (équation (4.6)),
- on fixe M_{k+1} et on tire aléatoirement avec remise M_{k+1} effets aléatoires parmi les M_k prédictions calculées.

Afin d'augmenter le nombre de couples à chaque itération, il est nécessaire de choisir une fonction $f()$ telle que $M_{k+1} = f(M_k)$ et que le nombre de couples soit égal ou supérieur entre deux itérations successives. Un exemple serait d'utiliser une fonction strictement croissante suivie d'une fonction constante ; voir également la section 2.1.3.2 pour des exemples de fonction.

Dans le cas où le nombre de séquences d'états simulées est constant au fil des itérations, la dernière étape n'est pas utile.

Ordre des étapes

Comme discuté par Neal et Hinton (1998) dans le cas des mélanges gaussiens, comme les paramètres de la chaîne de Markov sous-jacente et des modèles linéaires mixtes forment des ensembles disjoints et influencent séparément l'approximation de la log-vraisemblance des données complètes (équation (4.6)), les paramètres de la chaîne de Markov sous-jacente peuvent être estimés après que les séquences d'états soient simulées et les paramètres des modèles linéaires mixtes peuvent être estimés après que les effets aléatoires soient prédits. Il semble logique de ré-estimer immédiatement les paramètres avant d'effectuer l'étape de restauration conditionnelle pour la prochaine variable non observée.

Convergence de l'algorithme

Plusieurs méthodes sont proposées pour contrôler la convergence de l'algorithme MCEM ; une description détaillée est donnée dans la section 2.1.3.2. Dans le cas des combinaisons markoviennes de modèles linéaires mixtes, ces méthodes sont difficilement applicables car la log-vraisemblance des données observées ne peut pas être calculées et la méthode du "bridge sampling" ne peut pas se transposer facilement. Aussi, sous conditions de convergence des prédictions des effets aléatoires, nous choisissons de contrôler la

convergence de l'algorithme MCEM proposé en observant les fluctuations stationnaires autour de 0 de

$$g^{(k+1)} = \log P(Y = y|\xi^{(k+1)}; \theta^{(k+1)}) - \log P(Y = y|\xi^{(k)}; \theta^{(k)}), \tag{4.7}$$

où la quantité $\log P(Y = y|\xi; \theta^{(k)})$ est directement obtenue dans la récurrence avant (équation (4.5)).

Initialisation de l'algorithme

De nombreuses simulations, menées avec des valeurs initiales différentes, ont conduit à la conclusion que l'algorithme MCEM proposé était sensible aux valeurs initiales. Plus les valeurs initiales sont éloignées des vrais valeurs, moins bonnes sont les estimations des paramètres du modèle. Nous recommandons de choisir comme valeurs initiales les paramètres estimés par l'algorithme EM pour une simple combinaison markovienne de modèles linéaires (en fixant donc $\xi = 0$).

Prédiction finale des effets aléatoires

Comme suggéré par Shi et Lee (2000) pour les modèles à variables latentes, une estimation des moyennes des effets aléatoires peut être obtenue à la dernière itération de l'algorithme en moyennant pour chaque individu les M_k prédictions. La médiane peut également être envisagée. En effet, bien que la moyenne soit proche de la médiane sous l'hypothèse gaussienne, la médiane est plus robuste.

4.3.4.6 Simulations

Afin d'évaluer la capacité de l'algorithme MCEM proposé avec une étape E de simulation-prédiction à estimer correctement les paramètres des MS-LMM pour des mélanges markoviens dont les états sont plus ou moins bien séparés, nous avons effectué des simulations. Nous avons considéré une combinaison markovienne de modèles linéaires mixtes à deux états dont les paramètres de la chaîne de Markov sous-jacente sont :

- probabilités initiales : $\pi = (0.9 \quad 0.1)$,
- probabilités de transition : $P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$.

Les variances aléatoires et les variances résiduelles sont respectivement pour l'état 1, $\tau_1^2 = 1$ and $\sigma_1^2 = 2$, et pour l'état 2, $\tau_2^2 = 2$ and $\sigma_2^2 = 1$. Nous avons considéré un unique effet fixe (intercept) par état et nous avons simulé 20 individus sur 30 temps selon deux modalités :

- Modalité (A) défini par $\beta_1 = 2$ et $\beta_2 = 9$, les états sont bien séparés.
- Modalité (B) défini par $\beta_1 = 4$ et $\beta_2 = 7$, les états sont moins bien séparés.

Le tableau 4.1 donne les moyennes et les écart-types entre parenthèses des estimations obtenues sur 100 jeux de données simulés pour chaque modèle :

	param.	vraie valeur	modèle 1 (A)	modèle 2 (A)	vraie valeur	modèle 1 (B)	modèle 2 (B)
	π_1	0.9	0.901 (0.061)	0.907 (0.063)	0.9	0.887 (0.106)	0.876 (0.123)
	p_{11}	0.8	0.801 (0.019)	0.797 (0.022)	0.8	0.805 (0.067)	0.810 (0.071)
	p_{22}	0.7	0.701 (0.030)	0.701 (0.031)	0.7	0.707 (0.091)	0.702 (0.102)
	β_1	2	2.017 (0.257)	2.058 (0.274)	4	3.949 (0.524)	4.114 (0.532)
état 1	τ_1^2	1	0.912 (0.322)	0.969 (0.348)	1	0.987 (0.393)	1.104 (0.514)
	σ_1^2	2	2.045 (0.197)	1.941 (0.166)	2	2.036 (0.503)	1.935 (0.628)
	β_2	9	9.036 (0.331)	8.988 (0.421)	7	6.788 (0.730)	6.912 (1.052)
état 2	τ_2^2	2	1.844 (0.627)	2.096 (0.733)	2	1.994 (0.856)	1.802 (0.941)
	σ_2^2	1	1.006 (0.099)	0.972 (0.095)	1	1.066 (0.674)	1.245 (1.414)

TAB. 4.1 – *Hétérogénéité inter-individuelle : résultats d'estimation des paramètres d'une combinaison markovienne de modèles linéaires mixtes à 2 états avec un unique effet aléatoire pour toute la séquence observée (modèle 1) ou avec un effet aléatoire différent pour chaque état (modèle 2) par l'algorithme MCEM avec une étape E de simulation-prédiction sur 100 jeux de données simulés pour le cas d'états bien séparés (A) et pour le cas d'états moins bien séparés (B) : moyennes (écart-types).*

– Modèle 1 : un unique effet aléatoire pour toute la séquence observée (équation (4.1)),
– Modèle 2 : un effet aléatoire différent pour chaque état (équation (4.2)),

et pour chaque modalité (degré de séparabilité des états). Chaque estimation est basée sur une augmentation du nombre de séquences d'états simulées pour chaque individu au cours des itérations de l'algorithme : k séquences d'états ont été simulées à la $k^{ème}$ itération.

Nous notons un bon comportement de l'algorithme MCEM proposé avec une étape E de simulation-prédiction que ce soit pour les paramètres π et P de la chaîne de Markov sous-jacente ou pour les paramètres β, σ^2 et τ^2 des modèles linéaires mixtes associés à chaque état. Les résultats obtenus pour le modèle 1 sont légèrement meilleurs que ceux obtenus pour le modèle 2. Nous pouvons aussi noter que plus les variances sont fortes, plus les écart-types des estimations sont élevés et moins les estimations des paramètres sont précises.

Le degré de séparabilité des états a une influence significative sur la qualité des estimations des paramètres quel que soit le modèle : les écart-types des valeurs estimées augmentent avec la diminution du degré de séparabilité des états. Cette influence est beaucoup plus prononcée dans le cadre du modèle avec un effet aléatoire différent pour chaque état. Nous pouvons noter que l'écart-type de τ_1^2 est plus petit que celui de τ_2^2, ce

qui semble tout à fait logique puisque vu les probabilités initiales et de transition choisies, le temps passé dans l'état 1 est nécessairement plus long que le temps passé dans l'état 2. Il y a donc plus de mesures répétées dans l'état 1 que dans l'état 2 pour prédire les effets aléatoires et les variances aléatoires.

Nous avons également étudié l'influence du nombre d'individus simulés et l'influence de la longueur des séquences simulées sur les estimations des paramètres des combinaisons markoviennes de modèles linéaires mixtes avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle. Les résultats obtenus ont sans surprise montré que plus le nombre d'individus est élevé et plus les séquences sont longues, meilleures sont les estimations.

4.3.5 Algorithme MCEM avec une étape E de simulation-simulation

L'étape E de cet algorithme MCEM repose sur deux étapes de restauration conditionnelle par simulation :

- sachant les effets aléatoires, les séquences d'états sont simulées pour chaque individu par l'algorithme “avant-arrière” décrit dans la section 4.3.4.1,
- sachant les séquences d'états, les effets aléatoires sont simulés selon leurs distributions conditionnelles pour chaque individu.

Les formules de ré-estimation des paramètres sont obtenues par maximisation de l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées (équation (4.6)).

L'algorithme MCEM avec une étape E de simulation-simulation ne diffère de l'algorithme MCEM avec une étape E de simulation-prédiction que dans la restauration conditionnelle des effets aléatoires sachant les séquences d'états et les données observées. Nous la décrivons ci-dessous.

Dans le cas d'un effet aléatoire différent pour chaque état (équation (4.2)), sachant une séquence d'états $S_{a1}^{T_a} = s_{a1}^{T_a}$, les effets aléatoires sont simulés selon la loi (cf section 4.3.4.2 pour la loi jointe des effets aléatoires et des données observées sachant une séquence d'états) :

$$\xi_{a1}^{J} | S_{a1}^{T_a} = s_{a1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a} \sim \mathcal{N}_J \Bigg(\Omega {U_a}' \Big(U_a \Omega^2 {U_a}' + \mathrm{Diag}\{U_a \sigma^2\} \Big)^{-1} \Big(y_{a1}^{T_a} - \sum_{j=1}^{J} I_{aj} X_a \beta_j \Big), \\ \mathrm{Id}_J - \Omega {U_a}' \Big(U_a \Omega^2 {U_a}' + \mathrm{Diag}\{U_a \sigma^2\} \Big)^{-1} U_a \Omega \Bigg).$$

Dans le cas d'un unique effet aléatoire pour toute la séquence observée (équation (4.1)), sachant une séquence d'états $S_{a1}^{T_a} = s_{a1}^{T_a}$, les effets aléatoires sont simulés selon la loi (cf

section 4.3.4.4 pour la loi jointe des effets aléatoires et des données observées sachant une séquence d'états) :

$$\xi_a|S_{a1}^{T_a}=s_{a1}^{T_a},Y_{a1}^{T_a}=y_{a1}^{T_a} \sim \mathcal{N}\Bigg(\tau' {U_a}'\Big(U_a\tau\tau' {U_a}'+\mathrm{Diag}\{U_a\sigma^2\}\Big)^{-1}\Big(y_{a1}^{T_a}-\sum_{j=1}^{J} I_{aj}X_a\beta_j\Big), 1-\tau' {U_a}'\Big(U_a\tau\tau' {U_a}'+\mathrm{Diag}\{U_a\sigma^2\}\Big)^{-1}U_a\tau\Bigg).$$

Cette méthode n'ayant pas encore été implémentée et testée, nous n'avons pu la comparer avec les résultats de simulation obtenus pour l'algorithme MCEM avec une étape E de simulation-prédiction (section 4.3.4.6).

4.3.6 Algorithme MCEM avec une étape E de restauration probabiliste-simulation

Dans l'étape E de cet algorithme MCEM, on s'intéresse à l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées et aux effets aléatoires modélisant l'hétérogénéité inter-individuelle :

$$\begin{aligned}
Q(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log f(s,\xi,y;\theta)|\tilde{\xi},Y=y;\theta^{(k)}\Big)\\
Q(\theta|\theta^{(k)}) &= \sum_{a=1}^{N}\sum_{j=1}^{J} L'_{aj}(1)\log\pi_j\\
&+ \sum_{a=1}^{N}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} P(s_{at}=j,s_{at-1}=i|\tilde{\xi}_{a1}^{J},Y_{a1}^{T_a}=y_{a1}^{T_a};\theta^{(k)})\log p_{ij}\\
&- \frac{NJ}{2}\log 2\pi-\sum_{a=1}^{N}\sum_{j=1}^{J}\frac{\tilde{\xi}_{aj}^2}{2}-\sum_{a=1}^{N}\frac{T_a}{2}\log 2\pi\\
&- \sum_{a=1}^{N}\sum_{t=1}^{T_a}\sum_{j=1}^{J} L'_{aj}(t)\Big(\frac{\log\sigma_j^2}{2}+\frac{(y_{at}-X_{at}\beta_j-\tau_j\tilde{\xi}_{aj})^2}{2\sigma_j^2}\Big)
\end{aligned} \tag{4.8}$$

avec les probabilités lissées $L'_{aj}(t)=P(S_{at}=j|Y_{a1}^{T_a}=y_{a1}^{T_a},\tilde{\xi}_{a1}^{J})$ et $\tilde{\xi}_{a1}^{J}=\mathrm{E}(\xi_{a1}^{J}|Y_{a1}^{T_a}=y_{a1}^{T_a})$.

4.3.6.1 Algorithme "avant-arrière" sachant les effets aléatoires

L'algorithme "avant-arrière" proposé par Devijver (1985) pour les chaînes de Markov cachées classiques repose pour les MS-LMM sur la décomposition suivante des probabilités lissées pour chaque individu a :

$$\begin{aligned}
L'_{aj}(t) &= P(S_{at}=j|Y_{a1}^{T_a}=y_{a1}^{T_a},\tilde{\xi}_{a1}^{J})\\
&= \frac{P(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|S_{at}=j,\tilde{\xi}_{a1}^{J})}{P(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|Y_{a1}^{t}=y_{a1}^{t},\tilde{\xi}_{a1}^{J})}P(S_{at}=j|Y_{a1}^{t}=y_{a1}^{t},\tilde{\xi}_{a1}^{J})\\
&= B'_{aj}(t)F'_{aj}(t).
\end{aligned}$$

Les quantités $F'_{aj}(t)$ peuvent être calculées à l'aide d'une récurrence avant (c'est-à-dire de 1 à T_a) tandis que les quantités $B'_{aj}(t)$ ou $L'_{aj}(t)$ peuvent être calculées par une récurrence arrière (c'est-à-dire de T_a à 1) pour chaque individu a.

Récurrence avant

Dans le cas du MS-LMM, la récurrence avant est initialisée pour $t = 1$ par :

$$\begin{aligned} F'_{aj}(1) &= P(S_{a1} = j | Y_{a1} = y_{a1}, \tilde{\xi}^J_{a1}) \\ &= \frac{\phi(y_{a1}; X_{a1}\beta_j + \tau_j \tilde{\xi}_{aj}, \sigma_j^2)}{N'_{a1}} \ \pi_j \ , \end{aligned}$$

où $N'_{a1} = P(Y_{a1} = y_{a1} | \tilde{\xi}^J_{a1})$ est le facteur de normalisation tel que

$$N'_{a1} = \sum_{j=1}^{J} P(S_{a1} = j, Y_{a1} = y_{a1} | \tilde{\xi}^J_{a1}) = \sum_{j=1}^{J} \phi(y_{a1}; X_{a1}\beta_j + \tau_j \tilde{\xi}_{aj}, \sigma_j^2)\pi_j.$$

Pour $t = 2, \ldots, T_a$, la récurrence avant est donnée par :

$$\begin{aligned} F'_{aj}(t) &= P(S_{at} = j | Y^t_{a1} = y^t_{a1}, \tilde{\xi}^J_{a1}) \\ &= \frac{\phi(y_{at}; X_{at}\beta_j + \tau_j \tilde{\xi}_{aj}, \sigma_j^2)}{N'_{at}} \sum_{i=1}^{J} p_{ij} F'_{ai}(t-1). \end{aligned}$$

Le facteur de normalisation $N'_{at} = P(Y_{at} = y_{at} | Y^{t-1}_{a1} = y^{t-1}_{a1}, \tilde{\xi}^J_{a1})$ est obtenu directement durant la récurrence avant :

$$\begin{aligned} N'_{at} &= \sum_{j=1}^{J} P(S_{at} = j, Y_{at} = y_{at} | Y^{t-1}_{a1} = y^{t-1}_{a1}, \tilde{\xi}^J_{a1}) \\ &= \sum_{j=1}^{J} \phi(y_{at}; X_{at}\beta_j + \tau_j \tilde{\xi}_{aj}, \sigma_j^2) \sum_{i=1}^{J} p_{ij} F'_{ai}(t-1). \end{aligned}$$

La récurrence avant peut être utilisée pour calculer la log-vraisemblance des données observées sachant les effets aléatoires :

$$\begin{aligned} \log P(Y = y | \tilde{\xi}; \theta) &= \sum_{a=1}^{N} \log P(Y^{T_a}_{a1} = y^{T_a}_{a1} | \tilde{\xi}^J_{a1}; \theta) \\ &= \sum_{a=1}^{N} \left(\log P(Y_{a1} = y_{a1} | \tilde{\xi}^J_{a1}; \theta) \sum_{t=2}^{T_a} \log P(Y_{at} = y_{at} | Y^{t-1}_{a1} = y^{t-1}_{a1}, \tilde{\xi}^J_{a1}; \theta) \right) \\ &= \sum_{a=1}^{N} \sum_{t=1}^{T_a} \log N'_{at}. \end{aligned} \tag{4.9}$$

Récurrence arrière

La récurrence arrière est initialisée pour $t = T_a$ par :

$$L'_{aj}(T_a) = P(S_{aT_a} = j|Y_{a1}^{T_a} = y_{a1}^{T_a}, \tilde{\xi}_{a1}^{J}) = F'_{aj}(T_a).$$

Pour $t = T_a - 1, \ldots, 1$, la récurrence arrière est donnée par :

$$\begin{aligned} L'_{aj}(t) &= P(S_{at} = j|Y_{a1}^{T_a} = y_{a1}^{T_a}, \tilde{\xi}_{a1}^{J}) \\ &= \left\{ \sum_k \frac{L'_{ak}(t+1)}{G'_{ak}(t+1)} p_{jk} \right\} F'_{aj}(t) \end{aligned}$$

où la quantité $G'_{ak}(t+1) = P(S_{a,t+1} = k|Y_{a1}^{t} = y_{a1}^{t}, \tilde{\xi}_{a1}^{J}) = \sum_j p_{jk} F'_{aj}(t)$ peut être directement extraite et stockée en mémoire durant la récurrence avant.

4.3.6.2 Simulation des effets aléatoires sachant les séquences d'états

Le vecteur des prédictions des effets aléatoires modélisant l'hétérogénéité inter-individuelle est donné pour l'individu a par (voir équation (4.8)) :

$$\tilde{\xi}_{a1}^{J} = \mathrm{E}\Big(\xi_{a1}^{J}|Y_{a1}^{T_a} = y_{1a}^{T_a}\Big)$$

Cette quantité ne peut pas être calculée explicitement car la distribution de $Y_{a1}^{T_a}|\xi_{a1}^{J}$ n'est pas connue. Cependant, cette espérance peut être approximée par des méthodes MCMC (Monte Carlo par Chaînes de Markov). Dans le cas où la loi *a posteriori* n'est pas explicitement connue, l'algorithme de Metropolis-Hastings est la méthode alternative la plus connue et la plus usuelle (Gelman et al., 2004; Robert et Casella, 2004). Cet algorithme repose sur une méthode d'acceptation-rejet de la valeur proposée et se définit de la manière suivante :

Algorithme de Metropolis-Hastings

À l'itération k, étant donné $x^{(k)}$,

1. Générer $z_k \sim q(z|x^{(k)})$.
2. Prendre

$$x^{(k+1)} = \begin{cases} z_k & \text{avec probabilité } \rho(x^{(k)}, z_k) \\ x^{(k)} & \text{avec probabilité } 1 - \rho(x^{(k)}, z_k), \end{cases}$$

où

$$\rho(x^{(k)}, z_k) = \min\left\{1, \frac{f(z_k)q(x^{(k)}|z_k)}{f(x^{(k)})q(z_k|x^{(k)})}\right\}. \tag{4.10}$$

La loi $q()$ est appelée loi instrumentale ou loi conditionnelle.

Pour appliquer l'algorithme de Metropolis Hastings, il est nécessaire de choisir une densité conditionnelle $q()$. Le choix de la loi $q()$ est assez "subjectif". Plusieurs possibilités sont proposées dans la littérature dont (Robert et Casella, 2004) :

– $q()$ est indépendante de l'événement $x^{(k)}$: on parle d'algorithme de Metropolis-Hastings indépendant ;

– ou encore $q(z|x^{(k)}) = x^{(k)} + \iota^{(k)}$ où $\iota^{(k)}$ est une perturbation aléatoire indépendante de $x^{(k)}$: on parle d'algorithme de Metropolis-Hastings à marche aléatoire.

Le choix de la densité conditionnelle est motivé à la fois par des problèmes de parcours de l'espace des simulations, de mélangeance et de simplification de la probabilité d'acceptation-rejet (équation (4.10)).

Algorithme de Metropolis-Hastings indépendant

L'algorithme de Metropolis-Hastings indépendant est usuellement utilisé pour simuler des effets aléatoires modélisant l'hétérogénéité inter-individuelle ; voir McCulloch (1997) et Booth et Hobert (1999) dans le cas des modèles linéaires généralisés mixtes ou encore Lavergne et al. (2007) dans le cas des mélanges de modèles linéaires généralisés mixtes. Tous ces auteurs ont choisi comme densité conditionnelle $q()$ la loi de l'effet aléatoire ; c'est-à-dire $q(\xi|x^{(k)}) = f(\xi)$. Pour les combinaisons markoviennes de modèles linéaires mixtes, comme les individus sont supposées indépendants, cette loi est la loi multivariée gaussienne $\mathcal{N}_J(0, Id_J)$ pour chaque individu. Choisir cette densité conditionnelle permet à la probabilité d'acceptation-rejet de prendre une forme plus simple et d'être plus facile à calculer.

En effet, nous cherchons à simuler les effets aléatoires pour l'individu a selon la loi de $\xi^J_{a1}|Y^{T_a}_{a1} = y^{T_a}_{a1}$. Soit ξ^{J*}_{a1} la proposition et soit ξ^J_{a1} l'ancien vecteur des réalisations. La probabilité d'acceptation-rejet (équation (4.10)) s'écrit alors :

$$\begin{aligned}
\frac{f(\xi^{J*}_{a1}|Y^{T_a}_{a1} = y^{T_a}_{a1})q(\xi^J_{a1}|\xi^{J*}_{a1})}{f(\xi^J_{a1}|Y^{T_a}_{a1} = y^{T_a}_{a1})q(\xi^{J*}_{a1}|\xi^J_{a1})} &= \frac{f(\xi^{J*}_{a1}|Y^{T_a}_{a1} = y^{T_a}_{a1})f(\xi^J_{a1})}{f(\xi^J_{a1}|Y^{T_a}_{a1} = y^{T_a}_{a1})f(\xi^{J*}_{a1})} \\
&= \frac{f(y^{T_a}_{a1}|\xi^{J*}_{a1})f(\xi^{J*}_{a1})f(\xi^J_{a1})}{f(y^{T_a}_{a1}|\xi^J_{a1})f(\xi^J_{a1})f(\xi^{J*}_{a1})} \\
&= \frac{f(y^{T_a}_{a1}|\xi^{J*}_{a1})}{f(y^{T_a}_{a1}|\xi^J_{a1})} \\
&= \frac{\prod_{t=1}^{T_a} N'^{*}_{at}}{\prod_{t=1}^{T_a} N'_{at}}.
\end{aligned}$$

En choisissant la distribution des effets aléatoires comme loi instrumentale $q()$, la probabilité d'acceptation-rejet est simplifiée car son calcul nécessite seulement la spécification de la distribution conditionnelle des données observées sachant les effets aléatoires. De plus, la probabilité d'acceptation-rejet se calcule d'autant plus facilement dans notre cas car elle dépend uniquement des facteurs de normalisation N'_{at} et N'^{*}_{at} calculés dans la passe avant de l'algorithme “avant-arrière” (section 4.3.6.1). Notre choix pour simuler les effets aléatoires se porte donc sur l'algorithme de Metropolis-Hastings indépendant.

Le vecteur des effets aléatoires modélisant l'hétérogénéité inter-individuelle est ainsi calculé pour l'individu a par :

$$\tilde{\xi}_{a1}^{J} = \mathrm{E}\Big(\xi_{a1}^{J} | Y_{a1}^{T_a} = y_{a1}^{T_a}\Big) \approx \frac{1}{R}\sum_{r=1}^{R} \xi_{a1}^{J}(r)$$

où R est le nombre de simulations des effets aléatoires par l'algorithme de Metropolis-Hastings indépendant.

Remarques :

1. *Il est important de noter que les séquences d'états restaurées conditionnent implicitement les simulations des effets aléatoires au travers des facteurs de normalisation. C'est pourquoi, nous pouvons parler de simulations des effets aléatoires sachant les séquences d'états.*
2. *Nous avons choisi de simuler les effets aléatoires selon leur distribution conditionnelle sachant les données observées (et implicitement les séquences d'états). Une alternative est d'approcher* $\tilde{\xi}_{a1}^{J}$ *par* $E\Big(\xi_{a1}^{J} | Y_{a1}^{T_a} = y_{1a}^{T_a}, S_{a1}^{T_a} = s_{a1}^{T_a}\Big)$. *Cependant, cette approche est beaucoup plus coûteuse car elle nécessite le calcul du ratio* $\Big(P(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J*}) f(y_{a1}^{T_a} | \xi_{a1}^{J*})\Big) / \Big(P(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^{J}) f(y_{a1}^{T_a} | \xi_{a1}^{J})\Big)$. *De plus, cette alternative nous a amené à des problèmes d'estimation que l'on suppose dus à des effets de bords.*

4.3.6.3 Étape de maximisation

À l'itération k, les ré-estimations des paramètres du MS-LMM sont obtenues en maximisant les différents termes de l'équation (4.8), chaque terme dépendant d'un sous-ensemble de θ.

Pour les paramètres de la chaîne de Markov sous-jacente, nous obtenons :

– probabilités initiales

$$\pi_j^{(k+1)} = \frac{\sum_a L_{aj}^{'(k)}(1)}{N},$$

– probabilités de transition

$$p_{ij}^{(k)} = \frac{\sum_a \sum_{t=2}^{T_a} L_{aj}^{'(k)}(t) p_{ij}^{(k)} F_{ai}^{'(k)}(t-1) / G_{aj}^{'(k)}(t)}{\sum_a \sum_{t=2}^{T_a} L_{ai}^{'(k)}(t-1)}.$$

Pour les paramètres des J modèles linéaires mixtes, nous obtenons :

– paramètres d'effet fixe

$$\beta_j^{(k)} = \Big(\sum_a X_a' L_{aj}^{'(k)} X_a\Big)^{-1} \Big(\sum_a X_a' L_{aj}^{'(k)} \Big(y_{a1}^{T_a} - \tau_j^{(k)} \tilde{\xi}_{aj}^{(k)}\Big)\Big),$$

– variances résiduelles

$$\sigma_j^{2(k+1)} = \frac{\sum_a \left(y_{a1}^{T_a} - X_a\beta_j^{(k)} - \tau_j^{(k)}\tilde{\xi}_{aj}^{(k)}\right)' L_{aj}^{'(k)} \left(y_{a1}^{T_a} - X_a\beta_j^{(k)} - \tau_j^{(k)}\tilde{\xi}_{aj}^{(k)}\right)}{\sum_a \mathrm{tr}\{L_{aj}^{'(k)}\}},$$

– écart-type des effets aléatoires

$$\tau_j^{(k)} = \frac{\sum_a \sum_t L_{aj}^{'(k)}(t)\tilde{\xi}_{aj}^{(k)} \left(y_{at} - X_{at}\beta_j^{(k)}\right)}{\sum_a \sum_t L_{aj}^{'(k)}(t)\tilde{\xi}_{aj}^{(k)2}},$$

où

- X_a est la matrice de dimension $T_a \times Q$ des covariables,
- $L'_{aj} = \mathrm{Diag}\Big\{L'_{aj}(t), t = 1, \ldots, T_a\Big\}$ est une matrice diagonale de dimension $T_a \times T_a$.

La ré-estimation des paramètres des modèles linéaires mixtes est similaire à la ré-estimation des paramètres des modèles linéaires par la méthode des moindres carrés ordinaires ; les paramètres étant seulement pondérés par les pseudo-comptages en probabilités du nombre d'occurrences de chaque état le long de chaque séquence observée.

4.3.6.4 Transposition au cas d'un unique effet aléatoire pour toute la séquence observée

Pour les mêmes raisons que les autres algorithmes MCEM proposés, la transposition de cette méthode d'estimation aux combinaisons markoviennes de modèles linéaires mixtes avec un unique effet aléatoire pour toute la séquence observée (équation (4.1)) est directe. Dans l'étape de maximisation (section 4.3.6.3), la récurrence arrière et la récurrence avant (section 4.3.6.1), les effets aléatoires $\tilde{\xi}_{a1}^J$ sont remplacées par l'effet aléatoire $\tilde{\xi}_a$ tel que :

$$\tilde{\xi}_a = \mathrm{E}\Big(\xi_a | Y_{a1}^{T_a} = y_{a1}^{T_a}\Big) \approx \frac{1}{R}\sum_{r=1}^{R} \xi_a(r)$$

avec $\xi_a(r)$ simulé à partir de l'algorithme de Metropolis-Hastings indépendant avec pour probabilité d'acceptation-rejet :

$$\frac{f(\xi_a^* | Y_{a1}^{T_a} = y_{a1}^{T_a}) q(\xi_a | \xi_a^*)}{f(\xi_a | Y_{a1}^{T_a} = y_{a1}^{T_a}) q(\xi_a^* | \xi_a)} = \frac{\prod_{t=1}^{T_a} N_{at}'^*}{\prod_{t=1}^{T_a} N'_{at}}.$$

4.3.6.5 Remarques

Choix du nombre de réalisations des effets aléatoires à simuler

Il est nécessaire de s'intéresser au nombre de valeurs R à simuler pour obtenir un échantillon de valeurs indépendantes. Le choix de cette valeur est notamment discuté dans le livre de MacKay (2003). Il est important de noter que dans le cadre de l'algorithme de Metropolis-Hastings, tirer R valeurs successivement n'implique pas d'avoir tiré R valeurs

indépendantes. Il peut donc être nécessaire de tenir compte des impératifs d'indépendance entre les valeurs simulées. Cette étape est appelée thinning. La solution envisagée par Gelman et al. (2004) est de ne garder les simulations que tous les k tirages. Le nombre de tirages à partir duquel nous obtenons une approximation correcte de f fait débat. On parle d'étape de burn-in. Gelman et al. (2004) proposent de ne conserver que la seconde moitié des tirages de chaque séquence parallèle (en partant de points initiaux distincts). Leurs inférences finales sont fondées sur l'hypothèse que la distribution des valeurs simulées, pour un assez grand nombre de tirages, est près de la distribution cible. D'après Robert et Casella (2004), il n'existe pas de méthode optimale pour choisir la longueur du burn-in. Dans ce travail, nous avons omis les étapes de thinning et de burn-in. Utiliser une démarche identique à celle de la simulation des séquences d'états dans l'algorithme MCEM avec restauration par simulation pourrait s'avérer adéquate : on tire peu de valeurs pour les premières itérations puis on augmente le nombre au fur et à mesure des itérations.

Convergence de l'algorithme

Sous conditions de convergence des réalisations des effets aléatoires, nous choisissons, pour les mêmes raisons que pour les autres algorithmes MCEM proposés, de contrôler la convergence de l'algorithme MCEM avec une étape E de restauration probabiliste-simulation en observant les fluctuations stationnaires autour de 0 de

$$g^{(k+1)} = \log P(Y = y|\tilde{\xi}^{(k+1)}; \theta^{(k+1)}) - \log P(Y = y|\tilde{\xi}^{(k)}; \theta^{(k)}),$$

où la quantité $\log P(Y = y|\tilde{\xi}; \theta^{(k)})$ est directement obtenue dans la récurrence avant (équation (4.9)).

Initialisation de l'algorithme

L'algorithme MCEM proposé avec une étape E de restauration probabiliste-simulation est sensible aux valeurs initiales. Nous recommandons de choisir comme valeurs initiales les paramètres estimés par l'algorithme EM pour une simple combinaison markovienne ou semi-markovienne de modèles linéaires (en fixant $\xi = 0$).

4.3.7 Extension au SMS-LMM

Étant donné que les covariables et les effets aléatoires modélisant l'hétérogénéité inter-individuelle sont intégrés dans le processus d'observation du SMS-LMM, les observations successives pour un individu sont supposées conditionnellement indépendantes sachant les états non-observables et les effets aléatoires. Les algorithmes MCEM proposés peuvent donc être directement transposés aux combinaisons semi-markoviennes de modèles linéaires mixtes. Sachant les effets aléatoires, des séquences d'états sont soient simulées à partir de l'algorithme avant-arrière de simulation dérivé de celui pour les semi-chaînes de Markov cachées (Guédon, 2007), soient restaurées de manière probabiliste par l'algorithme

avant-arrière dérivé de celui pour les semi-chaînes de Markov cachées (Guédon, 2003). Sachant les séquences d'états simulées, les effets aléatoires sont soient prédits, soient simulés comme expliqué précédemment. Les paramètres de la semi-chaîne de Markov sous-jacente (probabilités initiales, probabilités de transition et lois d'occupation) et les paramètres des modèles linéaires mixtes (paramètres de regression, variances résiduelles et écart-types des effets aléatoires) sont obtenus en maximisant soit l'approximation de l'espérance de la log-vraisemblance des données complètes sachant les données observées, soit l'espérance de la log-vraisemblance des données complètes sachant les données observées et les effets aléatoires. Dans le premier cas, les formules de ré-estimation des probabilités initiales, des probabilités de transition et des lois d'occupation sont similaires aux formules de ré-estimation dans le cas des semi-chaînes de Markov cachées (Guédon (2003) et section 2.2.4.2), les probabilités lissées étant simplement remplacées par des comptages.

4.3.8 Application

	approche 1	approche 2
nombre d'itérations	70	80
π_1	0.95	1
π_2	0.05	0
p_{12}	1	1
p_{13}	0	0

TAB. 4.2 – *Comparaison de la “vitesse” de convergence, des probabilités initiales et des probabilités de transition de la semi-chaîne de Markov sous-jacente entre l'approche 1 : algorithme MCEM avec une étape E de simulation-prédiction et l'approche 2 : algorithme MCEM avec une étape E de restauration probabiliste-simulation.*

Nous avons estimé une combinaison semi-markovienne de modèles linéaires mixtes avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle sur la base des longueurs de pousses annuelles (en cm) des 4 sous-échantillons de pins Laricio (section 1.3.1 et figure 1.7). La semi-chaîne de Markov sous-jacente est supposée de type “gauche-droite” à 3 états avec 2 états successifs transitoires suivis par un état final absorbant (cf remarque 2.2.7). On s'est appuyé sur les travaux de Guédon et al. (2007) pour le choix du nombre d'états et du type de structure sous-jacente. En effet, Guédon et al. (2007) ont montré que la composante ontogénique pouvait être modélisée par un modèle de type “gauche-droite” où les états sont ordonnés et où chaque état ne peut être visité au maximum qu'une fois. Le modèle linéaire mixte associé à chaque état $S_{at} = s_{at}$ est défini pour chaque individu a au temps t par la relation :

$$y_{at}|_{S_{at}=s_{at}} = \beta_{s_{at}1} + X_t\beta_{s_{at}2} + \tau_{s_{at}}\xi_{as_{at}} + \epsilon_{at},$$

où y_{at} est la longueur de pousse annuelle pour l'arbre a au temps t, X_t est une covariable

climatique commune à tous les arbres et variant dans le temps et $\epsilon_{at}|_{S_{at}=s_{at}} \sim \mathcal{N}(0, \sigma^2_{s_{at}})$. Nous supposons qu'il y a un effet aléatoire différent pour chaque état.

Les tableaux 4.2 et 4.3 donnent les estimations obtenues avec l'algorithme MCEM avec une étape E de simulation-prédiction (approche 1) et avec l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (approche 2). L'approche 1 repose sur la simulation de $M_k = k$ séquences d'états par individu à l'itération k tandis que l'approche 2 repose sur 1000 tirages par l'algorithme de Metropolis-Hastings à chaque itération de l'algorithme MCEM.

	état 1		état 2		état 3	
	appr. 1	appr. 2	appr. 1	appr. 2	appr. 1	appr. 2
loi d'occupation	B(2,4,0.37)	B(1,4,0.53)	NB(1,73.29,0.94)	B(1,17,0.22)		
moy., et.	2.73, 0.68	2.58, 0.86	5.56, 2.20	4.57, 1.67		
paramètre LMM						
β_{j1}	7.09	7.72	25.79	28.96	50.25	49.49
β_{j2}	0.0027	0.0021	0.0165	0.0166	0.0309	0.0313
variance						
σ^2_j	4.74	4.47	39.95	41.10	76.86	81.72
τ^2_j	5.79	6.06	49.89	46.47	69.39	77.30
$\gamma^2_j = \sigma^2_j + \tau^2_j$	10.53	10.53	89.84	87.57	146.25	159.02
hétérogénéité						
$\frac{\tau^2_j}{\sigma^2_j+\tau^2_j}$	54.99%	57.55%	55.53%	53.07%	47.45%	48.61%

TAB. 4.3 – *Comparaison des estimations obtenues par l'algorithme MCEM avec une étape E de simulation-prédiction (approche 1) et par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (approche 2) pour chaque état.*

L'algorithme MCEM de l'approche 1 est plus rapide que l'algorithme MCEM de l'approche 2. Nous pouvons remarquer que ces deux méthodes d'estimation nous conduisent à des estimations des paramètres de la semi-chaîne de Markov sous-jacente légèrement différentes pour les probabilités initiales et les probabilités de transition. Les temps de séjour dans l'état 1 sont proches : la durée moyenne est pour l'approche 1 de 2.73 temps successifs avec un écart-type de 0.68 tandis que la durée moyenne pour l'approche 2 est de 2.58 temps successifs avec un écart-type de 0.86 (tableau 4.3). La différence principale se situe au niveau de la loi du temps de séjour dans l'état 2. L'approche 2 nous conduit à un temps de séjour plus court (4.57 temps successifs en moyenne pour l'approche 1 contre 5.56 temps successifs pour l'état 2). Cette différence au niveau des lois de temps de séjour entraîne une répartition de la variabilité différente principalement sur les deux premiers états. La moyenne des parts d'hétérogénéité inter-individuelle entre ces deux états reste cependant très proche : 55.26% pour l'approche 1 et 55.31% pour l'approche 2. Nous no-

tons également une diminution significative de la part d'hétérogénéité inter-individuelle entre les deux derniers états.

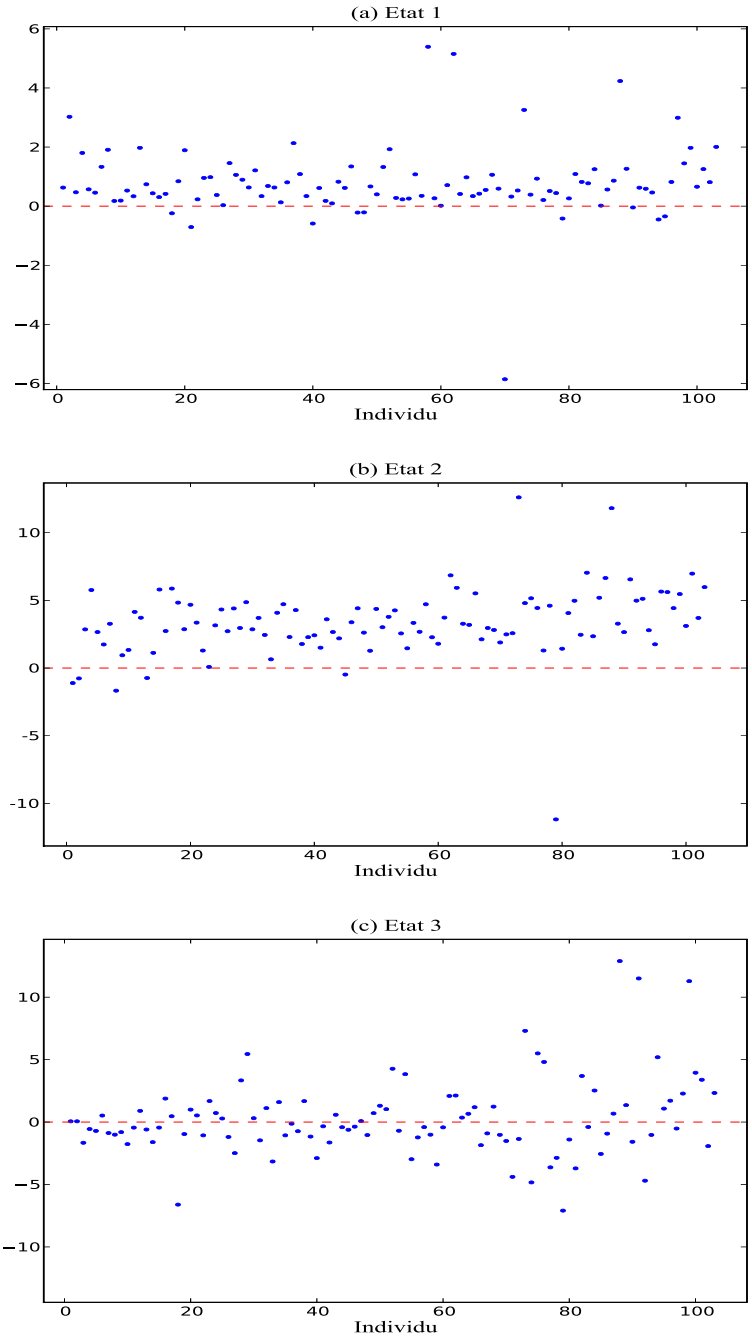

FIG. 4.1 – *Différence pour chaque individu et chaque état entre les effets aléatoires obtenus par l'algorithme MCEM avec une étape E de simulation-prédiction (approche 1) et les effets aléatoires obtenus par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (approche 2).*

L'influence de la covariable climatique sur les longueurs de pousses annuelles est la même pour chacune des approches, au vu des estimations des paramètres de régression β_{j2} dans le tableau 4.3. Dans les deux premiers états, l'approche 1 a tendance à sous-estimer la

constante β_{11} et la constante β_{21} par rapport à l'approche 2. Cette différence est compensée dans ces états par une sur-estimation des prédictions des effets aléatoires obtenus par l'algorithme MCEM avec une étape E de simulation-prédiction ; figure 4.1. La constante associée au troisième état est légèrement plus forte avec l'approche 1 qu'avec l'approche 2. Cette différence est compensée par une sous-estimation des prédictions des effets aléatoires obtenus par l'algorithme MCEM avec une étape E de simulation-prédiction.

4.4 Effet aléatoire "temporel"

Dans la section précédente, nous avons introduit dans le processus d'observation, des modèles linéaires mixtes avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle. Cependant, dans certaines de nos applications, les données environnementales communes à tous les arbres d'un même jeu de données ne sont pas connues. Une solution pour pallier ce problème est d'introduire dans le processus d'observation un effet aléatoire "temporel" modélisant l'environnement commun à tous les individus à chaque date d. Cet effet aléatoire permet de modéliser les corrélations entre les individus à un âge chronologique donné. Picard et al. (2007) ont introduit ce type d'effet aléatoire dans des modèles de détection de ruptures multiples pour l'analyse de profils CGH[1]. Cette approche trouve également son intérêt dans l'analyse de la production de sperme chez les taureaux en introduisant ce type d'effet aléatoire dans les modèles linéaires mixtes (David et al., 2007).
Dans la suite de cette section, nous omettrons les effets aléatoires modélisant l'hétérogénéité inter-individuelle et introduirons des effets aléatoires modélisant l'environnement commun à tous les individus à chaque date dans le processus d'observation des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes.

Comme nous nous intéressons à un effet aléatoire modélisant l'environnement commun à une date donnée d, il est nécessaire dans la suite d'introduire une double indexation afin de distinguer le temps t du processus sous-jacent de la date d des observations. Dans l'exemple des données de croissance des plantes, le temps t correspond à l'âge ontogénique tandis que la date d correspond à l'âge chronologique ou calendaire ; de plus amples explications sont données dans la section 1.1. Les notations adoptées sont les suivantes :

- h_a, date de la première observation pour l'individu a,
- H_a, date de la dernière observation pour l'individu a,
- $T_a = H_a - h_a + 1$, longueur de la séquence observée pour l'individu a,
- $t_a(d) = d - h_a + 1$ ($t_a(d)$ varie de 1 à T_a), fonction liant la date d au temps $t_a(d)$ pour un individu a, et pour tout d compris entre h_a et H_a,
- $T = \sum_{a=1}^{N} T_a = \sum_{a=1}^{N} H_a - h_a + 1$, la longueur cumulée de toutes les séquences,

[1] hybridation génomique comparative

- $D = \max_a H_a - \min_a h_a + 1$, intervalle de définition des dates,
- $Y_{ah_a}^{H_a} = y_{ah_a}^{H_a}$ est la suite des variables observées et de leurs réalisations $Y_{ah_a} = y_{ah_a}, \ldots, Y_{aH_a} = y_{aH_a}$ relatives à l'individu a.

4.4.1 Modèle d'observation

Nous allons définir les modèles linéaires mixtes qui lient le processus d'observation et le processus d'état. Conditionnellement à l'état $S_{a,t_a(d)} = s_{a,t_a(d)}$, l'observation y_{ad} de l'individu a à la date d (d varie de h_a à H_a) est modélisée par le modèle linéaire mixte suivant :

$$y_{ad}|_{S_{a,t_a(d)}=s_{a,t_a(d)}} = X_{ad}\beta_{s_{a,t_a(d)}} + \varsigma_{s_{a,t_a(d)}}\lambda_d + \epsilon_{ad},$$
$$\lambda_d \sim \mathcal{N}(0,1), \qquad \epsilon_{ad}|_{S_{a,t_a(d)}=s_{a,t_a(d)}} \sim \mathcal{N}(0, \sigma^2_{s_{a,t_a(d)}}).$$

Dans cette définition, X_{ad} est le vecteur ligne de taille Q des covariables pour l'individu a à la date d et λ_d est l'effet aléatoire à la date d. Sachant l'état $S_{a,t_a(d)} = s_{a,t_a(d)}$, $\beta_{s_{a,t_a(d)}}$ est le vecteur colonne de taille Q des paramètres d'effets fixes, $\varsigma_{s_{a,t_a(d)}}$ est l'écart-type de l'effet aléatoire et $\sigma^2_{s_{a,t_a(d)}}$ est la variance résiduelle. Par commodité, les effets aléatoires sont supposés suivre la loi gaussienne $\mathcal{N}(0,1)$.

Les dépendances entre les observations y_{ad} et $y_{a'd'}$ sont définies par les covariances suivantes :

$$\mathrm{cov}\big(Y_{ad}, Y_{a'd'}|\ S_{a,t_a(d)} = s_{a,t_a(d)}, S_{a',t_{a'}(d')} = s_{a',t_{a'}(d')}\big) = \begin{cases} \sigma^2_{s_{a,t_a(d)}} + \varsigma^2_{s_{a,t_a(d)}} & \text{si } a = a' \text{ et } d = d', \\ \varsigma_{s_{a,t_a(d)}}\varsigma_{s_{a',t_{a'}(d)}} & \text{si } a \neq a' \text{ et } d = d', \\ 0 & \text{sinon.} \end{cases}$$

Les effets aléatoires sont supposés indépendants entre deux dates ($\mathrm{cov}(\lambda_d,\ \lambda_{d'}) = 0; d \neq d'$). L'introduction d'effets aléatoires modélisant l'environnement commun à tous les individus à une date donnée invalide l'hypothèse selon laquelle les observations de plusieurs individus à une date donnée sont indépendantes. Les observations de plusieurs individus à une date donnée sont maintenant indépendantes sachant les effets aléatoires. Par analogie, les observations successives pour un individu sont maintenant indépendantes sachant les états non-observables et les effets aléatoires. L'introduction de ce type d'effet aléatoire permet de décomposer, dans l'état j, la variabilité totale γ_j^2 d'un individu à une date donnée en deux parts : la variabilité liée à l'environnement commun à tous les individus pour une date donnée ς_j^2 et la variabilité résiduelle σ_j^2.

La partie suivante est présentée dans le cadre de la famille des combinaisons markoviennes de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus à une date donnée. Comme les MS-LMM et les SMS-LMM ne diffèrent que dans le type de processus sous-jacent, la transposition au cas des combinaisons semi-markoviennes de modèles linéaires mixtes est directe.

4.4.2 Vraisemblance du MS-LMM

Les paramètres du MS-LMM peuvent être scindés en deux sous-ensembles : les paramètres $(\pi_j; j = 1, \ldots, J)$ et $(p_{ij}; i, j = 1, \ldots, J)$ de la chaîne de Markov sous-jacente et les paramètres $(\beta_j; j = 1, \ldots, J)$, $(\varsigma_j; j = 1, \ldots, J)$ et $(\sigma_j^2; j = 1, \ldots, J)$ des J modèles linéaires mixtes. Nous notons par $\theta = (\pi, P, \beta, \varsigma, \sigma^2)$ l'ensemble des paramètres à estimer.

Soit $\lambda = \left(\lambda_{\min_a h_a} \ldots \lambda_{\max_a H_a}\right)'$ le vecteur de taille D des effets aléatoires "temporels". La vraisemblance du vecteur de paramètres θ pour les données observées s'écrit :

$$\begin{aligned} L(\theta) &= \int_\lambda \left\{ \sum_s f(s, \lambda, y; \theta) \right\} d\lambda \\ &= \int_\lambda \left\{ \sum_s f(s; \theta) f(\lambda; \theta) f(y|s, \lambda; \theta) \right\} d\lambda, \end{aligned} \tag{4.11}$$

où $\sum_s$ signifie "somme sur toutes les séquences d'états possibles pour tous les individus observés".

La maximisation directe de cette vraisemblance (équation (4.11)) s'avère difficile. Les états de la chaîne de Markov sous-jacente et les effets aléatoires étant non-observables, l'algorithme EM semble être le candidat naturel pour estimer les paramètres de ce modèle. Cependant, la non-indépendance des données observées pour un individu sachant les séquences d'états empêche la décomposition et le calcul des probabilités lissées $L_{aj}(t_a(d)) = P(S_{a,t_a(d)} = j | Y_{ah_a}^{H_a} = y_{ah_a}^{H_a}; \theta)$ dans l'algorithme "avant-arrière" usuellement utilisé dans le cas des chaînes de Markov cachées classiques. De plus, les quantités $\mathrm{E}(\lambda_d^2 | Y = y; \theta)$, $\mathrm{E}(\lambda_d | S_{t(d)} = s_{t(d)}, Y = y; \theta)$ et $\mathrm{E}(\lambda_d^2 | S_{t(d)} = s_{t(d)}, Y = y; \theta)$ ne peuvent pas être calculées analytiquement pour chaque date d.

Pour les mêmes raisons que pour les effets aléatoires modélisant l'hétérogénéité inter-individuelle (section 4.3.3), nous proposons d'utiliser l'algorithme MCEM avec une étape E constituée de deux restaurations conditionnelles et une étape M de maximisation qui peut s'écrire pour chaque itération :

- Étape de restauration conditionnelle pour les séquences d'états : pour chaque individu a, on restaure $S_{a1}^{T_a}$ à partir de $P(S_{a1}^{T_a} = s_{a1}^{T_a} | \lambda, Y_{a1}^{T_a} = y_{a1}^{T_a})$.
- Étape de restauration conditionnelle pour des effets aléatoires : on restaure λ à partir de $P(\lambda | S = s, Y = y)$.
- Étape de maximisation : on maximise l'approximation de la log-vraisemblance des données complètes sachant les données observées.

Nous étudions la transposition des méthodes d'estimation proposées pour estimer les paramètres du MS-LMM avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle ; voir sections 4.3.4, 4.3.5 et 4.3.6.

4.4.3 Algorithme MCEM avec une étape E de simulation-prédiction ou une étape E de simulation-simulation

L'étape E de l'algorithme MCEM proposé avec simulation-prédiction ou simulation-simulation dans le cas de l'hétérogénéité inter-individuelle consiste en une restauration par simulation des séquences d'états sachant les effets aléatoires, par une transposition directe de l'algorithme "avant-arrière" de simulation proposé par Chib (1996), et en une restauration par prédiction ou par simulation des effets aléatoires sachant les séquences d'états.

Dans l'étape de maximisation (section 4.3.4.3), la récurrence avant et la passe arrière de simulation (section 4.3.4.1), les effets aléatoires "individuels" sont simplement remplacés par les effets aléatoires "temporels".

Intéressons-nous maintenant au calcul de l'espérance conditionnelle $\mathrm{E}(\lambda|S = s, Y = y)$ nécessaire pour l'étape de restauration conditionnelle des effets aléatoires. Un effet aléatoire n'étant pas associée à chaque individu mais à chaque date, il est nécessaire de faire la combinaison des états simulés $s_{a,t_a(d)}$ à chaque date d pour chaque individu. Le nombre de combinaisons possibles est de l'ordre de $M^{|A_d|}$ pour chaque date d où M est le nombre de séquences simulées pour chaque individu et $|A_d|$ est le cardinal de l'ensemble des individus observés à la date d. Il est important de noter que le nombre de combinaisons possibles augmente avec le nombre de séquences d'états restaurées par simulation et le nombre d'individus.

Ces problèmes de combinatoires et par suite de coût de calcul nous amène à privilégier, dans le cas d'effets aléatoires "temporels", l'algorithme MCEM avec une étape de restauration probabiliste-simulation :

- on restaure de manière probabiliste (donc non explicite) l'ensemble des séquences d'états possibles sachant les données observées et les effets aléatoires à partir de l'algorithme "avant-arrière" usuellement utilisé pour les chaînes de Markov cachées (Devijver, 1985),
- on restaure par simulation les effets aléatoires sachant les données observées à partir d'un algorithme de Metropolis-Hastings (Robert et Casella, 2004).

4.4.4 Algorithme MCEM avec une étape E de restauration probabiliste-simulation

Considérons la log-vraisemblance des données complètes où à la fois les observations y, les effets aléatoires λ et les états s de la chaîne de Markov sous-jacente sont observés :

$$\log f(s,\lambda,y;\theta) \;=\; \sum_{a=1}^{N}\left\{\log\Big(\pi_{s_{a1}}+\prod_{t=2}^{T_a}p_{s_{a,t-1},s_{a,t}}\Big)+\log\Big(\prod_{d=\min_a h_a}^{\max_a H_a}\phi(\lambda_d;0,1)\Big)\right.$$

$$
\begin{aligned}
&+\log\Big(\prod_{d=h_a}^{H_a}\phi(y_{ad};X_{ad}\beta_{s_{a,t_a(d)}}+\varsigma_{s_{a,t_a(d)}}\lambda_d,\ \sigma^2_{s_{a,t_a(d)}})\Big)\Big\}\\
&= \sum_{a=1}^{N}\sum_{j=1}^{J} I\Big(s_{a1}=j\Big)\log\pi_j+\sum_{a=1}^{N}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} I\Big(s_{a,t}=j,\ s_{a,t-1}=i\Big)\log p_{ij}\\
&+ \sum_{d=\min_a h_a}^{\max_a H_a}\log\phi(\lambda_d;0,1)\\
&+ \sum_{a=1}^{N}\sum_{d=h_a}^{H_a}\sum_{j=1}^{J} I\Big(s_{a,t_a(d)}=j\Big)\log\phi(y_{ad};X_{ad}\beta_j+\varsigma_j\lambda_d,\ \sigma_j^2).
\end{aligned}
$$

Dans l'étape E de l'algorithme MCEM proposé, on s'intéresse à l'espérance de la log-vraisemblance des données complètes conditionnellement aux données observées et aux effets aléatoires modélisant l'environnement commun :

$$
\begin{aligned}
Q(\theta|\theta^{(k)}) &= \mathrm{E}\Big(\log f(s,\lambda,y;\theta)|\tilde{\lambda},Y=y;\theta^{(k)}\Big)\\
Q(\theta|\theta^{(k)}) &= \sum_{a=1}^{N}\sum_{j=1}^{J} L'_{aj}(1)\log\pi_j\\
&+ \sum_{a=1}^{N}\sum_{t=2}^{T_a}\sum_{i,j=1}^{J} P\big(S_{a,t}=j,\ S_{a,t-1}=i|\tilde{\lambda}_{h_a}^{H_a},Y_{ah_a}^{H_a}=y_{ah_a}^{H_a};\theta^{(k)}\big)\log p_{ij}\\
&- \frac{D}{2}\log 2\pi-\sum_{d=\min_a h_a}^{\max_a H_a}\frac{\tilde{\lambda}_d^2}{2}-\sum_{a=1}^{N}\frac{T_a}{2}\log 2\pi\\
&- \sum_{a=1}^{N}\sum_{d=h_a}^{H_a}\sum_{j=1}^{J} L'_{aj}(t_a(d))\Big(\frac{\log\sigma_j^2}{2}+\frac{(y_{ad}-x_{ad}\beta_j-\varsigma_j\tilde{\lambda}_d)^2}{2\sigma_j^2}\Big) \qquad (4.12)
\end{aligned}
$$

avec les probabilités lissées $L'_{aj}(t_a(d)) = P\big(S_{a,t_a(d)}=j|Y_{ah_a}^{H_a}=y_{ah_a}^{H_a},\tilde{\lambda}_{h_a}^{H_a};\theta^{(k)}\big)$ et $\tilde{\lambda}=\mathrm{E}(\lambda|Y=y)$.

4.4.4.1 Algorithme "avant-arrière" sachant les effets aléatoires

L'algorithme "avant-arrière" proposé par Devijver (1985) pour les chaînes de Markov cachées classiques repose pour les MS-LMM sur la décomposition suivante des probabilités lissées pour chaque individu a :

$$
\begin{aligned}
L'_{aj}(t_a(d)) &= P(S_{a,t_a(d)}=j|Y_{ah_a}^{H_a}=y_{ah_a}^{H_a},\tilde{\lambda}_{h_a}^{H_a})\\
&= \frac{P(Y_{a,d+1}^{H_a}=y_{a,d+1}^{H_a}|S_{a,t_a(d)}=j,\tilde{\lambda}_{h_a}^{H_a})}{P(Y_{a,d+1}^{H_a}=y_{a,d+1}^{H_a}|Y_{ah_a}^{d}=y_{ah_a}^{d},\tilde{\lambda}_{h_a}^{H_a})}P(S_{a,t_a(d)}=j|Y_{ah_a}^{d}=y_{ah_a}^{d},\tilde{\lambda}_{h_a}^{H_a})\\
&= B'_{aj}(t_a(d))F'_{aj}(t_a(d)).
\end{aligned}
$$

Les quantités $F'_{aj}(t_a(d))$ peuvent être calculées à l'aide d'une récurrence avant (c'est-à-dire de 1 à T_a) tandis que les quantités $B'_{aj}(t_a(d))$ ou $L'_{aj}(t_a(d))$ peuvent être calculées par une récurrence arrière (c'est-à-dire de T_a à 1) pour chaque individu a.

Récurrence avant

Dans le cas du MS-LMM, la récurrence avant est initialisée pour $t_a(d) = 1$ par :

$$\begin{aligned} F'_{aj}(1) &= P(S_{a1} = j | Y_{ah_a} = y_{ah_a}, \tilde{\lambda}_{h_a}^{H_a}) \\ &= \frac{\phi(y_{ah_a}; X_{ah_a}\beta_j + \varsigma_j \tilde{\lambda}_{h_a}, \sigma_j^2)}{N'_{ah_a}} \ \pi_j, \end{aligned}$$

où $N'_{ah_a} = P(Y_{ah_a} = y_{ah_a} | \tilde{\lambda}_{h_a}^{H_a})$ est le facteur de normalisation tel que

$$N'_{ah_a} = \sum_{j=1}^{J} p(S_{a1} = j, Y_{ah_a} = y_{ah_a} | \tilde{\lambda}_{h_a}^{H_a}) = \sum_{j=1}^{J} \phi(y_{ah_a}; X_{ah_a}\beta_j + \varsigma_j \tilde{\lambda}_{h_a}, \sigma_j^2)\pi_j.$$

Pour $t_a(d) = 2, \ldots, T_a$, la récurrence avant est donnée par :

$$\begin{aligned} F'_{aj}(t_a(d)) &= P(S_{a,t_a(d)} = j | Y_{ah_a}^d = y_{ah_a}^d, \tilde{\lambda}_{h_a}^{H_a}) \\ &= \frac{\phi(y_{ad}; X_{ad}\beta_j + \varsigma_j \tilde{\lambda}_d, \sigma_j^2)}{N'_{ad}} \sum_{i=1}^{J} p_{ij} F'_{ai}(t_a(d) - 1). \end{aligned}$$

Le facteur de normalisation $N'_{ad} = P(Y_{ad} = y_{ad} | Y_{ah_a}^{d-1} = y_{ah_a}^{d-1}, \tilde{\lambda}_{h_a}^{H_a})$ est obtenu directement durant la récurrence avant :

$$\begin{aligned} N'_{ad} &= \sum_{j=1}^{J} P(S_{a,t_a(d)} = j, Y_{ad} = y_{ad} | Y_{ah_a}^{d-1} = y_{ah_a}^{d-1}, \tilde{\lambda}_{h_a}^{H_a}) \\ &= \sum_{j=1}^{J} \phi(y_{ad}; X_{ad}\beta_j + \varsigma_j \tilde{\lambda}_d, \sigma_j^2) \sum_{i=1}^{J} p_{ij} F'_{ai}(t_a(d) - 1). \end{aligned}$$

La récurrence avant peut être utilisée pour calculer la log-vraisemblance des données observées sachant les effets aléatoires :

$$\begin{aligned} \log P(Y = y | \tilde{\lambda}; \theta) &= \sum_{a=1}^{N} \log P(Y_{ah_a}^{H_a} = y_{ah_a}^{H_a} | \tilde{\lambda}_{h_a}^{H_a}; \theta) \\ &= \sum_{a=1}^{N} \Bigg(\log P(Y_{ah_a} = y_{ah_a} | \tilde{\lambda}_{h_a}^{H_a}; \theta) \\ &\qquad \times \sum_{d=h_a+1}^{H_a} \log P(Y_{ad} = y_{ad} | Y_{ah_a}^{d-1} = y_{ah_a}^{d-1}, \tilde{\lambda}_{h_a}^{H_a}; \theta) \Bigg) \\ &= \sum_{a=1}^{N} \sum_{d=h_a}^{H_a} \log N'_{ad}. \end{aligned} \tag{4.13}$$

Récurrence arrière

La récurrence arrière est initialisée pour $t_a(d) = T_a$ par :

$$L'_{aj}(T_a) = P(S_{aT_a} = j | Y_{ah_a}^{H_a} = y_{ah_a}^{H_a}, \tilde{\lambda}_{h_a}^{H_a}) = F'_{aj}(T_a).$$

Pour $t_a(d) = T_a - 1, \ldots, 1$, la récurrence arrière est donnée par :

$$\begin{aligned} L'_{aj}\big(t_a(d)\big) &= P(S_{a,t_a(d)} = j | Y_{ah_a}^{H_a} = y_{ah_a}^{H_a}, \tilde{\lambda}_{h_a}^{H_a}) \\ &= \left\{ \sum_k \frac{L'_{ak}\big(t_a(d)+1\big)}{G'_{ak}\big(t_a(d)+1\big)} p_{jk} \right\} F'_{aj}\big(t_a(d)\big) \end{aligned}$$

où la quantité $G'_{ak}\big(t_a(d)+1\big) = P(S_{a,t_a(d)+1} = k | Y_{ah_a}^{d} = y_{ah_a}^{d}, \tilde{\lambda}_{h_a}^{H_a}) = \sum_j p_{jk} F'_{aj}\big(t_a(d)\big)$ peut être directement extraite et stockée en mémoire durant la récurrence avant.

4.4.4.2 Simulation des effets aléatoires sachant les séquences d'états

Le vecteur des prédictions des effets aléatoires modélisant l'environnement commun à tous les individus est :

$$\tilde{\lambda} = \mathrm{E}\Big(\lambda | Y = y\Big).$$

Cette espérance peut être approximée par l'algorithme de Metropolis-Hastings (Robert et Casella, 2004).

Algorithme de Metropolis-Hastings indépendant

Pour les combinaisons markoviennes de modèles linéaires mixtes avec des effets aléatoires "temporels", la densité conditionnelle $q()$ est la densité de la loi multivariée gaussienne $\mathcal{N}_D(0, Id_D)$.

En effet, nous cherchons à simuler les effets aléatoires selon la loi de $\lambda | Y = y$. Soit λ^* la proposition et soit λ l'ancien vecteur des réalisations. La probabilité d'acceptation-rejet (équation (4.10)) s'écrit alors :

$$\begin{aligned} \frac{f(\lambda^*|Y=y)q(\lambda|\lambda^*)}{f(\lambda|Y=y)q(\lambda^*|\lambda)} &= \frac{f(\lambda^*|Y=y)f(\lambda)}{f(\lambda|Y=y)f(\lambda^*)} \\ &= \frac{f(y|\lambda^*)f(\lambda^*)f(\lambda)}{f(y|\lambda)f(\lambda)f(\lambda^*)} \\ &= \frac{f(y|\lambda^*)}{f(y|\lambda)} \\ &= \frac{\prod_{a=1}^{N}\prod_{d=h_a}^{H_a} N'^{*}_{ad}}{\prod_{a=1}^{N}\prod_{d=h_a}^{H_a} N'_{ad}}. \end{aligned}$$

La probabilité d'acceptation-rejet dépend uniquement des facteurs de normalisation N'_{ad} et N'^{*}_{ad} calculés dans la passe "avant" de l'algorithme "avant-arrière".

L'algorithme de Metropolis-Hastings repose sur un parcours aléatoire de l'espace des réalisations des effets aléatoires. Aussi, plus le nombre d'éléments à simuler conjointement est élevée, plus la probabilité d'acceptation-rejet sera faible. Les valeurs proposées à chaque tirage étant par conséquent souvent rejetées, ceci entraîne des gros problèmes de mélangeance. Or, contrairement au cas d'effets aléatoires “individuels” où au maximum J réalisations sont simulées à chaque tirage, il est nécessaire de simuler D effets aléatoires “temporels” à chaque tirage. C'est pourquoi, bien que cet algorithme soit efficace dans le cadre de combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes avec des effets aléatoires “individuels” (sections 4.3.6.2 et 4.3.6.4), il ne semble pas adapté aux modèles avec des effets aléatoires “temporels”. Une alternative à l'algorithme de Metropolis-Hastings indépendant est l'algorithme de Metropolis-Hastings à marche aléatoire.

Algorithme de Metropolis-Hastings à marche aléatoire

La densité conditionnelle utilisée $q(z|x^{(k)})$ est telle que $z_k = x^{(k)} + \iota^{(k)}$ où $\iota^{(k)}$ est une perturbation aléatoire symétrique indépendante de $x^{(k)}$ de type distribution gaussienne $\mathcal{N}(0, \omega^2)$ où ω^2 est fixée et connue. Choisir une telle distribution permet de profiter de sa propriété de symétrie ($q(z|x^{(k)}) = q(x^{(k)}|z)$, voir Robert et Casella, p. 287) et par conséquent, cette densité conditionnelle n'interviendra pas dans la probabilité d'acceptation-rejet. Comme précédemment, nous cherchons à simuler les effets aléatoires selon la loi de $\lambda|Y = y$. La probabilité d'acceptation-rejet (équation (4.10)) s'écrit alors :

$$\begin{aligned}
\frac{f(\lambda^*|Y=y)q(\lambda|\lambda^*)}{f(\lambda|Y=y)q(\lambda^*|\lambda)} &= \frac{f(\lambda^*|Y=y)}{f(\lambda|Y=y)} \\
&= \frac{f(y|\lambda^*)f(\lambda^*)}{f(y|\lambda)f(\lambda)} \\
&= \frac{f(\lambda^*)\prod_{a=1}^{N}\prod_{d=h_a}^{H_a} N'^{*}_{ad}}{f(\lambda)\prod_{a=1}^{N}\prod_{d=h_a}^{H_a} N'_{ad}}
\end{aligned}$$

La probabilité d'acceptation-rejet se calcule facilement ; elle dépend uniquement des facteurs de normalisation calculés dans la récurrence avant de l'algorithme “avant-arrière” et de la loi connue des effets aléatoires “temporels” (section 4.4.4.1). De plus, contrairement à l'algorithme de Metropolis-Hastings indépendant, l'algorithme de Metropolis-Hastings à marche aléatoire permet un parcours non aléatoire de l'espace des réalisations des effets aléatoires. Il correspond à chaque tirage à une exploration locale du voisinage des valeurs locales et permet d'éviter les problèmes de mélangeance. Nous allons par conséquent utiliser l'algorithme de Metropolis-Hastings à marche aléatoire pour simuler les effets aléatoires à chaque itération.

Le vecteur des effets aléatoires “temporels” est donc :

$$\tilde{\lambda} = \mathrm{E}\Big(\lambda|Y=y\Big) \approx \frac{1}{R}\sum_{r=1}^{R}\lambda(r)$$

où R est le nombre de simulations des effets aléatoires par l'algorithme de Metropolis-Hastings à marche aléatoire.

Remarque :
Il est important de noter que les séquences d'états restaurées conditionnent implicitement les simulations des effets aléatoires au travers des facteurs de normalisation. C'est pourquoi, nous pouvons parler de simulations des effets aléatoires sachant les séquences d'états.

4.4.4.3 Étape de maximisation

À l'itération k, les ré-estimations des paramètres du MS-LMM sont obtenues en maximisant les différents termes de l'équation (4.12), chaque terme dépendant d'un sous-ensemble de θ.

Pour les paramètres de la chaîne de Markov sous-jacente, nous obtenons :

– probabilités initiales

$$\pi_j^{(k+1)} = \frac{\sum_a L_{aj}^{'(k)}(1)}{N},$$

– probabilités de transition

$$p_{ij}^{(k)} = \frac{\sum_a \sum_{t=2}^{T_a} L_{aj}^{'(k)}(t) p_{ij}^{(k)} F_{ai}^{'(k)}(t-1)/G_{aj}^{'(k)}(t)}{\sum_a \sum_{t=2}^{T_a} L_{ai}^{'(k)}(t-1)}.$$

Pour les paramètres des J modèles linéaires mixtes, nous obtenons :

– paramètres d'effet fixe

$$\beta_j^{(k)} = \left(\sum_a X_a' L_{aj}^{'(k)} X_a \right)^{-1} \left(\sum_a X_a' L_{aj}^{'(k)} \left(y_{ah_a}^{H_a} - \varsigma_j^{(k)} \tilde{\lambda}_{h_a}^{H_a(k)} \right) \right),$$

– variances résiduelles

$$\sigma_j^{2(k+1)} = \frac{\sum_a \left(y_{ah_a}^{H_a} - X_a \beta_j^{(k)} - \varsigma_j^{(k)} \tilde{\lambda}_{h_a}^{H_a(k)} \right)' L_{aj}^{'(k)} \left(y_{ah_a}^{H_a} - X_a \beta_j^{(k)} - \varsigma_j^{(k)} \tilde{\lambda}_{h_a}^{H_a(k)} \right)}{\sum_a \mathrm{tr}\{L_{aj}^{(k)}\}},$$

– écart-type des effets aléatoires

$$\varsigma_j^{(k)} = \left(\sum_a \tilde{\lambda}_{h_a}^{'H_a(k)} L_{aj}^{'(k)} \tilde{\lambda}_{h_a}^{H_a(k)} \right)^{-1} \left(\sum_a \tilde{\lambda}_{h_a}^{'H_a(k)} L_{aj}^{'(k)} \left(y_{ah_a}^{H_a} - X_a \beta_j^{(k)} \right) \right),$$

où

- X_a est la matrice de dimension $T_a \times Q$ des covariables,
- $L'_{aj} = \mathrm{Diag}\left\{ L'_{aj}(t), t = 1, \ldots, T_a \right\}$ est une matrice diagonale de dimension $T_a \times T_a$,
- $\tilde{\lambda}_{h_a}^{'H_a}$ est le vecteur ligne de taille T_a des effets aléatoires.

4.4.4.4 Remarques

La remarque pour le choix du nombre d'effets aléatoires à simuler est la même que dans le cadre de MS-LMM avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle : il est conseillé d'augmenter le nombre de simulations au fil des itérations.

Choix de ω^2

Dans son livre, MacKay (2003) discute du choix de ω^2 et de son impact sur l'échantillon simulé. Plus la variance ω^2 est élevée, plus l'espace visité à chaque itération sera large. Plus la variance ω^2 est faible, plus l'espace visité à chaque itération est restreint : la perturbation proposée étant faible. Il faut par conséquent trouver un compromis afin que la perturbation envisagée ne soit ni trop forte, ni trop faible. Robert et Casella (2004) consacre un chapitre de leur livre à l'algorithme de Metropolis-Hastings, au choix de la densité conditionnelle et au choix de la puissance de la perturbation.

Convergence de l'algorithme

Sous conditions de convergence des effets aléatoires simulés, nous choisissons, pour les mêmes raisons que dans le cas du MS-LMM avec des effets aléatoires “individuels”, de contrôler la convergence de l'algorithme MCEM proposé en observant les fluctuations stationnaires autour de 0 de

$$g^{(k+1)} = \log P(Y = y|\tilde{\lambda}^{(k+1)}; \theta^{(k+1)}) - \log P(Y = y|\tilde{\lambda}^{(k)}; \theta^{(k)}), \tag{4.14}$$

où la quantité $\log P(Y = y|\tilde{\lambda}; \theta^{(k)})$ est directement obtenue dans la récurrence “avant” (équation (4.13)).

Initialisation de l'algorithme

L'algorithme MCEM proposé avec une étape E de restauration probabiliste-simulation est sensible aux valeurs initiales. Nous recommandons de choisir comme valeurs initiales les paramètres estimés par l'algorithme EM pour une simple combinaison markovienne de modèles linéaires (en fixant $\lambda = 0$).

4.4.5 Extension au SMS-LMM

Pour les mêmes raisons que pour les SMS-LMM avec des effets aléatoires “individuels” (section 4.3.7), l'algorithme MCEM proposé peut donc être directement transposé aux combinaisons semi-markoviennes de modèles linéaires mixtes. Sachant les effets aléatoires, l'ensemble des séquences d'états est restauré de manière probabiliste à partir de l'algorithme “avant-arrière” dérivé de celui pour les semi-chaînes de Markov cachées (Guédon, 2003). Sachant les séquences d'états restaurées, les effets aléatoires sont simulés à partir de l'algorithme de Metropolis-Hastings à marche aléatoire. Les paramètres de la semi-chaîne de Markov sous-jacente (probabilités initiales, probabilités de transition et lois d'occupation) et les paramètres des modèles linéaires mixtes (paramètres de regression, variances

résiduelles et écart-types des effets aléatoires) sont obtenus en maximisant l'approximation de la log-vraisemblance des données complètes sachant les données observées et les effets aléatoires.

4.4.6 Simulations

Afin de tester l'efficacité de l'algorithme MCEM proposé selon le degré de séparabilité des états de la chaîne de Markov sous-jacente, nous nous plaçons dans le même cadre que les simulations réalisées pour les MS-LMM avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle :

- probabilités initiales : $\pi = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix}$,
- probabilités de transition : $P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$,
- variances résiduelles : $\sigma^2 = \begin{pmatrix} 2 & 1 \end{pmatrix}$,
- variances des effets aléatoires : $\varsigma^2 = \begin{pmatrix} 1 & 2 \end{pmatrix}$.

Nous avons considéré un unique effet fixe par état et nous avons simulé 20 individus sur 30 temps selon deux cas :

- Modalité (A) défini par $\beta_1 = 2$ et $\beta_2 = 9$, les états sont bien séparés.
- Modalité (B) défini par $\beta_1 = 4$ et $\beta_2 = 7$, les états sont moins bien séparés.

	paramètre	vraie valeur	cas (A)	vraie valeur	cas (B)
	π_1	0.9	0.906 (0.074)	0.9	0.805 (0.135)
	p_{11}	0.8	0.799 (0.023)	0.8	0.792 (0.046)
	p_{22}	0.7	0.700 (0.033)	0.7	0.711 (0.048)
	β_1	2	1.930 (0.215)	4	4.161 (0.699)
état 1	ς_1^2	1	0.980 (0.294)	1	1.108 (0.450)
	σ_1^2	2	1.986 (0.161)	2	1.984 (0.315)
	β_2	9	8.849 (0.275)	7	6.734 (0.757)
état 2	ς_2^2	2	1.920 (0.549)	2	1.994 (0.667)
	σ_2^2	1	1.105 (0.117)	1	1.193 (0.316)

TAB. 4.4 – *Environnement commun : résultats d'estimation des paramètres d'une combinaison markovienne de modèles linéaires mixtes à 2 états par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation sur 100 jeux de données simulés pour le cas d'états bien séparés (A) et pour le cas d'états moins bien séparés (B) : moyennes (écart-types).*

Le tableau 4.4 donne les moyennes et les écart-types des estimations obtenues sur 100 jeux de données simulées selon le degré de séparabilité des états. Chaque simulation est basée sur 1000 itérations de l'algorithme de Metropolis-Hastings pour obtenir les effets aléatoires à chaque itération de l'algorithme MCEM proposé. Nous avons choisi comme variance $\omega^2 = 1$ pour la densité conditionnelle $q()$ utilisée dans l'algorithme de Metropolis-Hastings à marche aléatoire.

Nous pouvons noter un bon comportement de l'algorithme MCEM avec une étape E de restauration probabiliste-simulation pour l'ensemble des paramètres de la combinaison markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus. Cependant, le degré de séparabilité des états a une forte influence sur les estimations des paramètres. Cette influence se retrouve principalement dans l'augmentation des écart-types des estimations et dans les probabilités initiales de la chaîne de Markov sous-jacente. Plus les états sont distincts, meilleures sont les estimations.

4.4.7 Application

Nous avons estimé une combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun sur la base des longueurs de pousses annuelles (en cm) des 30 pins Laricios âgés de 18 ans (section 1.3.1 et figure 1.7). La semi-chaîne de Markov sous-jacente est supposée de type “gauche-droite” à 3 états avec 2 états successifs transitoires suivis par un état final absorbant (cf remarque 2.2.7). Le modèle linéaire mixte associé à chaque état $S_{a,t_a(d)} = s_{a,t_a(d)}$ est défini pour chaque individu a à l'âge chronologique d par la relation :

$$y_{ad}|_{S_{a,t_a(d)}=s_{a,t_a(d)}} = \beta_{s_{a,t_a(d)}} + \varsigma_{s_{a,t_a(d)}}\lambda_d + \epsilon_{ad},$$

où y_{ad} est la longueur de pousse pour l'individu a à l'âge chronologique (ou date) d et $\epsilon_{ad}|_{S_{a,t_a(d)}=s_{a,t_a(d)}} \sim \mathcal{N}(0, \sigma^2_{s_{a,t_a(d)}})$.

Nous avons comparé les paramètres estimés de la combinaison semi-markovienne de modèles linéaires mixtes avec les paramètres estimés d'une semi-chaîne de Markov cachée gaussienne (GHSMC, Gaussian Hidden Semi-Markov Chain) ; voir le tableau 4.5. Nous notons une modification de la loi de temps de séjour dans l'état 2 : bien que la durée moyenne reste proche, son écart-type a plus que doublé. Nous notons également une différence significative entre les constantes β_j et une augmentation significative de la variabilité sur le premier et le troisième état.

La différence entre les deux modèles est liée à l'introduction d'un effet aléatoire modélisant l'environnement commun à tous les individus à une date donnée. Afin de mieux comprendre l'origine de cette différence, nous observons les effets aléatoires obtenus par

	état 1		état 2		état 3	
	SMS-LMM	GHSMC	SMS-LMM	GHSMC	SMS-LMM	GHSMC
loi occup.	NB(1,1.97,0.45)	NB(1,2.95,0.56)	NB(2,0.99,0.15)	NB(1,169.73,0.96)		
moy., et.	3.44, 2.33	3.29, 2.01	7.62, 6.12	7.88, 2.68		
param. β_j	9.95	7.24	24.92	26.13	45.74	51.70
var. tot. γ_j^2	11.62	8.86	91.93	88.99	170.13	141.29

TAB. 4.5 – *Comparaison des paramètres estimés pour la semi-chaîne de Markov cachée gaussienne (GHSMC) et pour la combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun (SMS-LMM).*

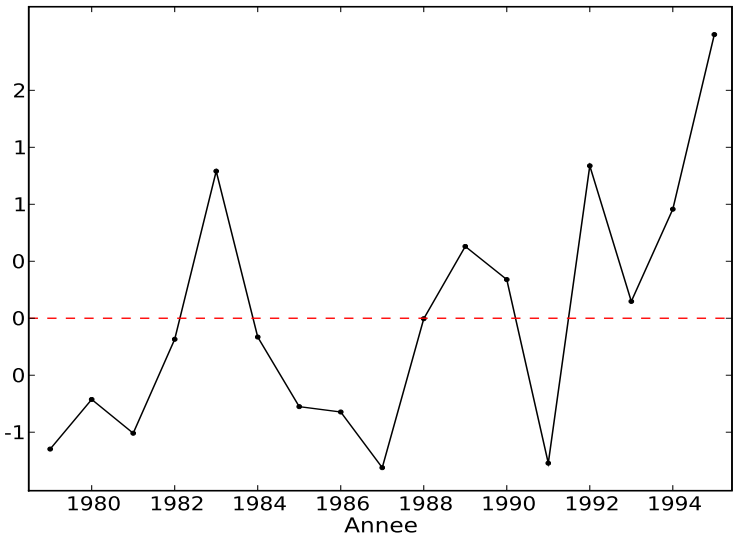

FIG. 4.2 – *Simulations des effets aléatoires modélisant l'environnement commun en fonction de l'âge chronologique (ou année) pour le SMS-LMM estimé par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation.*

l'algorithme de Metropolis-Hastings à marche aléatoire dans l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (cf figure 4.2).

Les effets aléatoires pour la période allant de 1990 à 1995 correspondent bien à des fluctuations synchrones entre arbres liées à l'environnement commun. Cependant, les fortes augmentations entre 1981 et 1983 et entre 1987 et 1988 ne correspondent pas à des fluctuations synchrones entre arbres liées à l'environnement mais à des années de changement d'état ; la succession des états étant ici caractérisée par une augmentation brutale des longueurs de pousses annuelles. Ce phénomène souligne un problème d'identifiabilité : si le degré de synchronisme entre les individus pour la date de changement d'état est élevé, l'effet aléatoire modélisant l'environnement commun "éponge" ce synchronisme, ce qui entraîne une certaine modification des estimations.

4.5 Conclusion et discussion

Dans ce chapitre, nous avons étudié la famille des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes pour deux types d'effets aléatoires : des effets aléatoires "individuels" et des effets aléatoires "temporels". Quel que soit le type des effets aléatoires des modèles linéaires mixtes associés à chaque état du processus sous-jacent, nous sommes en présence de deux structures non-observables :

- les états du processus markovien sous-jacent,
- les effets aléatoires.

Nous avons choisi de nous appuyer sur des approches de type MCEM pour estimer les paramètres des MS-LMM et des SMS-LMM dont l'étape E repose sur deux restaurations conditionnelles : une restauration des séquences d'états sachant les effets aléatoires (et les données observées) et une restauration des effets aléatoires sachant les séquences d'états (et les données observées). Les méthodes étudiées dans ce chapitre sont résumées dans le tableau 4.6.

démarche	1 simulation-prédiction	2 probabiliste-simulation	3 simulation-simulation
séquences d'états	restauration par simulation ***avant-arrière de simulation***	restauration probabiliste ***avant-arrière***	restauration par simulation ***avant-arrière de simulation***
effets aléatoires	restauration par prédiction ***espérance conditionnelle***	restauration par simulation ***Metropolis-Hastings***	restauration par simulation ***distribution conditionnelle***

Tab. 4.6 – *Variantes des restaurations conditionnelles dans l'étape E de l'algorithme MCEM pour les combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes.*

La démarche 1 (simulation-prédiction) est adéquate dans le cadre d'effets aléatoires modélisant l'hétérogénéité inter-individuelle. Les avantages de cette approche sont la rapidité de convergence et le calcul explicite et simple des effets aléatoires. Cependant, cette approche ne peut pas se transposer à d'autres types d'effets aléatoires nécessitant le calcul des réalisations sur un grand nombre de combinaisons possibles ou à des processus d'observations non gaussiens où l'écriture de l'espérance conditionnelle des effets aléatoires n'est pas aussi directe.

La démarche 2 (restauration probabiliste-simulation) a l'avantage de ne pas être contrainte par le type d'effets aléatoires : elle se transpose aussi bien au cas d'effets

aléatoires "individuels" qu'au cas d'effets aléatoires "temporels". Cette démarche n'est également pas contrainte par le type du processus d'observation : gaussien ou non. La restauration probabiliste de l'ensemble des séquences d'états à chaque itération permet de parcourir à chaque itération l'ensemble de l'espace des séquences d'états. Cependant, deux principaux inconvénients sont à souligner. L'approche avec une étape de Metropolis-Hastings repose pour chaque individu sur le calcul du ratio entre la log-vraisemblance des données observées sachant les effets aléatoires et la log-vraisemblance des données observées sachant les valeurs proposées des effets aléatoires à chaque tirage. Le calcul de la seconde quantité nécessite de devoir calculer les constantes de normalisation lors de la récurrence "avant" sachant les valeurs proposées des effets aléatoires. Par conséquent, cette approche est beaucoup plus lente que les autres approches. Le second inconvénient vient du choix du type d'algorithme de Metropolis-Hastings et dans le cas de l'algorithme de Metropolis-Hastings à marche aléatoire du choix de ω^2.

La démarche 3 (simulation-simulation) semble être un compromis entre les deux autres démarches. Cette approche n'a pas été implémentée et donc testée. Cependant, elle ne peut pas se transposer aux effets aléatoires "temporels" car elle nécessite de connaître explicitement des séquences d'états afin de pouvoir calculer simplement la distribution conditionnelle des effets aléatoires sachant les données observées et les séquences d'états.

Chapitre 5

Applications aux données de croissance d'arbres forestiers

Notre travail n'était pas seulement de répondre à une problématique statistique mais également de répondre à une problématique biologique bien définie. L'objectif biologique était de séparer et de caractériser les trois composantes principales de la croissance d'une plante (section 1.2) :

- la composante ontogénique, supposée être structurée en phases de croissance asynchrones entre individus,
- la composante environnementale principalement d'origine climatique, supposée prendre la forme de fluctuations synchrones entre individus et dont l'amplitude moyenne peut être modulée d'une phase de croissance à une autre,
- la composante individuelle dont la part est supposée varier d'une phase de croissance à une autre.

Ce chapitre est ainsi consacré à la modélisation de données de croissance (longueur de pousses annuelles, nombre de branches) d'arbres forestiers. Notre démarche s'est appuyée sur les travaux de Guédon et al. (2007) pour le choix du nombre d'états et du type de structure sous-jacente pour les pins Laricio, les chênes sessiles et les pins sylvestres. En effet, Guédon et al. (2007) ont montré que la composante ontogénique pouvait être modélisée par un modèle de type "gauche-droite" où les états sont ordonnés et où chaque état ne peut être visité au maximum qu'une fois. Dans le cadre de la modélisation de la croissance des plantes telle qu'envisagée ici, chaque état représente une phase de croissance. Comme la dernière année d'observation est arbitraire au regard du développement d'un arbre, la longueur de la dernière phase de croissance est supposée être systématiquement censurée à droite et ne peut donc pas être modélisée par une loi de temps de séjour. Le dernière phase de croissance sera donc modélisée par un état final absorbant.

La première partie de ce chapitre est dédiée à l'application des combinaisons semi-markoviennes de modèles linéaires mixtes avec des effets aléatoires modélisant l'hétérogé-

néité inter-individuelle aux jeux de données de pins Laricio, de chênes sessiles, de pins sylvestres et de noyers communs. Les combinaisons semi-markoviennes de modèles linéaires mixtes avec des effets aléatoires modélisant l'environnement commun sont appliquées dans une seconde partie aux jeux de données de pins Laricio et de chênes sessiles.

5.1 Analyse conjointe de la composante ontogénique, de la composante environnementale et de la composante individuelle

Cette partie est consacrée à la distinction et à la caractérisation des trois principales composantes de la croissance d'une plante en modélisant la composante ontogénique par une semi-chaîne de Markov sous-jacente, la composante environnementale par un effet fixe dans le processus d'observation et la composante individuelle par un effet aléatoire dans le processus d'observation.

Nous suivrons un schéma identique pour la présentation des différents résultats pour chaque espèce forestière. On s'intéressera dans un premier temps au modèle choisi pour l'analyse et aux caractéristiques de l'algorithme d'estimation proposé. Une deuxième partie sera consacrée à la population et aux paramètres estimés. Nous passerons ensuite de l'échelle de la population à l'échelle de l'arbre. Nous terminerons par une étude des prédictions des effets aléatoires.

5.1.1 Pins Laricio

5.1.1.1 Longueur de pousses annuelles

Cette partie fait suite à l'analyse exploratoire des données de pins Laricio où de nombreuses hypothèses biologiques ont été soulevées (section 1.3.5). Nous avons estimé une combinaison semi-markovienne de modèles linéaires mixtes avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle sur la base des longueurs de pousses annuelles des 4 sous-échantillons de pins Laricio (section 1.3.1 et figure 1.7). La semi-chaîne de Markov sous-jacente est supposée de type "gauche-droite" avec 3 états dont 2 états successifs transitoires suivis d'un état final absorbant (Guédon et al., 2007). Dans les régions tempérées, la pluie peut avoir un effet retardé d'un an (sur le nombre de feuilles) ou un effet immédiat (sur l'allongement de la pousse) selon qu'il se produit durant l'organogénèse (définition 1.1) ou l'allongement (définition 1.2) d'une plante. Aussi, nous avons choisi comme effets fixes, pour chaque modèle linéaire mixte, une constante et la pluie cumulée (en mm) sur une période recouvrant une période d'organogénèse et une période d'allongement. La covariable climatique est centrée (cf Fitzmaurice et al. (2004) pour une

discussion sur la question du centrage). Le modèle linéaire mixte attaché à l'état j est :

$$y_{at}|_{S_{at}=j} = \beta_{j1} + \beta_{j2} X_t + \tau_j \xi_{aj} + \epsilon_{at}, \qquad \xi_{aj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0, \sigma_j^2),$$

où y_{at} est la longueur de pousses annuelles pour l'individu a au temps t, β_{j1} est la constante, X_t est la pluie cumulée centrée au temps t ($\mathrm{E}(X_t) = 0$) et β_{j2} est le paramètre de régression de la pluie cumulée. Nous avons fait l'hypothèse d'un effet aléatoire différent pour chaque état. Grâce au centrage de la covariable climatique, la constante β_{j1} est directement interprétable comme la longueur moyenne des pousses annuelles successives dans l'état j.

L'algorithme MCEM avec une étape E de simulation-prédiction (section 4.3.4) a été initialisé avec les paramètres π, P, d, β et σ^2 estimés par l'algorithme EM pour combinaisons semi-markoviennes de modèles linéaires sans prendre en compte les effets aléatoires (i.e. $\xi = 0$). Après vérification de la convergence des prédictions des effets aléatoires, la convergence de l'algorithme est contrôlée par la différence entre deux itérations consécutives des log-vraisemblances des données observées sachant les effets aléatoires (équation (4.7)). Le tracé des valeurs de $g^{(k)}$ en fonction des itérations montre que l'algorithme d'estimation converge rapidement en environ 70 itérations (figure 5.1) avec $M_k = k$ séquences d'états simulées pour chaque individu à la $k^{ème}$ itération. Nous pouvons noter que les oscillations de $g^{(k)}$ sont fortes pour les premières itérations, diminuent nettement à la dixième itération puis progressivement par la suite.

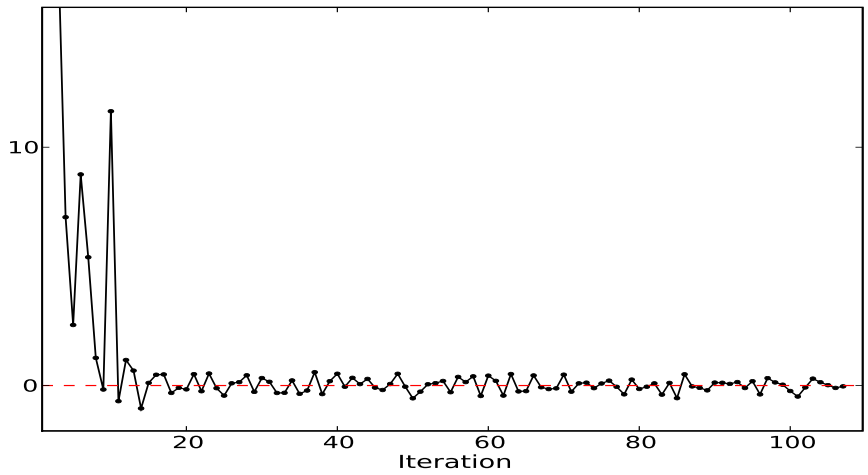

FIG. 5.1 – *Pins Laricio - hétérogénéité inter-individuelle : valeurs de la différence des log-vraisemblances des données observées sachant les effets aléatoires entre deux itérations consécutives de l'algorithme MCEM avec une étape E de simulation-prédiction.*

Propriétés de la population

Les états 1 et 2 sont les seuls états initiaux possibles (avec $\widehat{\pi}_1 = 0.95$ et $\widehat{\pi}_2 = 0.05$) de la semi-chaîne de Markov sous-jacente estimée ; cf figure 5.2. La matrice des probabilités

de transition est dégénérée ; pour chaque état transitoire i, $p_{i,i+1} = 1$ et $p_{ij} = 0$ pour $j \neq i+1$ (et pour l'état final absorbant $p_{ii} = 1$ et $p_{ij} = 0$ pour $j \neq i$). Par conséquent, la succession des états est déterministe pour la combinaison semi-markovienne de modèles linéaires mixtes de type "gauche-droite" dégénérée. Cette succession déterministe confirme l'hypothèse de succession de phases de croissance.

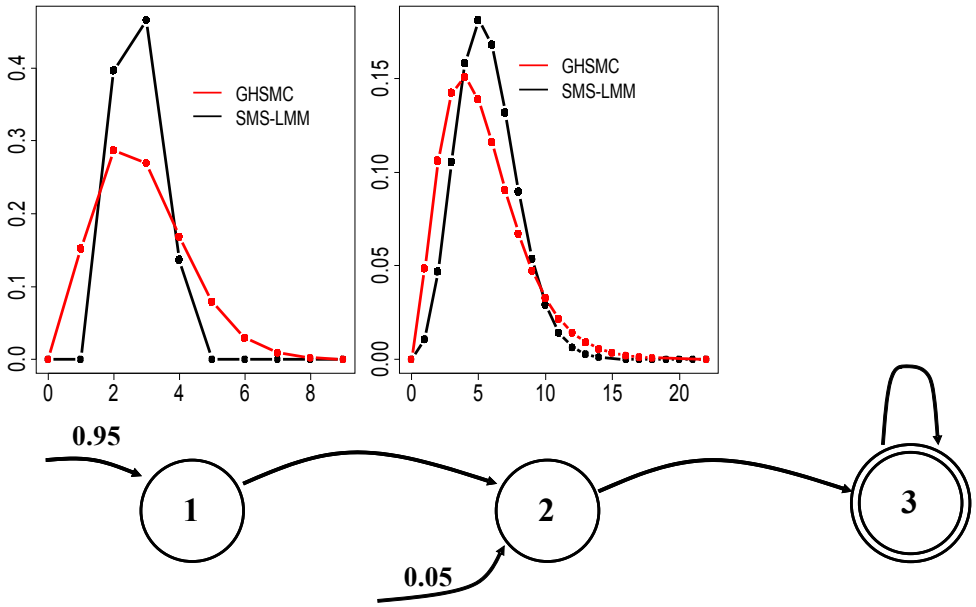

FIG. 5.2 – *Pins Laricio - hétérogénéité inter-individuelle : semi-chaîne de Markov sous-jacente estimée. Chaque état est représenté par un sommet numéroté. Les sommets représentant les états transitoires sont entourés par une simple ligne tandis que le sommet représentant l'état absorbant est entouré par une double ligne. Les transitions possibles entre états sont représentées par des arcs (les probabilités qui sont égales à 1 ne sont pas notées). Les arcs entrants dans les états indiquent les états initiaux. Les probabilités initiales associées sont notées à côté. Les lois d'occupation des états non absorbants sont figurées au dessus du sommet correspondant. Les lignes rouges correspondent aux lois d'occupation estimées pour une simple semi-chaîne de Markov cachée gaussienne (GHSMC) et les lignes noires correspondent aux lois d'occupation estimées pour une combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM).*

La distribution marginale du modèle linéaire mixte associé à l'état j est la distribution gaussienne $\mathcal{N}(\mu_j, \gamma_j^2)$ avec $\mu_j = \beta_{j1} + \beta_{j2} \mathrm{E}_j(X)$ et $\gamma_j^2 = \tau_j^2 + \sigma_j^2$, où $\mathrm{E}_j(X)$ est la moyenne de la pluie cumulée centrée X dans l'état j. La distribution marginale des observations représente la longueur de pousses annuelles moyenne dans l'état j. Les distributions marginales pour les différents états sont bien séparées (peu de recouvrement entre les distributions marginales correspondantes à deux états successifs) ; comparer la différence des moyennes $\mu_{j+1} - \mu_j$ entre états consécutifs et les écart-types γ_j et γ_{j+1} dans le tableau 5.1.

		état j		
		1	2	3
loi d'occupation moy., et.	GHSMC	P(1, 1.88)	NB(1, 4.36, 0.50)	
		2.88, 1.37	5.31, 2.93	
		B(2, 4, 0.37)	NB(1, 73.29, 0.94)	
	SMS-LMM	2.73, 0.68	5.56, 2.20	
paramètre de regression (SMS-LMM)	constante β_{j1}	7.09	25.79	50.25
	paramètre de la pluie cumulée β_{j2}	0.0027	0.0165	0.0309
	effet moyen de la pluie cumulée $\beta_{j2} \times sd_j(X)$	0.30	2.16	4.52
décomposition de la variance (SMS-LMM)	variance aléatoire τ_j^2	5.79	49.89	69.39
	variance résiduelle σ_j^2	4.74	39.95	76.86
	variance totale γ_j^2	10.53	89.84	146.25
	part d'hétérogénéité inter-individuelle	54.99%	55.53%	47.45%
distribution marginale μ_j, γ_j	GHSMC	6.97, 3.26	26.30, 9.12	54.35, 11.39
	SMS-LMM	6.99, 3.24	25.88, 9.48	50.32, 12.09

TAB. 5.1 – *Pins Laricio - hétérogénéité inter-individuelle : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante, le paramètre de régression, l'effet moyen de la pluie cumulée et la décomposition de la variance sont donnés.*

La prise en compte de l'influence de la covariable climatique et de l'hétérogénéité inter-individuelle rend les changements de phase de croissance plus synchrones entre arbres. Cette remarque se déduit de la comparaison des lois d'occupation des états estimées, et en particulier de leurs écarts-types, entre la semi-chaîne de Markov cachée gaussienne et la combinaison semi-markovienne de modèles linéaires mixtes (tableau 5.1 et figure 5.2). Nous pouvons noter que la probabilité de transition de l'état 1 à l'état 2 est égale à 1 : l'état 2 ne peut pas être "sauté". Cette succession déterministe des états est le résultat de l'estimation.

L'effet moyen de la pluie cumulée (i.e. l'amplitude moyenne des fluctuations climatiques) est donnée par $\beta_{j2} \times sd_j(X)$ pour chaque état j où $sd_j(X)$ représente l'écart-type de la pluie cumulée centrée sur l'état j. L'influence de la pluie cumulée est faible dans le premier état (où la croissance est lente) tandis qu'elle est forte dans les deux derniers états (un peu moins dans le deuxième état que dans le troisième état) ; voir le tableau 5.1.

La part d'hétérogénéité inter-individuelle, définie par le ratio entre la variance aléatoire τ_j^2 et la variance totale γ_j^2 dans l'état j, est plus forte au début de la vie de la plante (deux premiers états avec plus de 54%) et diminue légèrement dans le dernier état (environ 47%). Cette part importante d'hétérogénéité inter-individuelle peut être expliquée par l'effet de la transplantation des pins Laricio, par l'absence d'intervention sylvicole et par la stratégie d'échantillonnage (les arbres ont été choisis de façon à couvrir l'ensemble des classes de hauteur et de diamètre).

Comportement individuel

Les prédictions médianes des effets aléatoires sont calculées pour chaque individu et chaque état sur la base des prédictions des effets aléatoires à la dernière itération de l'algorithme MCEM avec une étape E de prédiction-simulation. La séquence d'états la plus probable sachant les effets aléatoires est calculée pour chaque séquence observée à l'aide d'une adaptation de l'algorithme de Viterbi pour semi-chaînes de Markov cachées (Guédon, 2003) aux combinaisons semi-markoviennes de modèles linéaires mixtes. La séquence d'états la plus probable restaurée peut être vue comme la segmentation optimale de la séquence observée correspondante en sous-séquences, chacune correspondant à un état donné.

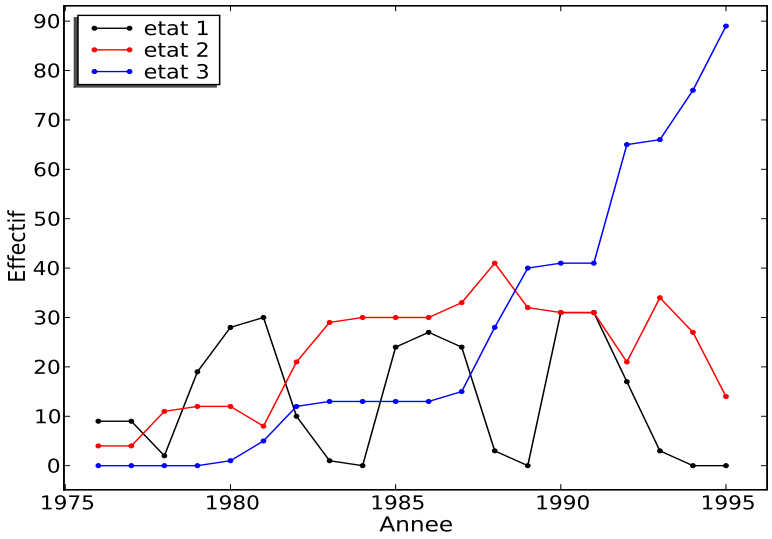

FIG. 5.3 – *Pins Laricio - hétérogénéité inter-individuelle : lois marginales empiriques des états par année extraites des séquences d'états les plus probables.*

Les lois marginales empiriques des états par année, extraites des séquences d'états les plus probables, sont représentées sur la figure 5.3. Cette représentation nous permet d'évaluer empiriquement le nombre de répétitions qui ont permis pour chaque état d'estimer les paramètres et l'hétérogénéité inter-individuelle dans la combinaison semi-markovienne de modèles linéaires mixtes. Le fort recouvrement des états sur les années permet d'obtenir des estimations de l'influence de la covariable climatique (qui varie dans le temps)

plus robuste. On retrouve tout de même la structure "gauche-droite" de la semi-chaîne de Markov sous-jacente. Les "vagues" pour l'état 1 correspondent au début de chaque sous-échantillon de pins Laricio. L'état 2 est représenté quasi-uniformément sur toutes les années. L'effectif pour l'état 3 augmente au fil des années.

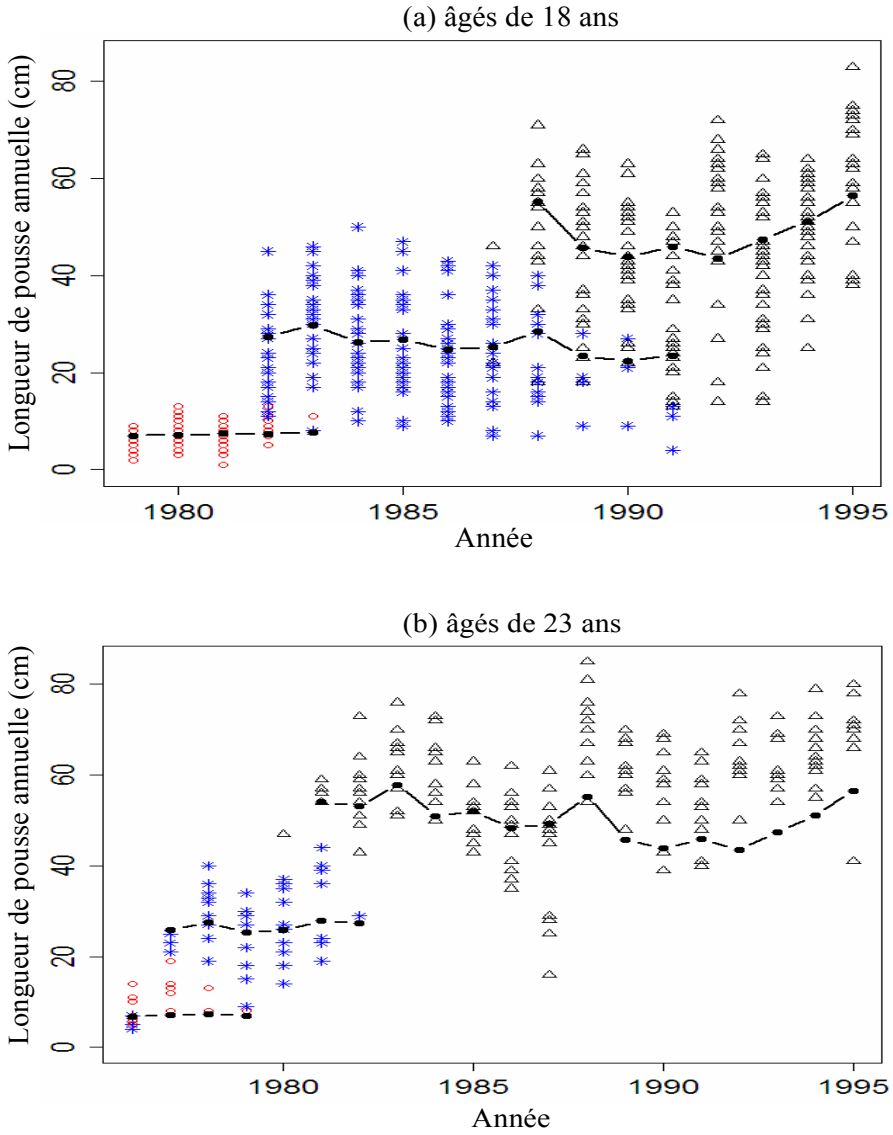

FIG. 5.4 – *Pins Laricio - hétérogénéité inter-individuelle : longueurs de pousses annuelles observées (état 1 : ○ ; état2 : ∗ ; état 3 : △) et partie fixe des trois modèles linéaires mixtes (i.e. $\beta_{j1} + \beta_{j2} X_t$ pour chaque état j) représentée par des lignes : (a)arbres âgés de 18 ans, (b)arbres âgés de 23 ans.*

Les parties fixes des trois modèles linéaires mixtes (i.e $\beta_{j1} + \beta_{j2} X_t$ pour chaque état j) pour les pins Laricio âgés de 18 ans et pour les pins Laricio âgés de 23 ans sont représentées sur la figure 5.4. Ceci confirme que les états sont bien séparés avec peu de recouvrement et que chaque changement de phase de croissance correspond à une

augmentation significative de la longueur de pousses annuelles moyenne. Le modèle mixte linéaire associé au troisième état sous-estime la longueur de pousses annuelles moyenne dans la phase de croissance correspondante pour les pins âgés de 23 ans. Ce comportement met l'accent sur un effet des sous-échantillons sur l'estimation des paramètres.

	âgés de 6 ans	âgés de 12 ans	âgés de 18 ans	âgés de 23 ans
GHSMC	1993 (0.56)	1988 (0.42)	1982 (1.93)	1978 (1.04)
état 2				
SMS-LMM	1993 (0.56)	1988 (0.50)	1982 (0.56)	1978 (1.17)
GHSMC	1995 (0.35)	1993 (1.00)	1988 (2.56)	1981 (0.78)
état 3				
SMS-LMM	1995 (0.38)	1992 (0.87)	1989 (1.27)	1982 (0.77)

TAB. 5.2 – *Pins Laricio - hétérogénéité inter-individuelle : première année médiane dans l'état 2 et dans l'état 3 pour chaque sous-échantillon déduite de la semi-chaîne de Markov cachée gaussienne estimée (GHSMC) et de la combinaison semi-markovienne de modèles linéaires mixtes estimée (SMS-LMM). Les écart-types correspondants sont indiqués entre parenthèses.*

Les caractéristiques (médiane et dispersion) de la première année dans chaque état sont extraites pour les quatre sous-échantillons de pins Laricio sur la base des séquences d'états les plus probables restaurées en utilisant soit la semi-chaîne de Markov cachée gaussienne estimée, soit la combinaison semi-markovienne de modèles linéaires mixtes estimée. La médiane de la première année dans le second état pour les quatre sous-échantillons est la même pour les deux modèles (cf tableau 5.2). La médiane de la première année dans le troisième état pour les pins Laricio âgés de 6 ans est la même pour les deux modèles. Il y a un décalage de 1 an de la médiane de la première année dans le troisième état entre les deux modèles pour les pins Laricio âgés de 12 ans, de 18 ans et de 23 ans. La dispersion de la première année dans le second état et dans le troisième état pour les pins âgés de 18 ans est fortement réduite dans le cas de la combinaison semi-markovienne de modèles linéaires mixtes par rapport au cas d'une semi-chaîne de Markov cachée gaussienne (tableau 5.2). Ceci confirme le fait que prendre en compte l'influence du climat et l'hétérogénéité inter-individuelle rend les changements de phase de croissance plus synchrones entre individus.

Etude des prédictions des effets aléatoires

Un intervalle de prédiction à 95% a été calculé afin de vérifier si l'influence de la prédiction médiane de chaque effet aléatoire pour chaque individu est significative ou pas. Cet intervalle de confiance à 95% est donné par (Hulting et Harville, 1991) :

$$[-t_{0.975}(N-1)\frac{sd(\xi_j)}{\sqrt{N}};\ t_{0.975}(N-1)\frac{sd(\xi_j)}{\sqrt{N}}]$$

groupe	comportement					nombre de
	état 1		état 2		état 3	pins Laricio
G1	$*$	$=$	$*$	$=$	$*$	29
G2	$*$	$=$	$*$	$\neq$	$*$	6
G3	$*$	$\neq$	$*$	$=$	$*$	13
G4	$*$	$\neq$	$*$	$\neq$	$*$	2
	non		$*$	$=$	$*$	11
	non		$*$	$\neq$	$*$	4
	$*$	$=$	$*$		non	9
	$*$	$\neq$	$*$		non	10
	$*$		non		$*$	6
	non		non		$*$	4
	non		$*$		non	2
	$*$		non		non	7

TAB. 5.3 – *Pins Laricio - hétérogénéité inter-individuelle : un intervalle de prédiction à 95% est calculé pour chaque état. Si les prédictions médianes de l'effet aléatoire n'appartiennent pas à cet intervalle, elles sont supposées significatives (*). Si les prédictions médianes de l'effet aléatoire appartiennent à cet intervalle, elles sont supposées non significatives (non). Le comportement, par rapport à l'individu moyen, entre deux états peut être identique (=) ou différent (≠). Cette classification est basée sur les prédictions médianes des effets aléatoires.*

où N est le nombre d'arbres, $sd(\xi_j)$ est l'écart-type empirique des prédictions médianes de l'effet aléatoire dans l'état j et $t_{0.975}(N-1)$ est le quantile d'ordre 0.975 de la loi de Student à $N-1$ degrés de liberté. Si la prédiction médiane de l'effet aléatoire pour un individu n'appartient pas à cet intervalle de prédiction, l'influence de cette prédiction est supposée significative ; c'est-à-dire qu'il y a une différence marquée entre cet individu et l'individu moyen dans l'état correspondant. Parmi les 103 pins Laricio, 50 ont une prédiction médiane de l'effet aléatoire significative pour chaque état (groupes G1+G2+G3+G4 dans le tableau 5.3) et parmi ces 50 individus, 29 (groupe G1) ont le même comportement sur tous les états par rapport à l'individu moyen (c'est-à-dire qu'ils sont soit toujours en dessous, soit toujours en dessus, soit toujours égaux). La corrélation entre les prédictions médianes de l'effet aléatoire pour l'état 1 et les prédictions médianes de l'effet aléatoire pour l'état 2 est égale à 0.26 alors que la corrélation entre les prédictions médianes de l'effet aléatoire pour l'état 2 et les prédictions médianes de l'effet aléatoire pour l'état 3 est égale à 0.62. Nous pouvons conclure que, comparé à l'individu moyen pour chaque état, chaque individu a un comportement plus étroitement lié sur les deux derniers états que sur les deux premiers. Par exemple, si un individu a une croissance plus lente par rapport à l'individu moyen sur le deuxième état, il y a de forte chance qu'il en soit de même sur le troisième état. La première phase de croissance correspond à l'établissement des arbres. La seconde phase de

croissance correspond à une chute ou à un rattrapage du statut social de l'arbre moyen. Le statut social de chaque arbre déclaré dans la deuxième phase de croissance peut être considéré comme irréversible et ne change pas dans la troisième phase de croissance. L'hypothèse générale d'un effet aléatoire différent pour chaque état comparée à un unique effet aléatoire pour toute la séquence observée est plus représentative du comportement des pins Laricio.

5.1.1.2 Nombre de branches par étage

On s'intéresse à une seconde caractéristique observée à l'échelle de la pousse annuelle pour chaque arbre des quatre sous-échantillons des pins Laricio : le nombre de branches par pousse annuelle (section 1.3.1). La distribution empirique du nombre de branches a pour moyenne 5.05 et pour variance 3.05. Bien que la plage des valeurs possibles ne soit pas relativement large, nous avons supposé que le nombre de branches par pousse annuelle suivait une loi gaussienne et nous avons donc considéré cette variable réponse comme continue.

Le nombre de branches par pousse annuelle est fortement corrélé à la longueur de pousses annuelles ($r^2 = 0.66$). C'est pourquoi, si nous introduisons la longueur de pousses annuelles comme covariable, il n'est pas nécessaire de modéliser explicitement les phases de croissance pour la variable réponse "nombre de branches". Un simple modèle linéaire mixte a ainsi été estimé où à la fois la longueur de pousses annuelles et le climat sont pris en compte comme covariables et modélisés comme effets fixes. Nous avons choisi comme covariable climatique, la pluie cumulée centrée (en mm) sur une période recouvrant la période d'organogénèse des pousses portées (Lanner, 1976). Nous avons vérifié que cette covariable climatique n'était pas corrélée avec la longueur de pousses annuelles ($r^2 = 0.07$). Le modèle linéaire mixte s'écrit :

$$y_{at} = \beta_1 + \beta_2 X_{at} + \beta_3 Z_t + \xi_a + \epsilon_{at}, \qquad \xi_a \sim \mathcal{N}(0, \tau^2), \qquad \epsilon_{at} \sim \mathcal{N}(0, \sigma^2),$$

où y_{at} est le nombre de branches pour l'individu a au temps t, X_{at} est la longueur de pousses annuelles pour l'individu a au temps t, Z_t est la pluie cumulée centrée au temps t, β_1 est la constante, β_2 et β_3 sont les paramètres de régression, ξ_a est l'effet aléatoire pour l'individu a, τ^2 est la variance aléatoire et σ^2 est la variance résiduelle. Ce modèle linéaire mixte a été estimé à l'aide de la fonction lme() du logiciel R[1].

La constante β_1 est significative (p-value $< 1.e-4$) et est égale à 3.08. L'influence de la pluie cumulée n'est pas significative (p-value ≈ 0.79) tandis que comme on pouvait s'y attendre, l'influence de la longueur de pousses annuelles est forte (p-value $< 1.e-4$). La non-significativité de la pluie cumulée montre qu'il n'y a pas de lien direct entre l'organogénèse des pousses portées et leur nombre. Par contre, la significativité de la

[1] http ://cran.r-project.org/

longueur de pousses annuelles permet d'émettre l'hypothèse que cette longueur joue le rôle de régulateur du nombre de branches : plus la pousse annuelle est longue, plus il y aura de branches. L'allongement de la pousse annuelle inhibe les bourgeons axillaires (qui sont à l'origine des rameaux portés) si les entre-noeuds sont trop courts. Le nombre de branches est déterminé lors du processus d'allongement. Comme la majeure partie de l'hétérogénéité inter-individuelle est déjà reflétée dans la covariable "longueur de pousses annuelles", la part d'hétérogénéité inter-individuelle estimée est inférieure à 14% : la variance résiduelle σ^2 est égale à 1.75 et la variance aléatoire τ^2 est égale à 0.27.

5.1.2 Chênes sessiles

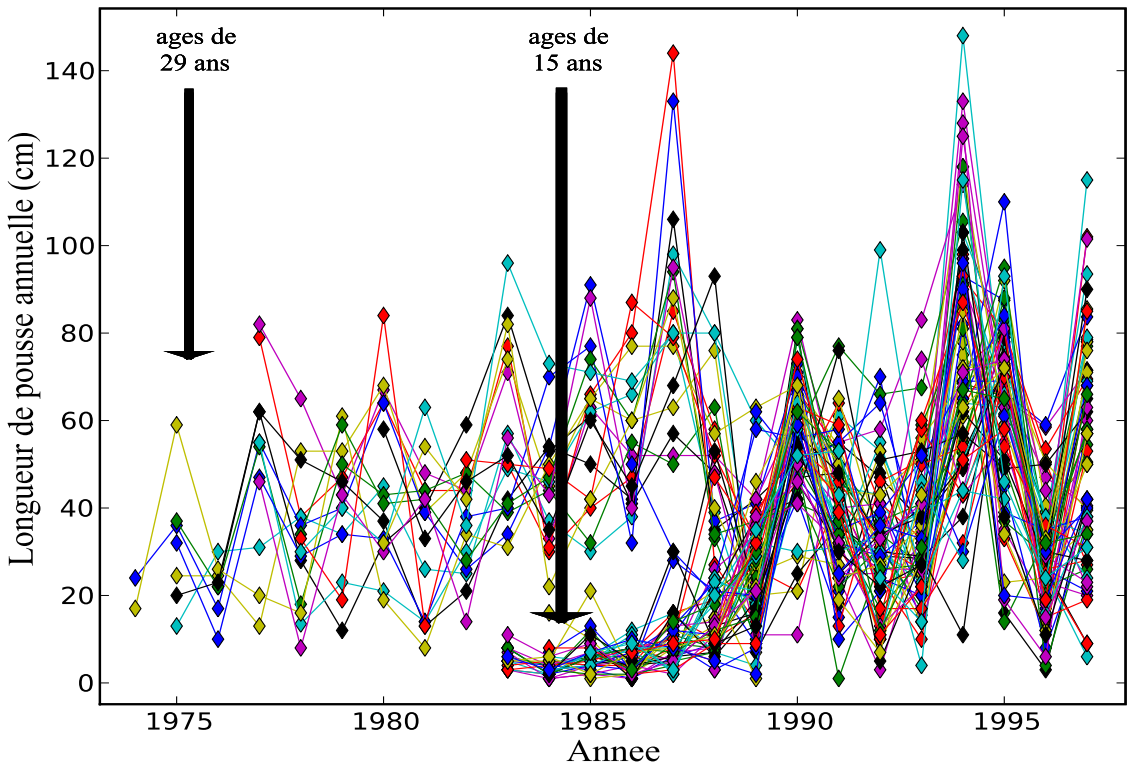

FIG. 5.5 – *Longueur des pousses annuelles successives le long du tronc des 2 sous-échantillons de chênes sessiles en fonction des années.*

Nous ne discutons pas dans cette partie de l'analyse exploratoire des données de chênes sessiles. Cette analyse nous a conduit aux mêmes hypothèses biologiques que pour les pins Laricio (section 1.3.5). Nous avons estimé une combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'hétérogénéité inter-individuelle sur la base des longueurs de pousses annuelles (en cm) des 2 sous-échantillons de chênes sessiles (section 1.3.2 et figure 5.5). La semi-chaîne de Markov sous-jacente est supposée de type "gauche-droite" à 2 états composés d'un état transitoire suivi d'un état absorbant (Guédon et al., 2007). Pour des raisons similaires aux pins Laricio, nous avons choisi comme effets

fixes, pour chaque modèle linéaire mixte, une constante et la pluie cumulée (en mm) centrée sur une période recouvrant l'organogénèse et l'allongement d'une pousse annuelle. Le modèle linéaire mixte associé à l'état j s'écrit :

$$y_{at}|_{S_{at}=j} = \beta_{j1} + \beta_{j2} X_t + \tau_j \xi_{aj} + \epsilon_{at}, \qquad \xi_{aj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0, \sigma_j^2),$$

où y_{at} est la longueur de pousses annuelles pour l'individu a au temps t, β_{j1} est la constante, X_t est la pluie cumulée centrée au temps t ($\mathrm{E}(X_t) = 0$) et β_{j2} est le paramètre de régression de la pluie cumulée. Nous avons fait l'hypothèse d'un effet aléatoire différent pour chaque état.

L'algorithme MCEM avec une étape E de simulation-prédiction (section 4.3.4) a été initialisé avec les paramètres π, P, d, β et σ^2 estimés par l'algorithme EM pour combinaisons semi-markoviennes de modèles linéaires sans prendre en compte les effets aléatoires (i.e. $\xi = 0$). Après vérification de la convergence des prédictions des effets aléatoires, la convergence de l'algorithme est contrôlée par la différence entre deux itérations consécutives des log-vraisemblances des données observées sachant les effets aléatoires. Le tracé des valeurs de $g^{(k)}$ en fonction des itérations montre que l'algorithme d'estimation converge rapidement en environ 45 itérations (figure 5.6) avec $M_k = k$ séquences d'états simulées pour chaque individu à la $k^{ème}$ itération.

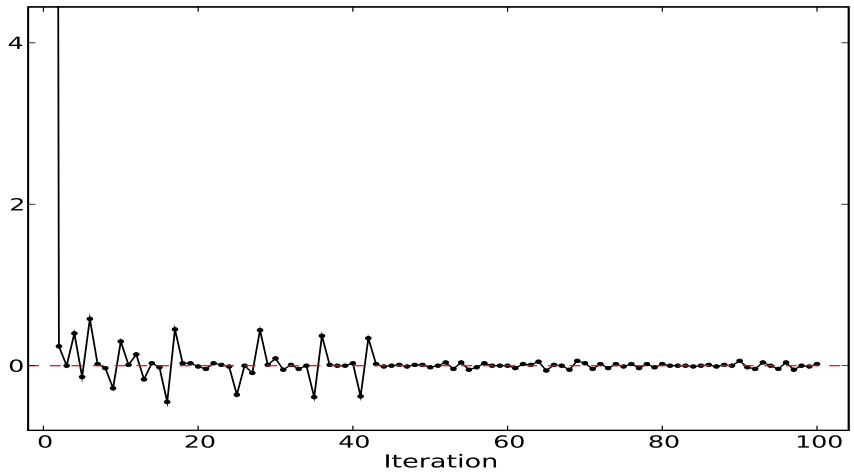

FIG. 5.6 – *Chênes sessiles - hétérogénéité inter-individuelle : valeurs de la différence des log-vraisemblances des données observées sachant les effets aléatoires entre deux itérations consécutives de l'algorithme MCEM avec une étape E de simulation-prédiction.*

Propriétés de la population

La forte séparabilité des distributions marginales pour les différents états est encore plus marquée chez les chênes sessiles que chez les pins Laricio (peu de recouvrement entre les distributions marginales) ; comparer la différence des moyennes $\mu_{j+1} - \mu_j$ entre états consécutifs et les écart-types γ_j et γ_{j+1} dans le tableau 5.4.

		état j	
		1	2
loi d'occupation moy., et.	GHSMC	B(1, 8, 0.52) 4.64, 1.32	
	SMS-LMM	B(1, 7, 0.63) 4.77, 1.18	
paramètre de regression (SMS-LMM)	constante β_{j1}	6.07	45.98
	paramètre de la pluie cumulée β_{j2}	0.0115	0.0739
	effet moyen de la pluie cumulée $\beta_{j2} \times sd_j(X)$	1.81	11.51
décomposition de la variance (SMS-LMM)	variance aléatoire τ_j^2	2.02	33.45
	variance résiduelle σ_j^2	10.64	416.33
	variance totale γ_j^2	12.66	449.78
	part d'hétérogénéité inter-individuelle	15.96%	7.4%
distribution marginale μ_j, γ_j	GHSMC	6.35, 3.79	45.18, 24.18
	SMS-LMM	6.32, 3.56	45.98, 21.21

TAB. 5.4 – *Chênes sessiles - hétérogénéité inter-individuelle : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante, le paramètre de régression, l'effet moyen de la pluie cumulée et la décomposition de la variance sont donnés.*

Comme pour les pins Laricio, les changements de phase de croissance sont plus synchrones entre individus pour la combinaison semi-markovienne de modèles linéaires mixtes que pour une simple semi-chaîne de Markov cachée gaussienne (tableau 5.4). Néanmoins, le gain en synchronisme est moins important que pour les pins Laricio ; cf les caractéristiques des lois d'occupation des états dans les tableaux 5.1 et 5.4.

L'influence de la pluie cumulée est faible dans le premier état (où la croissance est lente) tandis qu'elle est beaucoup plus forte dans le second état (tableau 5.4). Comme pour les pins Laricio, l'influence de la covariable climatique est à peu près proportionnelle au niveau de croissance. Ce résultat se déduit de l'observation des variations conjointes du paramètre de régression β_{j2} de la pluie cumulée centrée et de la constante β_{j1} avec l'état j dans le tableau 5.1 et le tableau 5.4. L'influence relative de la pluie cumulée est beaucoup plus forte pour les chênes sessiles que pour les pins Laricio ; en effet, les ratios

$\beta_{j2} \times sd_j(X)/\beta_{j1}$ dans le tableau 5.4 sont égaux à plus de 2,8 fois les ratios correspondant dans le tableau 5.1.

Contrairement aux pins Laricio, la variabilité totale est à peu près proportionnelle au niveau de croissance ; cf μ_j et γ_j dans le tableau 5.1 et le tableau 5.4. La part d'hétérogénéité inter-individuelle est plus forte au début de la vie de la plante (premier état). Cependant, cette part d'hétérogénéité inter-individuelle d'environ 16% dans le premier état (où la croissance est faible) et inférieure à 8% dans le second état (où la croissance est forte) reste faible. Cela peut être expliqué par le fait que les individus ont été échantillonnés parmi les arbres dominants et codominants (et non parmi l'ensemble des classes de hauteur et de diamètre pour les pins Laricio), par l'intervention sylvicole (coupe d'éclaircie) et par la régénération naturelle synchrone entre individus d'un même sous-échantillon (alors que les pins Laricio provenaient de pépinière et ont été transplantés après des durées d'élevage et selon des méthodes d'élevage différentes).

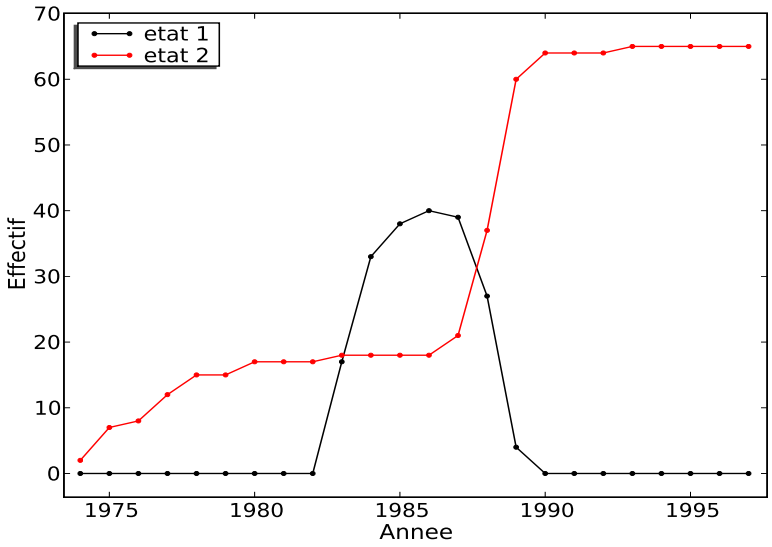

FIG. 5.7 – *Chênes sessiles - hétérogénéité inter-individuelle : lois marginales empiriques des états par année extraites des séquences d'états les plus probables.*

Comportement individuel

Comme pour les pins Laricio, la séquence d'états la plus probable a été restaurée pour chaque individu à l'aide de l'algorithme de Viterbi pour SMS-LMM. Cette restauration est basée sur les prédictions médianes des effets aléatoires pour chaque individu. Les lois marginales empiriques des états par année, extraites des séquences d'états les plus probables, sont représentées sur la figure 5.7. Nous pouvons remarquer que les arbres âgés de 29 ans sont dès le début de leur observation dans le second état. Ce phénomène peut avoir deux causes : soit les données n'ont pas été mesurées jusqu'au pied de l'arbre (et donc les premières années ont été omises), soit les marqueurs morphologiques du début

de la vie de la plante étaient peu visibles (et donc les données des premières années de vie des arbres correspondent à des cumuls de plusieurs années non repérées). De part le fort recouvrement des deux états, l'estimation de l'influence de la pluie cumulée et de l'hétérogénéité inter-individuelle semble robuste.

La partie fixe des deux modèles linéaires mixtes (i.e. $\beta_{j1} + \beta_{j2} X_t$ pour chaque état j) pour les arbres âgés de 15 ans et pour les arbres âgés de 29 ans est représentée sur la figure 5.8. Seul le second état est représenté pour les chênes sessiles âgés de 29 ans.

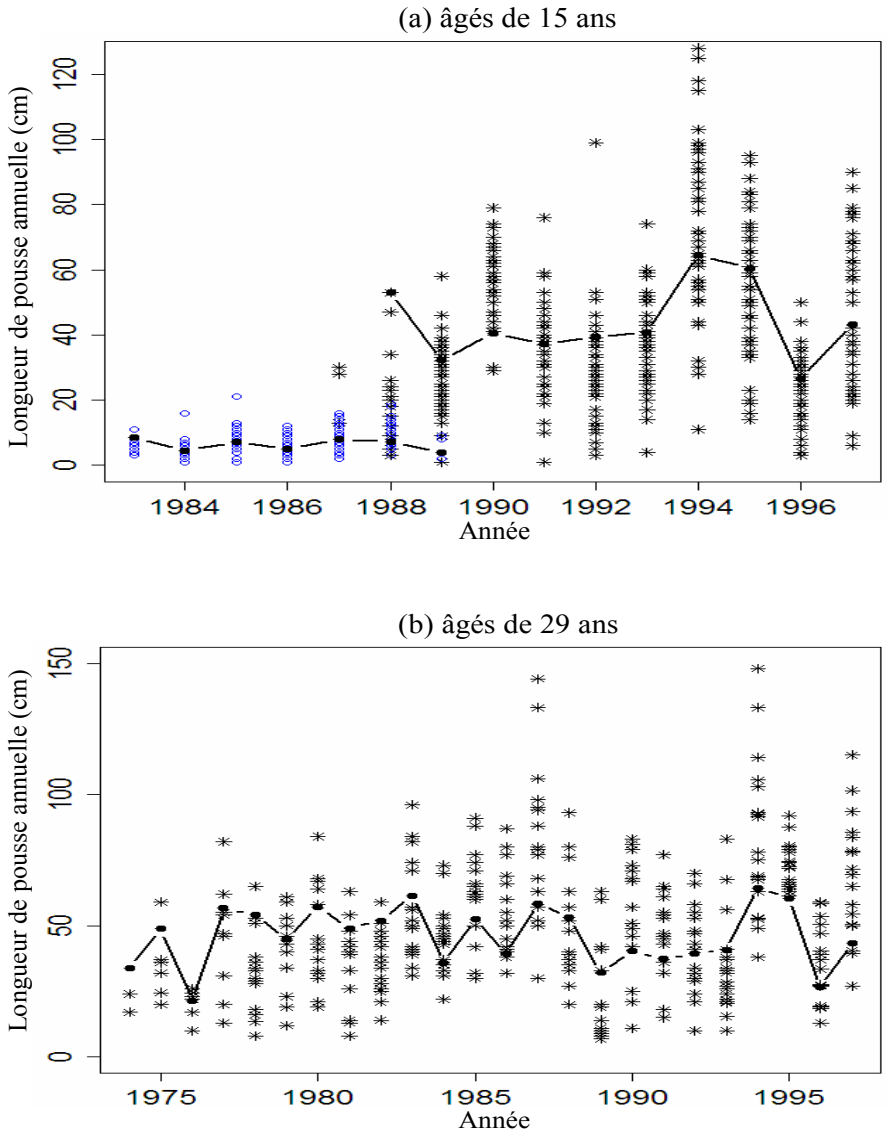

FIG. 5.8 – *Chênes sessiles - hétérogénéité inter-individuelle : longueurs de pousses annuelles observées (état 1 : ∘ ; état 2 : ∗) et partie fixe des trois modèles linéaires mixtes (i.e. $\beta_{j1} + \beta_{j2} X_t$ pour chaque état j) représentée par des lignes : (a)arbres âgés de 15 ans, (b)arbres âgés de 29 ans.*

Etude des prédictions des effets aléatoires

Nous avons également comparé le comportement des arbres âgés de 15 ans entre les deux états. Comme pour les pins Laricio, un intervalle de prédiction à 95% a été calculé pour chaque individu et chaque état sur la base des prédictions médianes des effets aléatoires. Parmi les 20 chênes sessiles qui ont une prédiction médiane de l'effet aléatoire significative pour chaque état, 12 ont le même comportement sur les deux états par rapport à l'individu moyen. La corrélation entre les prédictions médianes de l'effet aléatoire pour l'état 1 et les prédictions médianes de l'effet aléatoire pour l'état 2 est égale à 0.13. Comme pour les pins Laricio, l'hypothèse générale d'un effet aléatoire différent pour chaque état comparée à un unique effet aléatoire pour toute la séquence observée est plus représentative du comportement des chênes sessiles.

5.1.3 Pins sylvestres

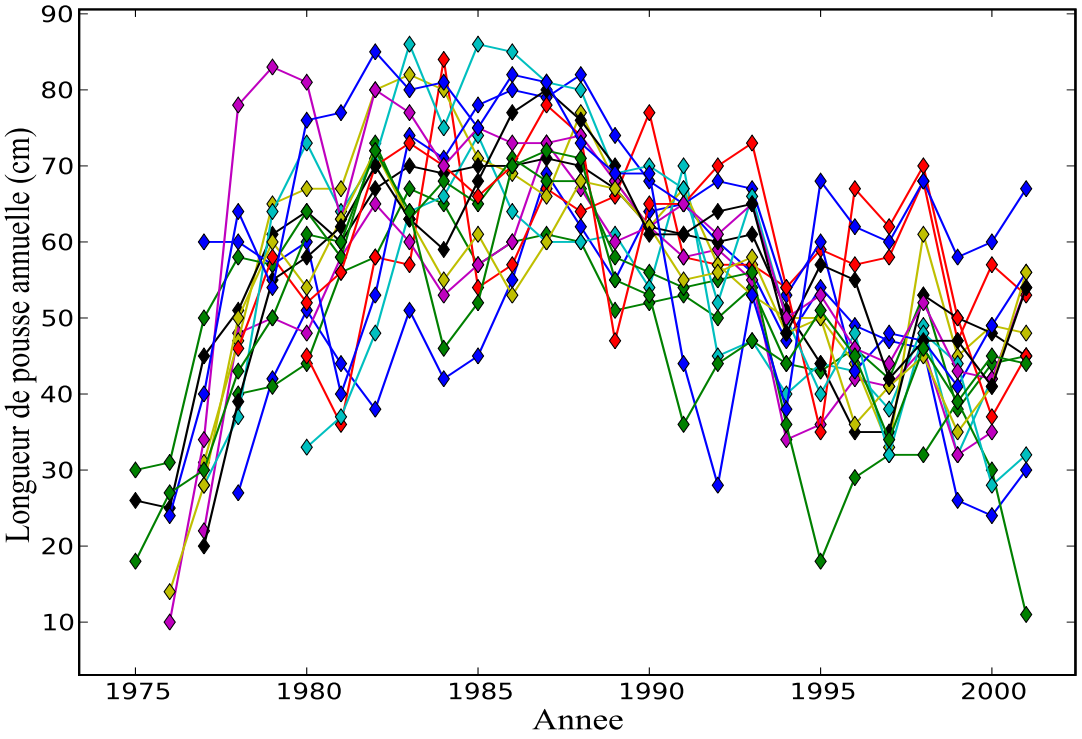

FIG. 5.9 – *Longueur des pousses annuelles successives le long des troncs des pins sylvestres en fonction des années.*

Comme pour les chênes sessiles, nous ne présentons pas l'analyse exploratoire des pins sylvestres précédant cette analyse. On s'intéresse aux longueurs de pousses annuelles de 16 pins sylvestres âgés de 30 ans (section 1.3.4 et figure 5.9). L'intérêt de cet exemple est le fait que les longueurs de pousses annuelles de ces arbres sont caractérisées par une succession de phases d'augmentation de la croissance moyenne suivie par une phase de diminution de la croissance moyenne (contrairement aux pins Laricio et aux chênes sessiles caractérisés

par seulement une succession de phases d'augmentation de la croissance moyenne). De plus, une intervention sylvicole (coupe d'éclaircie) a été appliquée à ces pins sylvestres.

Le choix de la modélisation et de l'estimation est similaire à celui fait pour les pins Laricio : une combinaison semi-markovienne de modèles linéaires mixtes de type "gauche-droite" à 3 états, un effet aléatoire "hétérogénéité inter-individuelle" différent pour chaque état et une estimation basée sur l'algorithme MCEM avec une étape E de simulation-prédiction. Après vérification de la convergence des prédictions des effets aléatoires, l'algorithme d'estimation converge rapidement en environ 20 itérations avec $M_k = k$ séquences d'états simulées pour chaque individu à la $k^{ème}$ itération. L'état 1 est le seul état initial possibles (avec $\widehat{\pi}_1 = 1$) de la semi-chaîne de Markov sous-jacente estimée. La matrice des probabilités de transition est dégénérée, ce qui entraîne une succession déterministe des états. Cette succession déterministe confirme l'hypothèse de succession de phases de croissance.

		état j		
		1	2	3
loi d'occupation moy., et.	GHSMC	P(1, 2.33) 3.33, 1.53	B(1, 25, 0.49) 12.72, 2.45	
	SMS-LMM	NB(2, 0.38, 0.32) 2.82, 1.60	B(1, 19, 0.72) 14.04, 1.90	
paramètre de regression (SMS-LMM)	constante β_{j1}	36.38	64.99	45.02
	paramètre de la pluie cumulée β_{j2}	0.0062	0.0102	0.008
	effet moyen de la pluie cumulée $\beta_{j2} \times sd_j(X)$	1.12	1.58	1.35
décomposition de la variance (SMS-LMM)	variance aléatoire τ_j^2	93.70	35.06	64.60
	variance résiduelle σ_j^2	49.95	51.78	53.66
	variance totale γ_j^2	143.65	86.84	118.26
	part d'hétérogénéité inter-individuelle	65.23%	40.37%	53.93%
distribution marginale μ_j, γ_j	GHSMC	40.11, 12.67	65.96, 8.76	45.95, 10.12
	SMS-LMM	36.41, 11.99	65.03, 9.32	44.95, 10.87

TAB. 5.5 – *Pins sylvestres - hétérogénéité inter-individuelle : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante, le paramètre de régression, l'effet moyen de la pluie cumulée et la décomposition de la variance sont donnés.*

L'influence de la pluie cumulée est proche dans les trois états (avec une légère augmentation entre les deux premiers états et une légère diminution entre les deux derniers états) ; cf le tableau 5.5. La part d'hétérogénéité inter-individuelle est forte dans le premier état (environ 65%), diminue dans le deuxième état (environ 40%) puis augmente dans le troisième état (environ 54%). La séquence d'états la plus probable a été restaurée pour chaque individu par l'algorithme de Viterbi pour SMS-LMM. Cette restauration est basée sur les prédictions médianes des effets aléatoires pour chaque individu. La première année dans le troisième état est 1994 pour 10 pins sylvestres parmi les 16 pins sylvestres. L'année 1994 suit la coupe d'éclaircie de 1993. La diminution synchrone entre les arbres de la longueur de pousses annuelles (figure 5.9) est liée au stress engendré par l'intervention sylvicole. Ce changement d'ambiance nécessite une période d'acclimatation qui peut varier d'un arbre à un autre ; période où l'acclimatation est privilégiée au dépend de l'influence du climat. Aussi, l'augmentation de la part d'hétérogénéité inter-individuelle et la légère diminution de l'influence du climat pourraient être expliquées par la coupe d'éclaircie. Cette intervention sylvicole a augmenté significativement l'hétérogénéité inter-individuelle et peut être interprétée comme une remise à zéro de la compétition entre les arbres.

La corrélation entre les prédictions médianes de l'effet aléatoire pour l'état 1 et les prédictions médianes de l'effet aléatoire pour l'état 2 est égale à 0.10 tandis que la corrélation entre les prédictions médianes de l'effet aléatoire pour l'état 2 et les prédictions médianes de l'effet aléatoire pour l'état 3 est égale à 0.54. Comme pour les pins Laricio, nous pouvons conclure que le comportement d'un individu par rapport à l'individu moyen est plus étroitement lié sur les deux derniers états que sur les deux premiers. L'hypothèse générale d'un effet aléatoire différent pour chaque état comparée à un unique effet aléatoire pour toute la séquence observée est plus représentative du comportement des pins sylvestres.

5.1.4 Noyers

Les noyers communs mesurés se répartissent en deux sous-échantillons : 22 noyers ayant au moins ramifié une fois au cours de leur vie et 116 noyers n'ayant jamais ramifié. Les longueurs de pousses annuelles (en cm) ont été mesurées rétrospectivement pour chaque arbre (section 1.3.3 et figure 5.10).

Par analogie avec les pins Laricio, les chênes sessiles et les pins sylvestres (Guédon et al., 2007), nous avons étudié la structure de la composante ontogénique sous-jacente. Cette analyse a été faite à l'aide du logiciel VPlants[2] (packages stat-tool et sequence-analysis [3]) et du langage AML[4]. Les résultats obtenus nous ont amené à choisir comme

[2]http ://www-sop.inria.fr/virtualplants/wiki/doku.php ?id=software
[3]http ://openalea.gforge.inria.fr/dokuwiki/doku.php ?id=packages :vplants :vplants
[4]http ://openalea.gforge.inria.fr/dokuwiki/doku.php ?id=packages :vplants :aml :aml

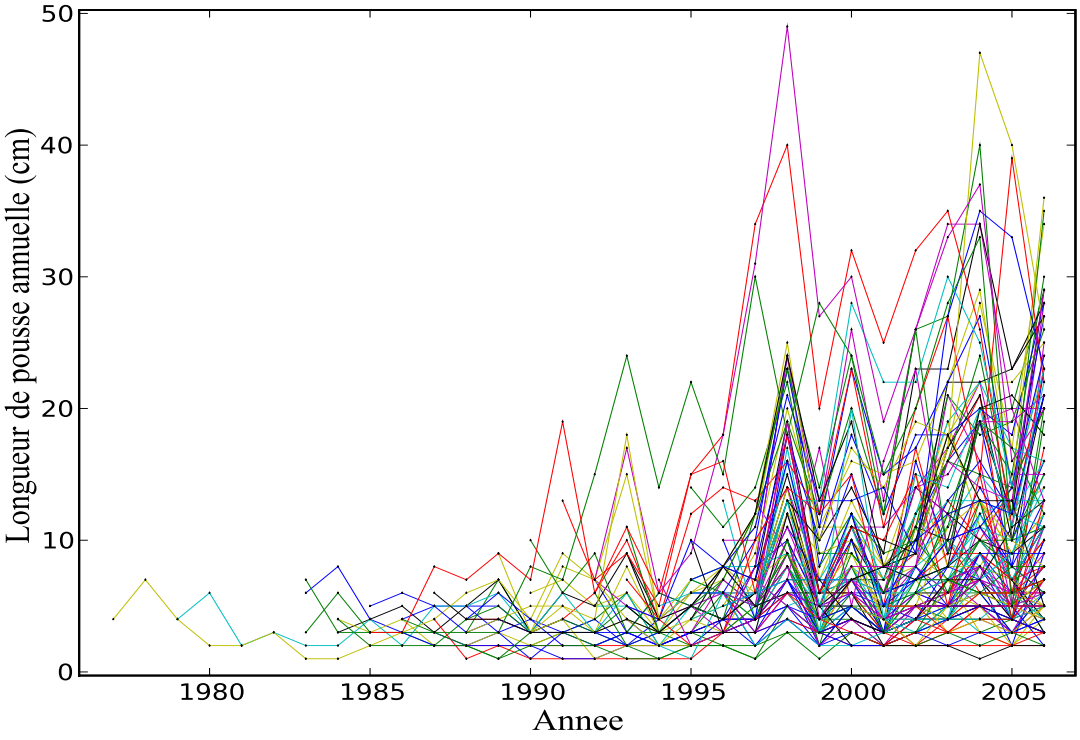

FIG. 5.10 – *Longueur des pousses annuelles successives le long du tronc des 2 sous-échantillons de noyers communs en fonction des années.*

processus sous-jacent, une semi-chaîne de Markov cachée de type "gauche-droite" à 3 états avec des changements de rythme de croissance nets entre les phases de croissance. Cette hypothèse a été confirmée par le calcul de la fonction d'autocorrélation à partir des séquences résiduelles obtenues en soustrayant des données observées la fonction en escalier déduite de la semi-chaîne de Markov cachée estimée. Nous avons ensuite extrait la fonction d'autocorrélation correspondant à chaque état. Ces fonctions ont montré que les sous-séquences résiduelles étaient stationnaires et proche de séquences de bruits blancs pour chaque état.

Nous avons estimé une combinaison semi-markovienne de modèles linéaires mixtes de type "gauche-droite" à 3 états sur la base des 2 sous-échantillons de noyers. L'intérêt de cet exemple est de déterminer si d'autres variables climatiques que la pluie cumulée peuvent jouer un rôle sur la croissance d'arbres. Nous avons choisi en plus de la constante, trois covariables : la présence/absence de ramification, la pluie cumulée (en mm) et la température maximale moyenne (en °C) sur une période couvrant l'organogénèse d'une pousse annuelle de noyer commun. Tout comme pour les exemples précédents, les covariables climatiques (précipitation et température) ont été centrées. Le modèle linéaire mixte attaché à l'état j s'écrit :

$$y_{at}|_{S_{at}=j} = \beta_{j1} + \beta_{j2}Z_a + \beta_{j3}X_{1t} + \beta_{j4}X_{2t} + \tau_j\xi_{aj} + \epsilon_{at},$$

$$\xi_{aj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2),$$

où y_{at} est la longueur de pousse de l'individu a au temps t, Z_a est la présence/absence de ramification, X_{1t} est la pluie cumulée centrée au temps t ($\mathrm{E}(X_{1t}) = 0$), X_{2t} est la température maximale moyenne centrée au temps t ($\mathrm{E}(X_{2t}) = 0$), $\beta_{j1}, \beta_{j2}, \beta_{j3}$ et β_{j4} sont les paramètres de régression. Nous avons fait l'hypothèse d'un effet aléatoire modélisant l'hétérogénéité inter-individuelle différent pour chaque état.

L'algorithme MCEM avec une étape E de simulation-prédiction a été initialisé avec les paramètres π, P, d, β et σ^2 estimés par l'algorithme EM pour combinaisons semi-markoviennes de modèles linéaires sans prendre en compte les effets aléatoires (i.e. $\xi = 0$). Après vérification de la convergence des prédictions des effets aléatoires, la convergence de l'algorithme est contrôlée par la différence entre deux itérations consécutives des log-vraisemblances des données observées sachant les effets aléatoires (équation (4.7)). L'algorithme d'estimation converge en environ 90 itérations avec $M_k = k$ séquences d'états simulées pour chaque individu à la $k^{ème}$ itération.

Propriétés de la population

Les états 1 et 2 sont les seuls états initiaux possibles (avec $\widehat{\pi}_1 = 0.64$ et $\widehat{\pi}_2 = 0.36$) de la semi-chaîne de Markov sous-jacente estimée. La matrice des probabilités de transition est dégénérée. Par conséquent, la succession déterministe des états confirme l'hypothèse de succession de phases de croissance.

La distribution marginale du modèle linéaire mixte associé à l'état j est la distribution gaussienne $\mathcal{N}(\mu_j, \gamma_j^2)$ avec $\mu_j = \beta_{j1} + \beta_{j3}\mathrm{E}_j(X_1) + \beta_{j4}\mathrm{E}_j(X_2)$ et $\gamma_j^2 = \tau_j^2 + \sigma_j^2$, où $\mathrm{E}_j(X_1)$ est la moyenne de la pluie cumulée centrée X_1 dans l'état j et $\mathrm{E}_j(X_2)$ est la moyenne de la température maximale moyenne centrée X_2 dans l'état j. La distribution marginale des observations représente la longueur de pousses annuelles moyenne dans l'état j. Les distributions marginales pour les différents états sont bien séparées (peu de recouvrement entre les distributions marginales correspondantes à deux états successifs) ; comparer la différence des moyennes $\mu_{j+1} - \mu_j$ entre états consécutifs et les écart-types γ_j et γ_{j+1} dans le tableau 5.6.

La prise en compte de l'hétérogénéité inter-individuelle et des covariables (pluie, température et présence/absence de ramification) modifie la semi-chaîne de Markov sous-jacente et notamment les temps de séjour dans chaque état par rapport à la semi-chaîne de Markov cachée gaussienne (tableau 5.6). En effet, les temps de séjour dans les deux premiers états sont plus longs ; une différence plus nette est à noter dans le deuxième état. Cependant, bien que les écart-types soient plus forts, les ratios de la moyenne sur l'écart-type sont conservés dans chaque état. L'augmentation de la variabilité dans les lois d'occupation est partiellement compensée par une diminution de la variabilité dans les distributions marginales des modèles linéaires mixtes ; comparer γ_j entre la combinaison

		état j		
		1	2	3
loi d'occupation moy., et.	GHSMC	NB(1, 1.79, 0.14) 12.40, 9.16	NB(1, 1.40, 0.17) 7.74, 6.27	
	SMS-LMM	NB(2, 1.96, 0.13) 15.01, 9.96	NB(1, 1.39, 0.09) 14.77, 12.27	
paramètre de regression (SMS-LMM)	constante β_{j1}	3.99	5.10	9.38
	ramification β_{j2}	0.78	1.86	4.97
	paramètre de la pluie cumulée β_{j3}	0.0007	0.0028	0.0040
	effet moyen de la pluie cumulée $\beta_{j3} \times sd_j(X_1)$	0.15	0.64	0.90
	paramètre de la température β_{j4}	0.13	2.17	4.58
	effet moyen de la température $\beta_{j4} \times sd_j(X_2)$	0.18	1.93	3.65
décomposition de la variance (SMS-LMM)	variance aléatoire τ_j^2	0.77	4.62	11.48
	variance résiduelle σ_j^2	1.58	3.96	28.18
	variance totale γ_j^2	2.35	8.58	39.66
	part d'hétérogénéité inter-individuelle	32.77%	53.85%	28.95%
distribution marginale μ_j, γ_j	GHSMC	3.99, 1.66	7.70, 3.08	17.62, 7.89
	SMS-LMM : sans ramif.	3.99, 1.53	6.92, 2.93	14.40, 6.30

TAB. 5.6 – *Noyers - hétérogénéité inter-individuelle : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante, les paramètres de régression, l'effet moyen de la pluie cumulée, l'effet moyen de la température et la décomposition de la variance sont donnés.*

markovienne de modèles linéaires mixtes et la semi-chaîne de Markov cachée gaussienne pour chaque état j dans le tableau 5.6.

L'effet moyen de la pluie cumulée (i.e. l'amplitude moyenne des fluctuations pluviométriques) est calculée par $\beta_{j3} \times sd_j(X_1)$ pour chaque état j où $sd_j(X_1)$ représente l'écart-type de la pluie cumulée centrée sur l'état j. L'influence de la pluie cumulée est faible dans le premier état et augmente lentement au fil de la vie de la plante (tableau 5.6). L'effet

moyen de la température maximale moyenne (i.e. l'amplitude moyenne des fluctuations thermiques) est calculée par $\beta_{j4} \times sd_j(X_2)$ pour chaque état j où $sd_j(X_2)$ représente l'écart-type de la température maximale moyenne centrée sur l'état j. L'influence de la température maximale moyenne est faible dans le premier état tandis qu'elle est forte sur les deux derniers états (plus forte dans le troisième état que dans le deuxième) (tableau 5.6). L'influence de la température maximale moyenne sur les longueurs de pousses annuelles est beaucoup plus prononcée que celle de la pluie cumulée sur les deux derniers états ; comparer $\beta_{j4} \times sd_j(X_2)$ et $\beta_{j3} \times sd_j(X_1)$ dans le tableau 5.6. La présence de ramification apporte plus de 20% de longueur en plus sur chaque état ; voir le ratio β_{j2}/μ_j pour chaque état j dans le tableau 5.6.

La part d'hétérogénéité inter-individuelle varie fortement au cours de la vie de ces noyers. La part d'hétérogénéité inter-individuelle augmente fortement entre les deux premiers états puis diminue d'autant entre les deux derniers états. Le bois dans lequel poussent ces noyers n'est pas entretenu et présente une grande diversité d'espèces (pins, chênes, tilleuls, ...). L'augmentation de la part d'hétérogénéité entre les deux premières phases de croissance peut être expliquée par la lutte pour survivre des noyers dans cet environnement. Les noyers n'ont pas la même façon d'arriver au même niveau de croissance ; ils se différencient énormément par leurs modalités de survie. Après avoir atteint une certaine hauteur (pouvant être vue comme un seuil de survie), le rythme de croissance s'homogénéise entre les arbres d'où une forte diminution de l'hétérogénéité inter-individuelle entre les deux dernières phases de croissance.

Comportement individuel

Les prédictions médianes des effets aléatoires sont calculées pour chaque individu sur la base des prédictions des effets aléatoires à la dernière itération de l'algorithme MCEM avec une étape E de prédiction-simulation. La séquence d'états la plus probable sachant les prédictions médianes des effets aléatoires est calculée pour chaque séquence observée à l'aide d'une adaptation de l'algorithme de Viterbi pour semi-chaînes de Markov cachées (Guédon, 2003) aux combinaisons semi-markoviennes de modèles linéaires mixtes. Parmi les 22 noyers ayant au moins ramifié une fois au cours de leur vie, 22 atteignent le troisième état tandis que parmi les 116 noyers n'ayant jamais ramifié, seulement 32 atteignent le troisième état. La ramification peut être considérée comme un marqueur du passage à un stade de différenciation architectural supérieur, phénomène déjà mis en évidence chez le noyer (Barthélémy et al., 1995) et chez le hêtre (Nicolini, 1998).

5.2 Analyse conjointe de la composante ontogénique et de la composante environnementale

Cette partie est consacrée à la distinction et à la caractérisation de seulement deux composantes de la croissance d'une plante en modélisant la composante ontogénique par une semi-chaîne de Markov sous-jacente et la composante environnementale variant dans le temps par un effet aléatoire dans le processus d'observation. Cet effet aléatoire représente en fait l'environnement commun à tous les individus chaque année. Nous avons vu dans la section 4.4.7 que la modélisation de l'environnement commun par un effet aléatoire posait des problèmes d'identifiabilité lorsque le degré de synchronisme du changement d'état entre individus était trop élevé ; les simulations de l'effet aléatoire avait tendance à "éponger" les changements d'état. Pour pallier ce problème, nous avons utilisé plusieurs jeux de données d'une même espèce mais à des âges ontogéniques différents afin de mélanger les phases de croissance pour chaque âge chronologique (ou année) ; voir la section 1.1 pour les notions d'âge ontogénique et d'âge chronologique.

Nous suivrons un schéma identique pour la présentation des différents résultats pour chaque espèce forestière. On s'intéressera dans un premier temps au modèle choisi pour l'analyse et aux caractéristiques de l'algorithme d'estimation proposé. Une deuxième partie sera consacrée à la population et aux paramètres estimés. Nous passerons ensuite à l'étude du comportement individuel. Nous terminerons par une étude des simulations des effets aléatoires et par une comparaison avec les covariables utilisées dans la section précédente pour les pins Laricio et les chênes sessiles.

5.2.1 Pins Laricio

Sous les hypothèses biologiques établies à la section 1.3.5, une combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus à chaque âge chronologique (ou date) est estimée sur la base des longueurs de pousses annuelles des 4 sous-échantillons de pins Laricio. Comme pour le cas des effets aléatoires modélisant l'hétérogénéité inter-individuelle (section 5.1.1.1), la semi-chaîne de Markov sous-jacente est supposée de type "gauche-droite" à 3 états. Seule une constante est choisie comme effet fixe pour chaque modèle linéaire mixte. Le modèle linéaire mixte attaché à l'état j est donné par :

$$y_{ad}|_{S_{a,t_a(d)}=j} = \beta_{j1} + \varsigma_j \lambda_d + \epsilon_{ad}, \qquad \lambda_d \sim \mathcal{N}(0,1), \qquad \epsilon_{ad}|_{S_{a,t_a(d)}=j} \sim \mathcal{N}(0, \sigma_j^2),$$

où y_{ad} est la longueur de pousses annuelles pour l'individu a à l'âge chronologique (ou date) d et β_{j1} est la constante associée à l'état j. Du fait de l'absence d'autres covariables, la constante β_{j1} est directement interprétable comme la longueur moyenne des pousses annuelles successives dans l'état j.

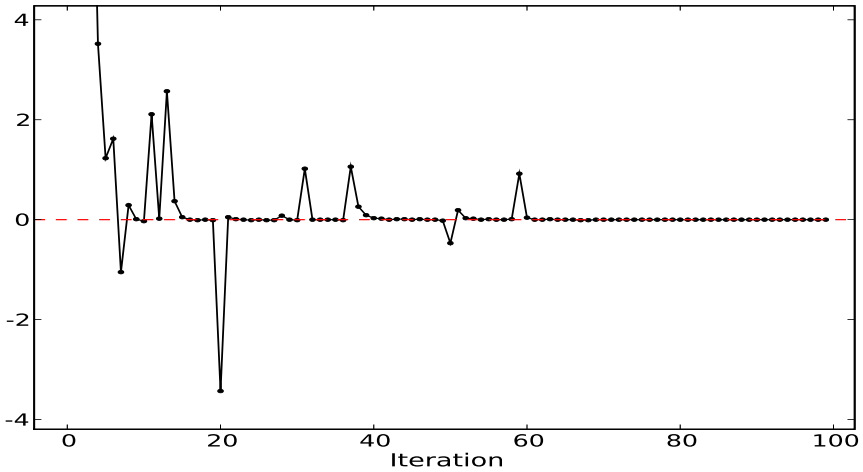

FIG. 5.11 – *Pins Laricio - environnement commun : valeurs de la différence des log-vraisemblances des données observées sachant les effets aléatoires entre deux itérations consécutives de l'algorithme MCEM avec une étape E de restauration probabiliste-simulation.*

Les paramètres du SMS-LMM ont été estimés par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (section 4.4.4) avec 1000 tirages par l'algorithme de Metropolis-Hastings à marche aléatoire à chaque itération. L'algorithme MCEM a été initialisé avec les paramètres β, σ^2, P, π et d estimés par l'algorithme EM pour semi-chaîne de Markov cachées gaussiennes sans prendre en compte d'effets aléatoires (i.e. $\lambda = 0$). Après vérification de la convergence des simulations des effets aléatoires, la convergence de l'algorithme est contrôlée par la différence des log-vraisemblances des données observées sachant les effets aléatoires entre deux itérations successives (équation (4.14)). Le tracé des valeurs de $g^{(k)}$ en fonction des itérations montre que l'algorithme MCEM converge en environ 60 itérations (figure 5.11).

Propriétés de la population

L'état 1 est le seul état initial possible (avec $\widehat{\pi}_1 = 1$) de la semi-chaîne de Markov sous-jacente estimée ; cf figure 5.12. La matrice des probabilités de transition est dégénérée ; pour chaque état transitoire i, $p_{i,i+1} = 1$ et $p_{ij} = 0$ pour $j \neq i+1$ (et pour l'état final absorbant $p_{ii} = 1$ et $p_{ij} = 0$ pour $j \neq i$). Par conséquent, la succession des états est déterministe pour la combinaison semi-markovienne de modèles linéaires mixtes de type "gauche-droite" dégénérée. Cette succession déterministe confirme l'hypothèse de succession de phases de croissance.

La distribution marginale du modèle linéaire mixte attaché à l'état j est la distribution gaussienne $\mathcal{N}(\mu_j, \gamma_j^2)$ où $\mu_j = \beta_{j1}$ et $\gamma_j^2 = \varsigma_j^2 + \tau_j^2$. Les distributions marginales pour les différents états sont bien séparées (peu de recouvrement entre les distributions marginales correspondantes à deux états successifs) ; comparer les différences des moyennes $\mu_{j+1} - \mu_j$ et les écart-types γ_{j+1} et γ_j entre états consécutifs dans le tableau 5.7. La prise en compte

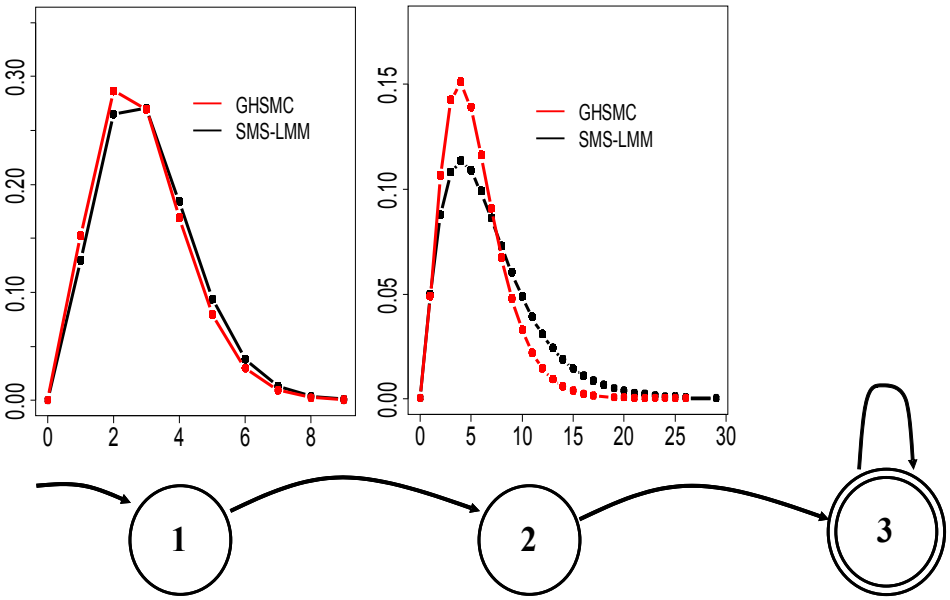

FIG. 5.12 – *Pins Laricio - environnement commun : semi-chaîne de Markov sous-jacente estimée. Chaque état est représenté par un sommet numéroté. Les sommets représentant les états transitoires sont entourés par une simple ligne tandis que le sommet représentant l'état absorbant est entouré par une double ligne. Les transitions possibles entre états sont représentées par des arcs (les probabilités qui sont égales à 1 ne sont pas notées). L'arc entrant dans l'état 1 indique l'état initial de probabilité égale à 1. Les lois d'occupation des états non absorbants sont figurées au dessus du sommet correspondant. Les lignes rouges correspondent aux lois d'occupation estimées pour une simple semi-chaîne de Markov cachée gaussienne (GHSMC) et les lignes noires correspondent aux lois d'occupation estimées pour une combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM).*

de l'environnement commun à tous les individus pour chaque année augmente le temps moyen passé dans l'état 2. Cette remarque se déduit de la comparaison des moyennes des loi d'occupation dans l'état 2 entre la semi-chaîne de Markov cachée gaussienne et la combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun (tableau 5.7 et figure 5.12). Cependant, nous pouvons remarquer que l'écart-type reste proportionnel à la moyenne dans chaque état par rapport au cas de la semi-chaîne de Markov cachée gaussienne.

La part d'environnement commun, définie par le ratio entre la variance aléatoire ς_j^2 et la variance totale γ_j^2 dans l'état j est plus forte au début de la vie de plante (deux premiers états avec plus de 39%) et diminue de quasiment 1/3 dans le dernier état (environ 27%). Ce phénomène de décroissance semble contradictoire avec les résultats où l'influence du climat modélisé comme un effet fixe augmentait le long de la vie de la plante (section 5.1.1.1). Nous justifierons par la suite les raisons de cette contradiction.

		état j		
		1	2	3
loi d'occupation moy., et.	GHSMC	P(1, 1.88) 2.88, 1.37	NB(1, 2.57, 0.31) 5.31, 2.93	
	SMS-LMM	P(1, 2.04) 3.04, 1.43	B(2, 4, 0.37) 6.64, 4.24	
paramètre de regression (SMS-LMM)	constante β_{j1}	8.75	26.44	51.98
décomposition de la variance (SMS-LMM)	variance aléatoire ς_j^2	6.19	45.23	39.08
	variance résiduelle σ_j^2	9.36	63.28	105.91
	variance totale γ_j^2	15.55	108.51	144.99
	part d'environnement commun	39.81%	41.68%	26.95%
distribution marginale μ_j, γ_j	GHSMC	6.97, 3.26	26.30, 9.12	54.35, 11.39
	SMS-LMM	8.75, 3.94	26.44, 10.42	51.98, 12.04

TAB. 5.7 – *Pins Laricio - environnement commun : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante et la décomposition de la variance sont donnés.*

Comportement individuel

La séquence d'états la plus probable sachant les simulations des effets aléatoires est calculée pour chaque séquence observée à partir d'une adaptation de l'algorithme de Viterbi pour semi-chaînes de Markov cachées (Guédon, 2003) aux combinaisons semi-markoviennes de modèles linéaires mixtes. Les lois marginales empiriques des états par année, extraites des séquences d'états les plus probables, sont représentées sur la figure 5.13. Comme pour le SMS-LMM avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle et la pluie cumulée comme effet fixe (figure 5.3), on retrouve un recouvrement des états sur les différentes années et quasiment la même structure sous forme de "vagues" pour l'état 1. Cependant, l'état 2 et l'état 3 sont moins distincts ; les effectifs de l'état 2 et de l'état 3 sont quasi égaux chaque année. La différence de répartition entre ces deux états est principalement liée à la loi d'occupation dans l'état 2 ; comparer les lois d'occupation entre SMS-LMM avec hétérogénéité inter-individuelle (tableau 5.1) et SMS-LMM avec environnement commun (tableau 5.7).

Les caractéristiques (médiane et dispersion) de la première année dans chaque état sont extraites pour les quatre sous-échantillons de pins Laricio sur la base des séquences d'états restaurées en utilisant soit la semi-chaîne de Markov cachée gaussienne estimée,

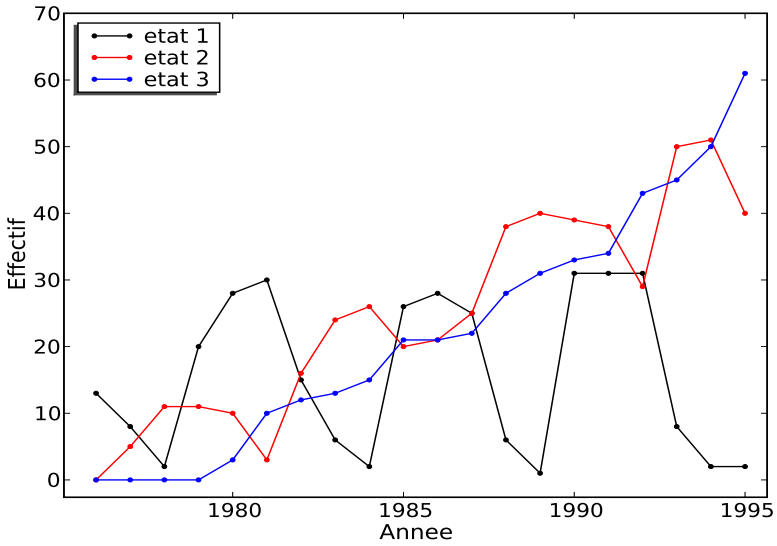

FIG. 5.13 – *Pins Laricio - environnement commun : lois marginales empiriques des états par année extraites des séquences d'états les plus probables.*

	âgés de 6 ans	âgés de 12 ans	âgés de 18 ans	âgés de 23 ans
état 2 GHSMC	1993 (0.56)	1988 (0.42)	1982 (1.93)	1978 (1.04)
état 2 SMS-LMM	1993 (0.41)	1988 (0.60)	1982 (1.63)	1978 (1.04)
état 3 GHSMC	1995 (0.35)	1993 (1.00)	1988 (2.56)	1981 (0.78)
état 3 SMS-LMM	1995 (0.30)	1992 (1.28)	1988 (3.06)	1981 (0.86)

TAB. 5.8 – *Pins Laricio - environnement commun : première année médiane dans l'état 2 et dans l'état 3 pour chaque sous-échantillon déduite de la semi-chaîne de Markov cachée gaussienne estimée (GHSMC) et de la combinaison semi-markovienne de modèles linéaires mixtes estimée (SMS-LMM). Les écart-types correspondants sont indiqués entre parenthèses.*

soit la combinaison semi-markovienne de modèles linéaires mixtes estimée. La première année médiane dans le deuxième état pour les quatre sous-échantillons est la même pour les deux modèles (tableau 5.8). La première année médiane dans le troisième état est la même pour les deux modèles dans le cas des pins Laricio âgés de 6 ans, de 18 ans et de 23 ans. Il y a un décalage de 1 an entre la première année médiane dans le troisième état pour les pins Laricio âgés de 12 ans. La dispersion de la première année dans le deuxième état est diminuée dans le cas du SMS-LMM pour les pins Laricio âgés de 6 ans et 18 ans et augmentée dans le cas du SMS-LMM pour les pins Laricio âgés de 12 ans. La dispersion de la première année dans le troisième état est fortement augmentée dans le cas de la combinaison semi-markovienne de modèles linéaires mixtes par rapport à la semi-chaîne de Markov cachée gaussienne excepté pour les pins âgés de 6 ans. Cette analyse

montre que la prise en compte de l'environnement commun augmente l'asynchronisme du changement d'état entre les individus (principalement entre le deuxième et le troisième état).

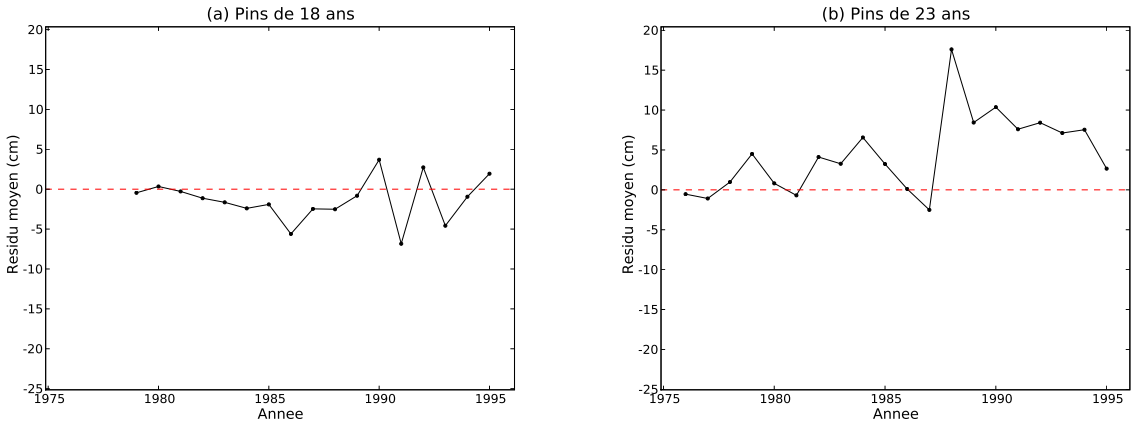

FIG. 5.14 – *Pins Laricio - environnement commun : moyennes annuelles calculées à partir des séquences résiduelles (a) pins âgés de 18 ans, (b) pins âgés de 23 ans.*

Pour chaque arbre, une séquence résiduelle a été calculée en soustrayant de la séquence originale, les lois empiriques des observations par année, déduites de la combinaison semi-markovienne de modèles linéaires mixtes estimée, des simulations des effets aléatoires et des séquences d'états les plus probables. Les moyennes annuelles sont calculées à partir des séquences résiduelles pour chaque sous-échantillon de pins Laricio (figure 5.14 pour les pins âgés de 18 ans et de 23 ans). Nous avons comparé ces moyennes résiduelles aux moyennes résiduelles obtenues pour chaque sous-échantillon dans le cas d'une semi-chaîne de Markov cachée gaussienne (figure 1.9). Les amplitudes entre les moyennes de deux années consécutives ont diminué identiquement pour les pins âgés de 18 ans et les pins de 23 ans entre le GHSMC et le SMS-LMM. Cependant, l'amplitude entre 1987 et 1988 étant beaucoup plus élevée pour les pins âgés de 23 ans que pour les pins âgés de 18 ans, elle n'a pas été entièrement compensée dans les simulations des effets aléatoires ; les simulations des effets aléatoires se basant uniquement sur les amplitudes annuelles communes à l'ensemble des pins Laricio. Les amplitudes entre 1990 et 1993 étant beaucoup plus fortes pour les pins âgés de 18 ans par rapport aux pins de 23 ans, l'introduction d'un effet aléatoire modélisant l'environnement commun ne permet d'évaluer que les amplitudes communes à tous les échantillons. Par conséquent, les amplitudes entre 1990 et 1993 ont diminué identiquement par rapport au GHSMC pour tous les échantillons mais pas suffisamment pour les pins âgés de 18 ans. Cet effet des sous-échantillons se retrouve dans les simulations des effets aléatoires (figure 5.15) et par conséquent dans la proportion d'environnement commun. Ceci peut expliquer la contradiction avec les résultats obtenus dans le cas d'une covariable climatique (section 5.1.1.1).

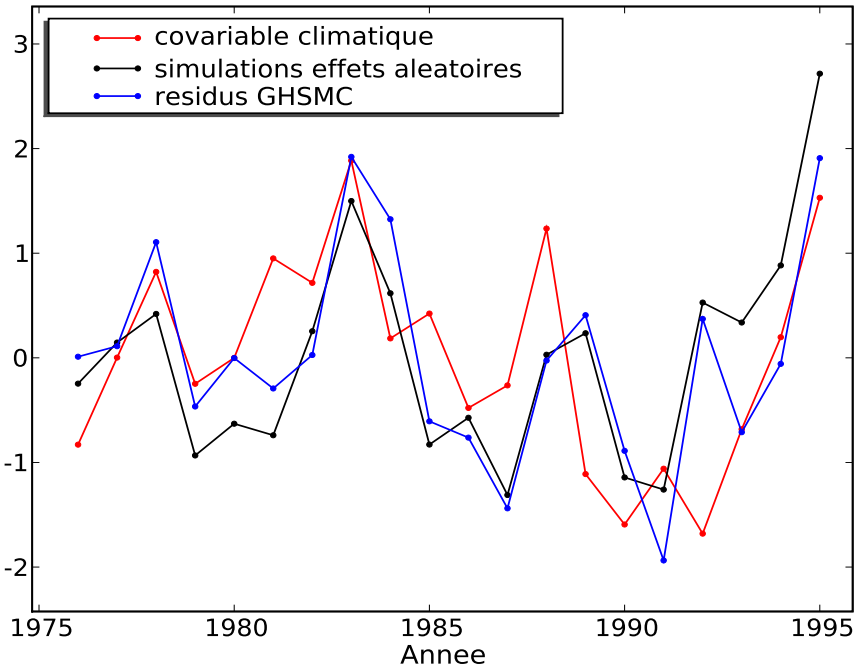

FIG. 5.15 – *Pins Laricio - environnement commun : comparaison des résidus moyens standardisés calculés à partir des séquences d'états restaurées pour la semi-chaîne de Markov gaussienne (GHSMC) avec les simulations standardisés des effets aléatoires modélisant l'environnement commun en fonction des années pour le SMS-LMM estimé avec l'algorithme MCEM avec une étape E de restauration probabiliste-simulation, et avec la covariable climatique standardisée utilisée dans la section 5.1.1.1.*

Etude des simulations des effets aléatoires

Nous avons calculé les moyennes résiduelles annuelles pour l'ensemble des pins Laricio (tout échantillon confondu) dans le cas de la semi-chaîne de Markov cachée gaussienne (en bleu sur la figure 5.15). Ces moyennes ont été comparées aux simulations des effets aléatoires (en noir sur la figure 5.15). Toutes les valeurs ont été centrées et réduites afin d'être sur une même échelle et donc de pouvoir les comparer. La différence observée les premières années peut être expliquée par le faible nombre de répétitions par année (seuls les 13 pins âgés de 23 ans sont mesurés) et par le peu de mélangeance des états (figure 5.13). La différence observée les dernières années est expliquée par la non-compensation des amplitudes synchrones pour tous les sous-échantillons (cf paragraphe précédent). Cette différence vient également du fait que l'estimation des 4 sous-échantillons conjointement n'a pas suffisamment fait diminuer le synchronisme entre individus du passage de la deuxième phase de croissance à la troisième ; comparer les lois marginales empiriques des états par année entre le SMS-LMM avec des effets aléatoires modélisant l'environnement commun (figure 5.13) et le SMS-LMM avec des effets aléatoires modélisant l'hétérogénéité inter-individuelle (figure 5.3).

Les simulations des effets aléatoires ont été comparées à la covariable climatique choisie

dans la section 5.1.1.1 : la pluie cumulée sur une période recouvrant une période d'organogénèse et d'allongement pour chaque pousse annuelle. Comme précédemment, toutes les valeurs ont été centrées et réduites afin d'être sur une même échelle et donc de pouvoir les comparer (figure 5.15). De nombreuses fluctuations synchrones entre individus sont liées à cette covariable climatique : une faible pluie cumulée inhibe la croissance (années 1986, 1990, 1991 par exemple) tandis qu'une forte pluie cumulée favorise la croissance (années 1978,1983 et 1995 par exemple). Cependant, à partir de 1984, les différences d'amplitudes entre les simulations et les valeurs de la covariable climatique semblent indiquer que d'autres facteurs influencent la vigueur commune de croissance des individus.

5.2.2 Chênes sessiles

Une combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus pour chaque âge chronologique (ou date) est estimée sur la base des longueurs de pousses annuelles des 2 sous-échantillons de chênes sessiles. Comme dans le cas d'effets aléatoires modélisant l'hétérogénéité inter-individuelle (section 5.1.2), la semi-chaîne de Markov sous-jacente est supposée de type "gauche-droite" à 2 états. Nous avons choisi un unique effet fixe (une constante) pour chaque modèle linéaire mixte qui s'écrit pour l'état j :

$$y_{ad}|_{S_{a,t_a(d)}=j} = \beta_{j1} + \varsigma_j \lambda_d + \epsilon_{ad}, \qquad \lambda_d \sim \mathcal{N}(0,1), \qquad \epsilon_{ad}|_{S_{a,t_a(d)}=j} \sim \mathcal{N}(0, \sigma_j^2),$$

où y_{ad} est la longueur de pousses annuelles pour l'individu a à l'âge chronologique (ou date) d et β_{j1} est la constante associée à l'état j.

Les paramètres du SMS-LMM ont été estimés par l'algorithme MCEM avec une étape E de restauration probabiliste-simulation (section 4.4.4) avec 1000 tirages par l'algorithme de Metropolis-Hastings à marche aléatoire à chaque itération. L'algorithme MCEM a été initialisé avec les paramètres β, σ^2, P, π et d estimés par l'algorithme EM pour semi-chaîne de Markov cachées gaussiennes sans prendre en compte les effets aléatoires (i.e. $\lambda = 0$). Après vérification de la convergence des simulations des effets aléatoires, l'algorithme converge en environ 20 itérations. La convergence est contrôlée par la différence des log-vraisemblances des données observées sachant les effets aléatoires entre deux itérations successives (équation (4.14)).

Propriétés de la population

Les distributions marginales des observations pour les différents états sont fortement séparées (peu de recouvrement entre les distributions marginales) ; comparer la différence des moyennes $\mu_{j+1} - \mu_j$ entre états consécutifs et les écart-types γ_j et γ_{j+1} dans le tableau 5.9. La prise en compte de l'environnement commun comme effet aléatoire a peu d'influence sur la semi-chaîne de Markov sous-jacente. En effet, les lois d'occupation et notamment leurs écart-types sont quasi-identiques entre la combinaison semi-markovienne

		état j	
		1	2
loi d'occupation moy., et.	GHSMC	B(1, 8, 0.52) 4.64, 1.32	
	SMS-LMM	B(1, 8, 0.51) 4.54, 1.32	
paramètre de regression (SMS-LMM)	constante β_{j1}	5.50	45.70
décomposition de la variance (SMS-LMM)	variance aléatoire ς_j^2	1.59	215.10
	variance résiduelle σ_j^2	12.83	367.77
	variance totale γ_j^2	14.42	582.87
	part d'environnement commun	10.61%	36.90%
distribution marginale μ_j, γ_j	GHSMC	6.35, 3.79	45.18, 24.18
	SMS-LMM	5.50, 3.80	45.70, 24.14

TAB. 5.9 – *Chênes sessiles - environnement commun : comparaison des paramètres de la semi-chaîne de Markov cachée gaussienne (GHSMC) avec les paramètres de la combinaison semi-markovienne de modèles linéaires mixtes (SMS-LMM) (lois d'occupation et distributions marginales des observations). Pour chaque modèle linéaire mixte, la constante et la décomposition de la variance sont donnés.*

de modèles linéaires mixtes et la semi-chaîne de Markov cachée gaussienne dans le tableau 5.4. Ceci montre que la modélisation conjointe de ces deux sous-échantillons de chênes sessiles fait disparaître les problèmes d'identifiabilité évoqué dans le chapitre précédent. Les simulations des effets aléatoires modélisant l'environnement commun confirment cette hypothèse. En effet, les simulations des effets aléatoires entre 1988 et 1989 (figure 5.16) diminuent alors que ces années correspondent aux années de changement d'état (changement qui correspond à une augmentation significative de la longueur de pousses annuelles moyenne).

Contrairement aux pins Laricio, la part d'environnement commun est faible au début de la vie de la plante (environ 11% sur le premier état) puis augmente fortement sur le deuxième état (environ 37%) (tableau 5.9). La forte augmentation de la part d'environnement commun au cours de la vie de l'arbre est en adéquation avec la forte augmentation de l'influence de la pluie cumulée dans le cas de la combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'hétérogénéité inter-individuelle et la pluie cumulée comme effet fixe (section 5.1.2).

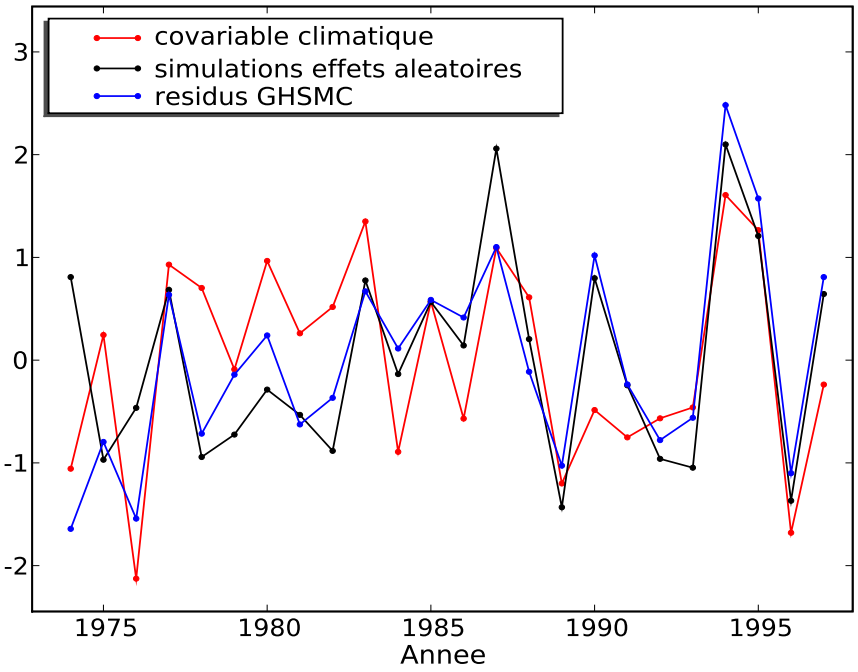

FIG. 5.16 – *Chênes sessiles - environnement commun : comparaison des résidus moyens standardisés calculés à partir des séquences d'états restaurées pour la semi-chaîne de Markov gaussienne (GHSMC) avec les simulations standardisés des effets aléatoires modélisant l'environnement commun en fonction des années pour le SMS-LMM estimé avec l'algorithme MCEM avec une étape E de restauration probabiliste-simulation, et avec la covariable climatique standardisée utilisée dans la section 5.1.2.*

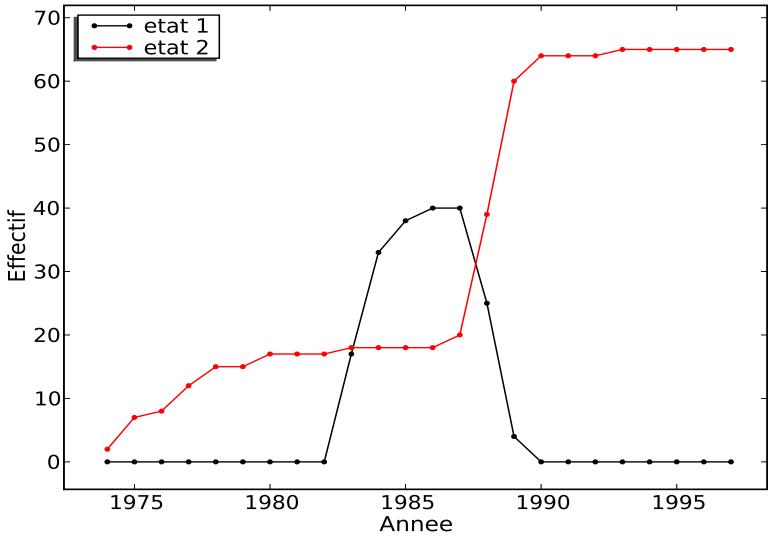

FIG. 5.17 – *Chênes sessiles - environnement commun : lois marginales empiriques des états par année extraites des séquences d'états les plus probables.*

Comportement individuel

Comme pour les pins Laricio, la séquence d'états la plus probable sachant les effets aléatoires est restaurée pour chaque individu à l'aide de l'algorithme de Viterbi pour SMS-LMM. Les lois marginales empiriques des états par année, extraites des séquences d'états les plus probables, sont représentées sur la figure 5.17. Ces effectifs sont identiques à ceux obtenus dans le cas de l'estimation d'une combinaison semi-markovienne de modèles linéaires mixtes avec un effet aléatoire modélisant l'hétérogénéité inter-individuelle et la pluie cumulée comme effet fixe (figure 5.7).

Etude des simulations des effets aléatoires

Pour chaque arbre, une séquence résiduelle a été calculée en soustrayant de la séquence originale, la fonction en escalier de la semi-chaîne de Markov cachée gaussienne. Les moyennes annuelles sont calculées à partir des séquences résiduelles de l'ensemble des chênes sessiles (en bleu sur la figure 5.16). Ces moyennes ont été comparées aux simulations des effets aléatoires (en noir sur la figure 5.16). Toutes les valeurs ont été centrées et réduites afin d'être sur une même échelle et donc de pouvoir les comparer. Seule la première année (année 1974) diffère fortement entre sa valeur simulée et sa valeur résiduelle. Cette différence s'explique par le faible nombre de répétitions (seulement 2 arbres ont une mesure en 1974) utilisés pour simuler les effets aléatoires (figure 5.17). La forte concordance entre les simulations des effets aléatoires et les résidus moyens du GHSMC montre que les effets aléatoires ne modélisent bien que l'environnement commun à tous les individus chaque année (et n'inclut pas le certain synchronisme entre individus du changement de phase de croissance).

Les simulations des effets aléatoires ont été comparées à la covariable climatique choisie dans la section 5.1.2 : la pluie cumulée sur une période recouvrant une période d'organogénèse et d'allongement pour chaque pousse annuelle. Comme précédemment, toutes les valeurs ont été centrées et réduites afin d'être sur une même échelle et donc de pouvoir les comparer (figure 5.16). Comme pour les pins Laricio, de nombreuses fluctuations synchrones entre individus sont liées à cette covariable climatique : une faible pluie cumulée inhibe la croissance (années 1989 et 1996 par exemple) tandis qu'une forte pluie cumulée favorise la croissance (années 1987, 1994 et 1995 par exemple). Cependant, les différences entre les simulations et les valeurs de la covariable climatique de 1978 à 1982 semblent indiquer que d'autres facteurs environnementaux influencent la croissance locale des individus. La différence d'amplitude entre les simulations et les valeurs de la covariable climatique de 1989 et 1990 n'est pas due à l'influence d'une autre facteur mais à une intervention sylvicole sur les chênes sessiles âgés de 15 ans en 1989 (première coupe secondaire (Heuret et al., 2000)).

5.3 Conclusion et discussion

Les combinaisons semi-markoviennes de modèles linéaires mixtes ont permis d'apporter une réponse à notre problématique biologique : identifier et caractériser les composantes de la croissance d'une plante. Les sorties des modèles ont souvent été justifiées par des explications biologiques fondées.

Nous avons dans un premier temps considéré des effets aléatoires modélisant l'hétérogénéité inter-individuelle. La composante ontogénique prend la forme d'une succession d'états où le statut d'un arbre par rapport à l'arbre "moyen" peut évoluer entre les états. Par exemple, un arbre peut croître plus lentement dans le premier état que l'arbre moyen puis plus rapidement dans le deuxième état. Dans le cas des chênes sessiles et des pins Laricio, la prise en compte de l'influence du climat (pluie cumulée) et de l'hétérogénéité inter-individuelle rend les changements de phase de croissance plus synchrones entre les arbres que pour une simple semi-chaîne de Markov cachée gaussienne. Les états identifiés par l'estimation d'une combinaison semi-markovienne de modèles linéaires mixtes ne sont pas seulement définis par la longueur de pousses annuelles moyenne, mais aussi par l'amplitude des fluctuations climatiques synchrones entre les individus et par la part d'hétérogénéité inter-individuelle.

Les poids des composantes environnementales et individuelles sont très différents pour les chênes sessiles et les pins Laricio :

- L'influence de la pluie cumulée est beaucoup plus faible dans le cas du pin Laricio que dans le cas du chêne sessile. La plus forte sensibilité aux facteurs climatiques des chênes sessiles peut être expliquée par leur plasticité en raison de leur capacité à produire plus d'une unité de croissance au cours d'une saison de croissance (Barthélémy et Caraglio, 2007) ; la production d'unités de croissance supplémentaires étant en partie contrôlée par les conditions environnementales et par la réactivité à des traumatismes.
- La part de l'hétérogénéité inter-individuelle est beaucoup plus forte pour les pins Laricio que pour les chênes sessiles.

L'influence de la covariable climatique (c'est-à-dire les pluies cumulées au cours d'une période recouvrant l'organogénèse et l'allongement) est faible dans le premier état (ce qui correspond au début de la vie de la plante), puis augmente nettement avec la croissance moyenne alors que la part d'hétérogénéité inter-individuelle diminue plus légèrement. La plus petite part d'hétérogénéité inter-individuelle pour les chênes sessiles par rapport aux pins Laricio peut être expliquée par l'origine des arbres (régénération naturelle pour les chênes sessiles au lieu de plants issus de pépinière et transplantés pour les pins Laricio), les interventions sylvicoles (éclaircie pour les chênes sessiles) et la stratégie d'échantillonnage (chênes sessiles sélectionnés parmi les arbres dominants ou codominants et les pins Laricio choisis afin de couvrir l'ensemble des classes de hauteur et de diamètre).

Dans le cas des pins sylvestres, l'année du changement d'état entre l'état de plus forte croissance et l'état suivant est clairement influencée par l'intervention sylvicole (éclaircie). Cette intervention sylvicole entraîne une augmentation marquée de la part d'hétérogénéité inter-individuelle. Il serait très utile d'étudier la part d'hétérogénéité inter-individuelle de diverses espèces dans des conditions similaires ou pour une espèce donnée dans des conditions variées telles que des densités différentes de plantation ou des interventions sylvicoles différentes.

Dans le cas des noyers, la pluie cumulée a peu d'influence sur les longueurs de pousses annuelles tandis que l'influence de la température maximale moyenne augmente significativement d'un état au suivant (double entre les deux derniers états). La prise en compte de la présence d'au moins une ramification au cours de la vie de la plante apporte près de 20% de longueur supplémentaire sur chaque état. Ce phénomène s'explique par une forte corrélation entre une pousse annuelle longue et la présence de ramification la même année.

La seconde partie de ce chapitre était consacrée à la modélisation des longueurs de pousses annuelles des pins Laricio et des chênes sessiles par des combinaisons semi-markoviennes de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus pour chaque date. Chez les pins Laricio, la prise en compte de l'environnement commun modifie la semi-chaîne de Markov sous-jacente et augmente l'asynchronisme des changements de phase de croissance entre les individus. *A contrario*, la prise en compte de l'environnement commun chez le chêne sessile n'entraîne aucune modification de la structure ontogénique sous-jacente.

Pour les pins Laricio, la part d'environnement commun est forte au début de la vie de la plante puis diminue dans le troisième état. Ce résultat semble en opposition avec les résultats obtenus sur l'influence de la pluie cumulée lorsque cette dernière est modélisée comme un effet fixe. La principale raison vient du fait que l'introduction d'un effet aléatoire modélisant l'environnement commun ne permet d'évaluer que les amplitudes communes à tous les sous-échantillons. Pour les chênes sessiles, les résultats obtenus vont dans le même sens que dans le cas de la modélisation avec la variable explicative pluie cumulée comme effet fixe. La part d'environnement commun augmente fortement entre le premier état et le deuxième état. Il serait très utile d'étudier la part d'environnement commun pour des jeux de données où le degré d'asynchronisme des changements de phases de croissance entre les individus est très élevé.

La comparaison des simulations des effets aléatoires et des covariables climatiques pour les pins Laricio et les chênes sessiles a montré que bien que les pluies cumulées soient fortement liées aux fluctuations synchrones entre individus, d'autres facteurs d'origine climatique ou sylvicole pouvaient jouer un rôle sur la croissance des arbres. Il pourrait

être intéressant d'introduire de nouvelles covariables variant dans le temps dans le processus d'observation des combinaisons semi-markoviennes de modèles linéaires mixtes. Nous avons vu que pour les noyers, la température maximale moyenne jouait un rôle important sur la croissance des arbres. Il serait utile de prendre en compte l'influence de la température maximale moyenne sur la croissance des pins Laricio, des chênes sessiles et des pins sylvestres. D'autres facteurs environnementaux influencent le développement des plantes tels que le temps thermique sur la biomasse[5] de petits pois (Lecoeur et Ney, 2003) ou encore la somme des températures ou la pluviométrie en juin sur la croissance des pousses et la phénologie[6] du pin pignon (Mutke et al., 2003).

[5]Masse dur pied d'un organisme à un temps donné.

[6]Étude de l'apparition d'événements périodiques dans le monde vivant, déterminée par les variations saisonnières du climat. En botanique, ces événements sont par exemple la floraison, la feuillaison, la fructification et la coloration des feuilles des végétaux.

Chapitre 6

Conclusions et perspectives

Nous avons étudié dans ce travail de thèse une famille de modèles statistiques : les combinaisons markoviennes et semi-markoviennes de modèles de régression. Nous avons évalué les performances des modèles et méthodes d'estimation proposés par le biais d'études de simulations et d'analyses de données réelles. Ce travail était motivé par une problématique biologique bien définie concernant la croissance d'arbres forestiers. La modélisation des composantes de la croissance d'arbres forestiers nous a permis de quantifier le rôle joué par chacune de ces composantes au cours de la croissance. Dans ce chapitre, nous allons revenir sur les points forts des approches proposées mais aussi sur leurs limites et sur les développements futurs.

6.1 Conclusions

6.1.1 Au niveau statistique

L'objectif statistique de cette thèse était de modéliser des données de type séquence ou série chronologique. Ces séquences observées étaient structurées en phases successives asynchrones entre individus, étaient influencées dans chaque phase par des covariables pouvant varier dans le temps et pouvaient présenter une hétérogénéité inter-individuelle. Nous avons alors étudié un certain nombre de combinaisons markoviennes et semi-markoviennes de modèles de regression. Ces modèles peuvent être vu comme des modèles de mélange fini de modèles de regression avec des dépendances markoviennes ou semi-markoviennes.

Le schéma 6.1 résume les différents modèles présentés ainsi que leur lien et leur origine :

- les combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés (MS-GLM et SMS-GLM) présentées dans le chapitre 3,
- les combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes (MS-LMM et SMS-LMM) présentées dans le chapitre 4.

Les combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés mixtes (MS-GLMM et SMS-GLMM) n'ont pas été traitées dans le cadre de cette thèse. Nous en discuterons ultérieurement dans les perspectives.

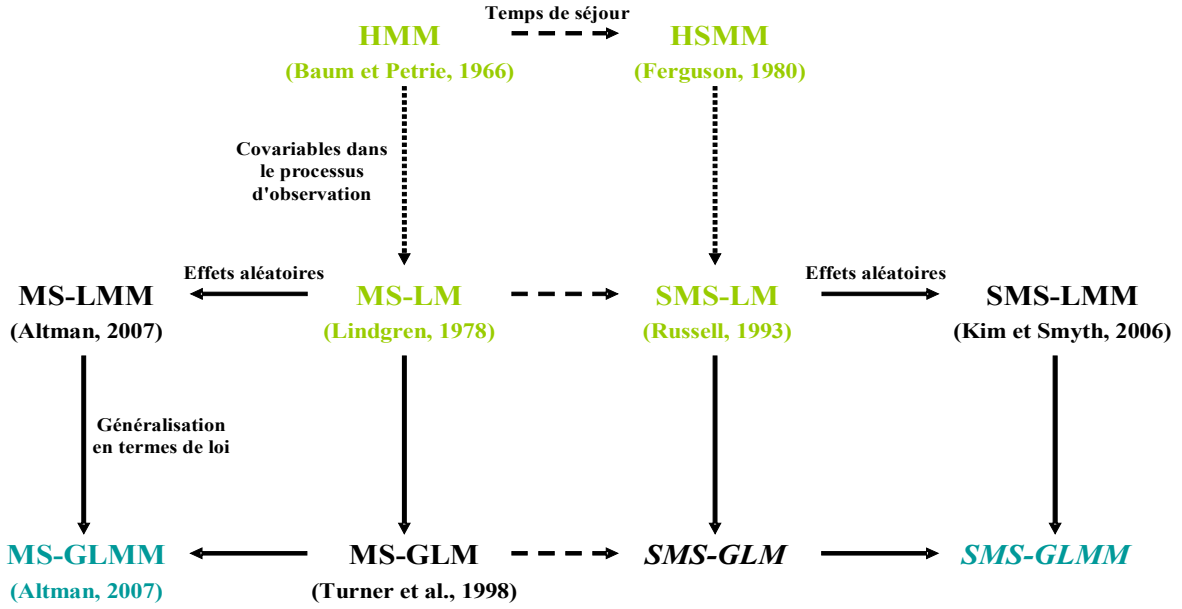

FIG. 6.1 – *Schéma des liens entre les différents modèles étudiés (en noir) et les modèles connus (en vert). Les modèles en bleus (MS-GLMM et SMS-GLMM) n'ont pas été étudié dans le cadre de cette thèse. L'article d'origine des différents modèles est référencé entre parenthèses.*

Ces modèles sont des modèles de type markovien ou semi-markovien caché où le processus d'observation est lié au processus d'état par un modèle de regression. Les états du processus markovien ou semi-markovien sous-jacent étant non-observables, l'estimation des paramètres de ces combinaisons pouvant être vue comme la résolution d'un problème aux données incomplètes, les méthodes d'estimation proposées sont basées sur le principe :

- de l'algorithme du gradient EM pour les combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés (chapitre 3),
- de l'algorithme MCEM pour les combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes (chapitre 4).

Le chapitre 3 est consacré aux combinaisons (semi-)markoviennes de modèles linéaires généralisés où un modèle linéaire généralisé est associé à chaque état. Ces modèles prennent seulement l'effet de covariables comme des effets fixes dans le processus d'observation. Nous avons proposé des algorithmes de type gradient EM ou MCEM pour estimer les paramètres des MS-GLM et des SMS-GLM. L'étape E consiste soit en une restauration probabiliste de toutes les séquences d'états possibles par l'algorithme "avant-arrière", soit

en une restauration par simulation de séquences d'états par l'algorithme "avant-arrière" de simulation. L'étape M repose sur une maximisation directe de la quantité calculée à l'étape E pour les paramètres du processus d'état sous-jacent et sur les équations des scores de Fisher pour les paramètres des modèles linéaires généralisés. Ces équations sont similaires aux équations pour GLM classiques hormis le fait que les matrices hessiennes sont pondérées soit par les probabilités lissées, soit par les comptages extraits des séquences d'états simulées. Il pourrait être intéressant d'envisager d'autres approches numériques telles que les méthodes de quasi-Newton pour estimer les paramètres des GLM associés à chaque état du processus d'état sous-jacent. Les approches proposées ont l'avantage d'être assez rapides, peu sensibles aux valeurs initiales et de respecter la propriété d'accroissement monotone de l'algorithme EM. Cependant, dans le cas des données binaires, les simulations ont montré que l'estimation des constantes des GLM n'était pas suffisamment robuste.

Les combinaisons (semi-)markoviennes de modèles linéaires mixtes où le processus d'observation est lié au processus d'état par des modèles linéaires mixtes ont été étudiées dans le chapitre 4. Nous avons distingué deux types d'effets aléatoires possibles dans le processus d'observation : des effets aléatoires "individuels" et des effets aléatoires "temporels". Quel que soit le type d'effet aléatoire, nous avons proposé d'estimer les paramètres des MS-LMM et des SMS-LMM par des algorithmes de type MCEM où l'étape E consiste en deux étapes conditionnelles : une restauration des séquences d'états sachant les effets aléatoires et une restauration des effets aléatoires sachant les séquences d'états. Les méthodes étudiées dans le chapitre 4 sont résumées dans le tableau 4.6. Nous avons proposé deux approches :

- Une première approche consiste en une restauration par simulation des séquences d'états sachant les effets aléatoires, et soit en une restauration déterministe soit en une restauration par simulation des effets aléatoires sachant les séquences d'états. L'étape de restauration conditionnelle des effets aléatoires nécessite de calculer explicitement la loi des effets aléatoires sachant les données observées et les séquences d'états. C'est pourquoi, cette approche est particulièrement adaptée au cas d'effets aléatoires "individuels" mais ne peut pas se transposer aisément au cas d'effets aléatoires "temporels".
- Une seconde approche consiste en une restauration probabiliste des séquences d'états sachant les effets aléatoires et en une restauration par simulation des effets aléatoires sachant les séquences d'états. L'étape de restauration conditionnelle des effets aléatoires repose sur des simulations par l'algorithme de Metropolis-Hastings indépendant ou à marche aléatoire selon le type d'effet aléatoire. Du fait de ne pas devoir connaître explicitement la loi des effets aléatoires sachant les séquences d'états, cette approche peut être utilisée pour tout type d'effet aléatoire. Cependant, l'approche avec une étape de Metropolis-Hastings repose pour chaque individu sur le calcul du ratio entre la log-vraisemblance des données observées sachant les effets aléatoires et la log-vraisemblance des données observées sachant les valeurs proposées des ef-

fets aléatoires à chaque tirage. Le calcul de la seconde quantité nécessite de devoir calculer les constantes de normalisation lors de la récurrence "avant" sachant les valeurs proposées des effets aléatoires. Par conséquent, cette approche est beaucoup plus lente que les autres approches. De plus, des problèmes de mélangeance peuvent être rencontrés. Une solution serait de simuler les effets aléatoires par un algorithme de Metropolis-Hastings par blocs (Robert et Casella, 2004).

Dans le cas d'effets aléatoires "individuels", les simulations ont montré la robustesse de l'algorithme MCEM avec une étape E de simulation-prédiction pour différents degrés de séparabilité des états. Nous avons également constaté que cet algorithme conduit aux mêmes résultats que l'algorithme MCEM avec une étape E de restauration probabiliste-simulation. La préférence pour le premier vient principalement du fait de sa plus grande rapidité et de son plus faible coût de calcul.

Dans le cas d'effets aléatoires "temporels", nous avons noté un bon comportement de l'algorithme MCEM avec une étape E de restauration probabiliste-simulation avec cependant, une forte influence du degré de séparabilité des états. L'application sur les longueurs de pousses annuelles des 30 pins Laricio âgés de 18 ans a mis en évidence des problèmes d'identifiabilité. En effet, si le degré de synchronisme des changements d'état est trop fort entre individus, ce synchronisme est "épongé" par les effets aléatoires. La solution proposée pour pallier ce problème a été de combiner plusieurs jeux de données d'une même espèce à des âges ontogéniques différents afin de mélanger les états pour chaque année.

6.1.2 Au niveau informatique

Les modèles et méthodes d'estimation proposés ont été implémentés en C++. On s'est appuyé sur le code C++ de Y. Guédon pour les semi-chaînes de Markov cachées classiques, et sur la librairie numérique GSL[1] (Galassi et al., 2006) pour les calculs d'algèbre linéaire. Les combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés ont été implémentées hormis le cas des données de type catégorielles ou ordinales. L'algorithme MCEM avec une étape E de simulation-simulation pour les combinaisons (semi-)markoviennes de modèle linéaires mixtes (section 4.3.5) n'a pas été implémenté. Il est envisagé de le faire assez rapidement. Bien que les modèles et les méthodes d'estimation soient implémentés en C++, nous avons utilisé le logiciel R pour une grande partie des analyses *a posteriori* des résultats obtenus et la librairie matplotlib[2] de Python (Martelli, 2004) pour les sorties graphiques.

[1] http ://www.gnu.org/software/gsl/
[2] http ://matplotlib.sourceforge.net/

Le code C++ résultant de ce travail de thèse sera intégré dans le logiciel VPlants[3] dans le cadre du projet Virtual Plants[4], dédié à la modélisation et à l'analyse du développement des plantes à différentes échelles. Ce logiciel fait partie de la plateforme logiciel OpenAlea[5]. L'intégration de notre code dans le logiciel VPlants nécessite d'écrire une interface en Python. Grâce au module RPy[6], notre code sera alors directement interfacé en R.

6.1.3 Au niveau biologique

L'objectif applicatif de cette thèse était d'identifier et de caractériser les trois principales composantes de la croissance d'arbres forestiers : la composante ontogénique, la composante environnementale et la composante individuelle, sur la base des hypothèses biologiques présentées au chapitre 1. La modélisation des composantes de la croissance d'arbres forestiers par des combinaisons semi-markoviennes de modèles de régression nous a permis de caractériser le rôle joué par chacune de ces composantes au cours de la croissance. La composante ontogénique, caractérisée par une succession de phases de croissance, a été modélisée par une semi-chaîne de Markov de type "gauche-droite". Un modèle de régression a été associé à chaque état et modélisait dans chaque état, l'influence de covariables par des effets fixes et soit l'hétérogénéité inter-individuelle, soit l'environnement commun par des effets aléatoires (hormis pour les SMS-GLM où il n'y avait pas d'effets aléatoires dans la modélisation).

Les combinaisons semi-markoviennes de modèle linéaires généralisés (chapitre 3) ont été peu utilisées pour modéliser les données de croissance d'arbres forestiers. Un unique exemple a été traité. Il concernait le nombre d'unités de croissance par pousse annuelle chez les pins Laricio de 18 ans. Nous avons cependant pu mettre en avant que le nombre d'unités de croissance par pousse annuelle était influencé par les conditions climatiques et notamment la pluviométrie. Cependant, il serait intéressant de traiter le cas de variables catégorielles (rameau/fleur/feuille) ou ordinales (1, 2 ou plus de 3 unités de croissance) à l'échelle de la pousse annuelle.

Le chapitre 5 était consacré à la modélisation de la croissance d'arbres forestiers par des combinaisons semi-markoviennes de modèles linéaires mixtes. Le post-traitement des résultats nous a permis, en collaboration avec des biologistes, de justifier nos résultats par des explications biologiques fondées. L'étude de plusieurs espèces dans des contextes différents a permis de mettre en évidence les similitudes entre espèces mais également les différences en terme de poids de chaque composante. La distinction et la caractérisation des trois principales composantes de la croissance est un résultat nouveau au sein de la communauté botanique.

[3] http ://www-sop.inria.fr/virtualplants/wiki/doku.php ?id=software
[4] http ://www-sop.inria.fr/virtualplants/wiki/doku.php ?id=home
[5] http ://openalea.gforge.inria.fr/dokuwiki/doku.php
[6] http ://rpy.sourceforge.net/

Nous avons dans un premier temps considéré des effets aléatoires modélisant l'hétérogénéité inter-individuelle. Nous avons montré que la composante ontogénique prenait la forme d'une succession de phases de croissance où le statut d'un arbre par rapport à l'arbre "moyen" pouvait évoluer entre les phases. Les phases identifiées par l'estimation d'une combinaison semi-markovienne de modèles linéaires mixtes ne sont pas seulement définies par la longueur de pousses annuelles moyenne, mais aussi par l'amplitude des fluctuations climatiques synchrones entre les individus et par la part d'hétérogénéité inter-individuelle. La prise en compte de l'influence du climat (composante environnementale) et de l'hétérogénéité inter-individuelle (composante individuelle) rend les changements de phase de croissance plus synchrones entre les arbres que pour une simple semi-chaîne de Markov cachée gaussienne.

L'influence du climat est faible dans la première phase de croissance (début de la vie de la plante), puis augmente nettement avec la croissance moyenne alors que la part d'hétérogénéité inter-individuelle diminue plus légèrement. Les différences de part d'hétérogénéité inter-individuelle entre espèces peuvent être expliquées par l'origine des arbres (régénération naturelle ou plants issus de pépinière et transplantés), les interventions sylvicoles (éclaircie ou coupe secondaire) et la stratégie d'échantillonnage. Il serait très utile d'étudier la part d'hétérogénéité inter-individuelle de diverses espèces dans des conditions similaires ou pour une espèce donnée dans des conditions variées telles que des densités différentes de plantation, des lieux différents ou des interventions sylvicoles différentes.

Nous avons dans un second temps modélisé les longueurs de pousses annuelles par des combinaisons semi-markoviennes de modèles linéaires mixtes avec un effet aléatoire modélisant l'environnement commun à tous les individus à chaque date. Des résultats très contrastés ont été obtenus selon l'espèce étudiée. Pour les pins Laricio, la part d'environnement commun est forte au début de la vie de la plante puis diminue dans la troisième phase de croissance (croissance maximum). Ce résultat est en contradiction avec les résultats obtenus sur l'influence de la pluie cumulée lorsque cette dernière est modélisée comme un effet fixe. La principale raison vient du fait que les effets aléatoires "temporels" épongent les amplitudes communes à tous les sous-échantillons mais n'épongent pas suffisamment les amplitudes synchrones au sein de chaque sous-échantillon. Pour les chênes sessiles, les résultats obtenus sont cohérents avec ceux de la modélisation avec la variable explicative pluie cumulée comme effet fixe. Il serait très utile d'étudier la part d'environnement commun pour des jeux de données où le degré d'asynchronisme des changements de phases de croissance entre les individus est très élevé.

Ce type de données trouvent, dans notre cas, son intérêt dans la modélisation de la croissance d'arbres. Cependant, il existe de nombreux autres domaines dans lesquels on peut rencontrer des données de type séquence ou série chronologique de même nature : en biologie sur les réseaux de gènes (Gupta et al., 2007) ou sur les coliformes fécaux dans l'eau de mer (Turner et al., 1998), en médecine sur les tumeurs du cerveau (Rijmen et al.,

2008), sur les données d'IRM de patients atteints de sclérose en plaques (Altman et Petkau (2005) et chapitre 3) ou sur les crises d'épilepsie (Wang et Puterman, 1999), en économie sur les courbes de rendement (Chopin et Pelgrin, 2004) ou encore en traitement du signal sur les formes d'ondes (Kim et Smyth, 2006). Il serait donc intéressant, comme cela a été fait dans le chapitre 3, d'appliquer les modèles proposés à d'autres domaines.

6.2 Perspectives

6.2.1 Prise en compte d'un effet aléatoire "groupe"

Nous avons noté dans l'exemple de la section 5.1.1.1 que le modèle linéaire mixte associé au troisième état sous-estimait la longueur de pousses annuelles moyenne dans la phase de croissance correspondante pour les pins Laricio âgés de 23 ans. Ce comportement a mis en évidence un effet "sous-échantillon". Pour prendre en compte cet effet, une variante possible des combinaisons semi-markoviennes de modèles linéaires mixtes serait d'ajouter un effet aléatoire "groupe" aux effets fixes dans chaque état. La prise en compte d'un effet aléatoire "groupe" peut être utile dans différentes situations. En biologie, la différence entre groupes (c'est-à-dire l'hétérogénéité inter-groupe) peut être due à une origine génétique (Segura et al., 2008), à une densité de plantation (Uzoh et Oliver, 2006) ou à des propriétés du sol (Meng et al., 2007).

Deux hypothèses pourraient être envisagées concernant la modélisation de l'hétérogénéité inter-groupe :

- un unique effet aléatoire pour toute la séquence observée. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \alpha_j\zeta_k + \epsilon_{at},$$

$$\zeta_k \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

- un effet aléatoire différent pour chaque état. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \alpha_j\zeta_{kj} + \epsilon_{at},$$

$$\zeta_{kj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

où ζ_k est l'effet aléatoire groupe, α_j est l'écart-type de l'effet aléatoire groupe dans l'état j et ζ_{kj} est l'effet aléatoire groupe associé à l'état j.

Comme dans le cas d'effets aléatoires individuels ou d'effets aléatoire temporels, nous serions en présence de deux structures cachées : les états du processus markovien sous-jacent et les effets aléatoires groupes. De plus, les individus d'un même groupe ne seraient

indépendants que sachant les effets aléatoires. L'algorithme d'inférence consisterait alors en un algorithme MCEM avec deux étapes de restauration conditionnelle dans l'étape E : une restauration des séquences d'états sachant les effets aléatoires et une restauration des effets aléatoires sachant les séquences d'états. L'algorithme MCEM avec une étape E de simulation-prédiction ou de simulation-simulation ne semble pas adapté pour l'estimation des paramètres. En effet, un effet aléatoire n'étant pas associé à chaque individu mais à chaque groupe d'individus, la restauration par prédiction des effets aléatoires reposerait sur la combinaison des séquences d'états simulées pour les individus d'un même groupe. Il est important de noter que le nombre de combinaisons possibles augmente avec le nombre de séquences d'états simulées par individu et avec le nombre d'individus constituant chaque groupe. Une solution alternative *a priori* plus adaptée serait d'utiliser l'algorithme MCEM avec une étape E de restauration probabiliste-simulation. Mais, quel algorithme de Metropolis-Hastings serait le plus adéquat pour simuler les effets aléatoires "groupes" dans l'algorithme MCEM pour les MS-LMM et les SMS-LMM ?

6.2.2 Plusieurs types d'effets aléatoires

Dans chaque modèle proposé, nous faisons l'hypothèse d'un unique type d'effets aléatoires dans les processus d'observation. Il serait intéressant de combiner de manière additive les effets aléatoires "individuels" avec les effets aléatoires "groupes" :

- un unique effet aléatoire individuel et un unique effet aléatoire groupe pour toute la séquence observée. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \tau_j\xi_a + \alpha_j\zeta_k + \epsilon_{at},$$

$$\xi_a \sim \mathcal{N}(0,1), \qquad \zeta_k \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

- un effet aléatoire individuel différent pour chaque état et un unique effet aléatoire groupe pour toute la séquence observée. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \tau_j\xi_{aj} + \alpha_j\zeta_k + \epsilon_{at},$$

$$\xi_{aj} \sim \mathcal{N}(0,1), \qquad \zeta_k \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

- un unique effet aléatoire individuel pour toute la séquence observée et un effet aléatoire groupe différent pour chaque état. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \tau_j\xi_a + \alpha_j\zeta_{kj} + \epsilon_{at},$$

$$\xi_a \sim \mathcal{N}(0,1), \qquad \zeta_{kj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

- un effet aléatoire individuel et un effet aléatoire groupe différents pour chaque état. Conditionnellement à l'état $S_{at} = j$, l'observation y_{at} d'un individu a du groupe k au temps t est modélisée par le modèle linéaire mixte suivant :

$$Y_{at}|_{S_{at}=j} = X_{at}\beta_j + \tau_j\xi_{aj} + \alpha_j\zeta_{kj} + \epsilon_{at},$$

$$\xi_{aj} \sim \mathcal{N}(0,1), \qquad \zeta_{kj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2).$$

Cette modélisation trouve par exemple son intérêt pour l'étude conjointe de la croissance de plusieurs espèces d'arbres forestiers. Nous pourrions étudier à la fois la composante individuelle au travers des effets aléatoires individuels et les différences de vigueur de croissance entre espèces au travers des effets aléatoires groupes. Cependant, la première question qui se pose est : quelle(s) modélisation(s) est la plus adéquate pour modéliser ces deux caractéristiques ?

Nous serions alors en présence de trois structures cachées : les états du processus sous-jacent, les effets aléatoires individuels et les effets aléatoires groupes. Si nous suivons le même raisonnement que dans le cas de deux structures cachées (sections 4.3.3 et 4.4.2), il serait nécessaire d'envisager, un algorithme MCEM avec trois restaurations conditionnelles dans l'étape E pour estimer les paramètres des (S)MS-LMM. L'étape de restauration conditionnelle des séquences d'états sachant les effets aléatoires individuels et les effets aléatoires groupes consisterait soit en une restauration par simulation par un algorithme "avant-arrière" de simulation, soit en une restauration probabiliste par un algorithme "avant-arrière". L'étape de restauration conditionnelle des effets aléatoires individuels consisterait soit en une restauration par simulation par un algorithme MCMC, soit en une restauration par prédiction. Nous avons vu dans la section précédente, que les effets aléatoires groupes ne pouvaient être prédits mais devaient être simulés par un algorithme de Metropolis-Hastings. Comme les deux types d'effet aléatoire sont indépendants, l'algorithme de Metropolis-Hastings se transposerait facilement pour restaurer par simulation les effets aléatoires groupes sachant les séquences d'états et les effets aléatoires individuels. L'enchaînement des étapes de restauration conditionnelles soulèvent de nombreuses questions parmi lesquelles : Y-a-il des combinaisons de restauration meilleures que d'autres ? Comment initialiser les algorithmes ? Comment contrôler leur convergence ?

6.2.3 Extension aux arbres de Markov cachés

Ce travail de thèse était centré sur deux processus d'état sous-jacents : les chaînes de Markov et les semi-chaînes de Markov. Nous avons vu que, comme les effets fixes et les effets aléatoires étaient intégrés dans le processus d'observation, les relations d'indépendance conditionnelle entre les processus d'état et les processus d'observation n'étaient pas modifiées. Les algorithmes MCEM proposés pouvaient alors être directement transposés du cas markovien au cas semi-markovien.

Les modèles proposés nous ont permis de modéliser des zones homogènes (phases de croissance) le long du tronc d'arbres forestiers. Cependant, ils ne permettent pas de caractériser la structure globale d'un arbre, formée d'une répétition d'entités (troncs, branches, rameaux) car les dépendances entre ces entités ne sont pas modélisées. Durand et al. (2005) ont proposé d'utiliser le cadre statistique des arbres de Markov cachés, introduit par Crouse et al. (1998) en traitement du signal, pour modéliser efficacement des zones homogènes à l'intérieur de la structure globale d'un arbre. Ces modèles sont basés sur des états cachés qui correspondent à des zones homogènes et qui sont obtenus en définissant des dépendances locales entre les états attachés à des entités adjacentes (rameau avec sa branche porteuse par exemple). L'intérêt biologique serait de combiner les arbres de Markov cachés avec les modèles de regression. L'écriture de tels modèles ne poserait pas de problèmes particuliers. Les relations d'indépendances conditionnelles entre les processus d'état et les processus d'observation n'étant pas modifiées, les algorithmes MCEM proposés pour les combinaisons (semi-)markoviennes de modèles de régression pourraient facilement s'adapter aux combinaisons de modèles de regression par arbre de Markov. La seule modification concernerait l'étape de restauration conditionnelle des séquences d'états sachant les effets aléatoires. L'algorithme "avant-arrière" serait remplacée par l'algorithme " ascendant-descendant" propre aux arbres de Markov cachés (Crouse et al., 1998; Durand et al., 2004).

6.2.4 Combinaisons markoviennes et semi-markoviennes de modèles linéaires généralisés mixtes

Dans les combinaisons (semi-)markoviennes de modèles linéaires mixtes (chapitre 4), la variable réponse (longueur de pousses annuelles par exemple) est supposée suivre un mélange de lois gaussiennes avec des dépendances (semi-)markoviennes. Dans les combinaisons (semi-)markoviennes de modèles linéaires généralisés (chapitre 3), la variable réponse (1 ou 2 unités de croissance, nombre de lésions dans le cerveau par exemple) est supposée suivre un mélange de lois appartenant à la famille exponentielle. Bien que les MS-GLM et les SMS-GLM généralisent en termes de lois les MS-LM et les SMS-LM, ils ne prennent pas en compte l'hétérogénéité inter-individuelle dans les processus d'observation (figure 6.1). Altman (2007) a introduit les combinaisons markoviennes de modèles linéaires généralisés mixtes (MS-GLMM) où un modèle linéaire généralisé mixte (McCulloch et al., 2008) est associée à chaque état du processus markovien sous-jacent. Cependant, les approches d'estimation proposées par Altman (2007) et Rijmen et al. (2008) ne se transposent pas au cas semi-markovien (section 4.2). Il serait intéressant de développer les combinaisons semi-markoviennes de modèles linéaires généralisés mixtes (SMS-GLMM) et de proposer des approches d'estimation pouvant être facilement appliquées quel que soit le processus markovien sous-jacent. Cette famille de combinaisons semi-markoviennes trouverait son

intérêt applicatif dans l'analyse des structures de ramification d'arbres forestiers (Guédon et al., 2001) et en serait un prolongement naturel.

Les algorithmes MCEM pour estimer les (S)MS-GLMM combineraient l'étape E des algorithmes MCEM pour (S)MS-LMM et l'étape M des algorithmes du gradient EM pour (S)MS-GLM :

- **Étape E :**
  - une restauration conditionnelle des séquences d'états sachant les effets aléatoires,
  - une restauration conditionnelle des effets aléatoires sachant les séquences d'états.
- **Étape M :**
  - une ré-estimation des paramètres de la (semi-)chaîne de Markov sous jacente par maximisation directe,
  - une ré-estimation itérative des paramètres des modèles linéaires généralisés mixtes par un algorithme des scores de Fisher.

La restauration des séquences d'états sachant les effets aléatoires pourrait aussi bien se faire de manière probabiliste que par simulation. La loi des effets aléatoires sachant les séquences d'états et les données observées ne pouvant pas être calculée analytiquement, l'étape de restauration des effets aléatoires sachant les séquences d'états ne pourrait pas se faire par prédiction. L'alternative serait une étape de restauration conditionnelle par simulation par un algorithme de Metropolis-Hastings ; voir notamment Martinez (2006) dans le cadre des mélanges de modèles linéaires généralisés mixtes. Cependant, dans le cas des données binaires, l'introduction d'effets aléatoires va-t-il accentuer le manque de robustesse de l'estimation des constantes dans chaque état (section 3.4.2) ?

6.2.5 Critère de sélection de modèles

Dans nos applications (chapitre 5), le choix du nombre d'états s'est appuyé sur les travaux de Guédon et al. (2007). Cependant, il serait nécessaire de définir un critère pour choisir le nombre d'états des combinaisons markoviennes et semi-markoviennes de modèles de régression. Dans les applications sur données de croissance d'arbres forestiers, les processus markoviens sous-jacents sont supposées de type "gauche-droite". Cette hypothèse entraîne la non-ergodicité des modèles proposés. Les critères classiques de sélection de modèles (AIC, BIC) ne peuvent pas être utilisés pour choisir le nombre d'états et le meilleur modèle. Le critère de Chopin et Pelgrin (2004) pour les combinaisons markoviennes de modèles linéaires ne convient pas car ils se placent dans un cadre ergodique. Aussi, les critères proposés par Lavielle (2005), Lebarbier (2005) et Zhang et Siegmund (2007) pour déterminer le nombre optimal de ruptures en détection de ruptures peuvent-ils être transposés aux (S)MS-LMM de type "gauche-droite" ? De plus, comment gérer le fait que nous ne connaissons que la log-vraisemblance des données observées sachant les effets aléatoires ?

Les résultats biologiques obtenus ont montré que l'influence du climat était faible au début de la plante et était plus forte sur les dernières phases de croissance. Il serait intéressant de définir un critère pour déterminer la significativité des covariables dans chaque état. Ce critère pourrait-il s'appuyer sur l'adaptation de la méthode de Louis utilisée pour calculer la matrice d'information des observations dans le cadre de l'algorithme MCEM (McLachlan et Krishnan, 2008, chap. 6) ?

Bibliographie

A. Agresti. *Categorical data analysis. 2nd Edition.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Chichester : Wiley, 2002. (Cité pages 58 et 60.)

P.S. Albert. A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 47:1371–1381, 1991. (Cité pages 38, 43 et 84.)

R.M. Altman. *Hidden Markov Models : Multiple Processes and Model Selection.* Thèse, 115 p., 2003. (Cité pages 45, 85, 86 et 87.)

R.M. Altman. Mixed hidden Markov models : An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statististical Association*, 102: 201–210, 2007. (Cité pages 4, 6, 7, 84, 96, 98 et 184.)

R.M. Altman et A.J. Petkau. Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, 24:2335–2344, 2005. (Cité pages 27, 84 et 181.)

G.E.B. Archer et D.M. Titterington. Parameter estimation for hidden Markov chains. *Journal of Statistical Planning and Inference*, 108:365–390, 2002. (Cité pages 46, 54, 73, 93 et 103.)

J.K. Baker. The Dragon system - an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):24–29, 1975. (Cité page 2.)

V. Barbu et N. Limnios. Maximum likelihood estimation for hidden semi-Markov models. *Comptes Rendus Mathématique Académie des Sciences de Paris, Series I*, 342:201–205, 2006. (Cité page 57.)

D. Barthélémy et Y. Caraglio. Plant architecture : a dynamic, multilevel and comprehensive approach of plant form and ontogeny. *Annals of Botany*, 99(3):375–407, 2007. (Cité pages 1, 2, 12, 18 et 172.)

D. Barthélémy, S. Sabatier et O. Pascal. Le développement architectural du Noyer commun *Juglans regia L.*. *Forêt Entreprise*, 103:61–68, 1995. (Cité page 160.)

L.E. Baum et T. Petrie. Statistical inference for probalistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966. (Cité page 2.)

L.E. Baum, T. Petrie, G. Soules et N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970. (Cité pages 46 et 73.)

P.J. Bickel, Y. Ritov et T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, 26(4):1614–1635, 1998. (Cité page 57.)

J.G. Booth et J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61(1):265–285, 1999. (Cité pages 35 et 117.)

C. Burge et S. Karlin. Prediction of complete gene structures in human genomic DNA. *Statistical Science*, 13(2):142–162, 1997. (Cité pages 3 et 43.)

B.S. Caffo, W. Jank et G.L. Jones. Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society, Series B*, 67(2):235–251, 2005. (Cité page 36.)

S.G. Candy. Estimation in forest yield models using composite link functions with random effects. *Biometrics*, 53(1):146–160, 1997. (Cité page 1.)

0. Cappé, V. Buchoux et E. Moulines. Quasi-Newton method for maximum likelihood estimation of hidden Markov models. Dans *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 2265–2268, 1998. (Cité page 46.)

O. Cappé, E. Moulines et T. Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. New York, NY : Springer, 2005. (Cité pages 38, 43 et 57.)

G. Celeux et J. Diebolt. The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, 2: 73–82, 1985. (Cité page 34.)

G. Celeux, O. Martin et C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5:243–267, 2005. (Cité page 103.)

P. Champagnat, P. Barnola et S. Lavarenne. Quelques modalités de la croissance rythmique endogène des tiges chez les végétaux ligneux. Dans *Comptes rendus du Colloque International sur l'Arbre*, pages 279–302, Montpellier, Septembre 1986. Naturalia Monspeliensia, n° hors série. (Cité pages 12 et 22.)

S. Chib. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75:79–97, 1996. (Cité pages 55, 78, 104, 105 et 127.)

N. Chopin et F. Pelgrin. Bayesian inference and state number determination for hidden Markov models : an application to the information content of the yield curve about inflation. *Journal of Econometrics*, 123:327–344, 2004. (Cité pages 5, 27, 73, 181 et 185.)

G.A. Churchill. Stochastics models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51:79–94, 1989. (Cité page 43.)

I.B. Collings et T. Rydén. A new maximum likelihood gradient algorithm for on-line hidden Markov model identification. Dans *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 2261–2264, 1998. (Cité page 46.)

S.R. Cosslett et L.F. Lee. Serial correlation in latent discrete variable models. *Journal of Econometrics*, 27:79–97, 1985. (Cité pages 5 et 73.)

R.G. Cowell, A.P. Dawid, S.L. Lauritzen et D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Statistics for Engineering and Information Science. New York, NY : Springer, 1999. (Cité pages 7 et 98.)

M.S. Crouse, R.D. Nowak et R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998. (Cité page 184.)

J. Danusevicius. Use of introduced provenances to increase genetic diversity in local Scots pine populations. *Biologija*, 1:59–61, 2001. (Cité page 19.)

I. David, X. Druart, G. Lagriffoul, E. Manfredi, C. Robert-Granié et L. Bodin. Genetic and environmental effects on semen traits in Lacaune and Manech tête rousse AI rams. *Genetics Selection Evolution*, 39(4):405–419, 2007. (Cité page 124.)

B. Delyon, M. Lavielle et E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1):94–128, 1999. (Cité page 36.)

A.P. Dempster, N.M. Laird et D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977. (Cité pages 29, 31 et 32.)

P.A. Devijver. Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3:369–373, 1985. (Cité pages 49, 50, 52, 114, 127 et 128.)

P.J. Diggle, P. Heagerty, K.Y. Liang et S.L. Zeger. *Analysis of longitudinal data.* Oxford University Press, 2002. (Cité page 63.)

A.J. Dobson et A.G. Barnett. *An introduction to generalized linear models. Third Edition.* London etc. : Chapman & amp, 2008. (Cité page 3.)

J. Dolezal, H. Ishii, V.P. Vetrova, A. Sumida et T. Hara. Tree growth and competition in a Betula platyphylla-Larix cajanderi post-fire forest in central Kamchatka. *Annals of Botany*, 94:333–343, 2004. (Cité page 19.)

J.B. Durand, P. Goncalvès et Y. Guédon. Computational methods for hidden Markov tree models - An application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, 2004. (Cité page 184.)

J.B. Durand, Y. Guédon, Y. Caraglio et E. Costes. Analysis of the plant architecture via tree-structured statistical models : the hidden Markov tree models. *New Phytologist*, 166(3):813–825, 2005. (Cité pages 1 et 184.)

R. Durbin, S.R. Eddy, A. Krogh et G.J. Mitchinson. *Biological Sequence Analysis : Probalistic Models of Protein and Nucleic Acids.* Cambridge : Cambridge University Press, 1998. (Cité page 2.)

J.C. Eickhoff, J. Zhu et Y. Amemiya. On the simulation size and the convergence of the Monte Carlo EM algorithm via likelihood-based distances. *Statistics and Probability Letters*, 67:161–171, 2004. (Cité pages 35 et 36.)

Y. Ephraim et N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569, 2002. (Cité pages 2, 5, 38 et 43.)

L. Fahrmeir et H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, 13:342–368, 1985. (Cité page 62.)

P. Fearnhead. Computational methods for complex stochastic systems : A review of some alternatives to MCMC. *Statistics and Computing*, 18:151–171, 2008. (Cité page 99.)

J.D. Ferguson. Variable duration models for speech. Dans *Proceedings of the Symposium on the Applications of Hidden Markov Models to text and Speech*, volume 61, pages 143–179. ed. J.D. Ferguson, Princeton, New Jersey, 1980. (Cité pages 3 et 38.)

G.M. Fitzmaurice, N.M. Laird et J.H. Ware. *Applied longitudinal analysis.* Wiley Series in Probability and Statistics. Hoboken, NJ : John Wiley & amp ; Sons, 2004. (Cité pages 63 et 140.)

F. Fontaine, H. Chaar, F. Colin, C. Clément, M. Burus et J.L. Druelle. Preformation and neoformation of growth units on 3-year-old seedlings of Quercus petraea. *Canadian Journal of Botany*, 77:1623–1631, 1999. (Cité page 22.)

L.A. Foreman. Generalization of the Viterbi algorithm. *IMA Journal of Management Mathematics*, 4(4):351–367, 1993. (Cité pages 57 et 87.)

G.D. Forney. The Viterbi algorithm. Dans *Proceedings of IEEE*, volume 61, pages 268–278, 1973. (Cité pages 6, 57 et 98.)

J.L. Foulley. Algorithme EM : théorie et application au modèle mixte. *Journal de la Société Française de Statistique*, 143(3-4):1–65, 2002. (Cité pages 32, 33, 37 et 66.)

J.L. Foulley, F. Jaffrézic et C. Robert-Granié. EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data. *Genetics Selection Evolution*, 32:129–141, 2000. (Cité page 76.)

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching models.* Springer Series in Statistics. New York, NY : Springer, 2006. (Cité pages 4, 5 et 70.)

H.C. Fritts. *Tree Rings and Climate.* London : Academic Press, 1976. (Cité page 1.)

M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth et F. Rossi. *GNU Scientific Library. Reference manual.* GNU Free Documentation License, 2006. (Cité page 178.)

A. Gelman, J.B. Carlin, H.S. Stern et D.B. Rubin. *Bayesian data analysis. 2nd Edition.* Boca Raton, FL : Chapman & Hall/CRC, 2004. (Cité pages 104, 105, 116 et 120.)

Y. Guédon. Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3):604–639, 2003. (Cité pages 24, 38, 41, 42, 45, 47, 49, 51, 54, 121, 133, 144, 160 et 164.)

Y. Guédon. Hidden hybrid Markov/semi-Markov chains. *Computational Statistics and Data Analysis*, 49(3):663–688, 2005. (Cité page 51.)

Y. Guédon. Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics and Data Analysis*, 51(5):2379–2409, 2007. (Cité pages 21, 55, 56, 57, 93 et 120.)

Y. Guédon, D. Barthélémy, Y. Caraglio et E. Costes. Pattern analysis in branching and axillary flowering sequences. *Journal of Theoretical Biology*, 212:481–520, 2001. (Cité pages 3, 43 et 185.)

Y. Guédon, Y. Caraglio, P. Heuret, E. Lebarbier et C. Meredieu. Analyzing growth components in trees. *Journal of Theoretical Biology*, 248(3):418–447, 2007. (Cité pages 1, 3, 5, 18, 20, 23, 24, 42, 58, 121, 139, 140, 149, 156 et 185.)

M. Gupta, P. Qu et J.G. Ibrahim. A temporal hidden Markov regression model for the analysis of gene regulatory networks. *Biostatistics*, 8(4):805–820, 2007. (Cité pages 4, 27, 70 et 180.)

F. Hallé et R. Oldeman. *Essai sur l'architecture et la dynamique de croissance des arbres tropicaux.* Masson, Paris, 1970. (Cité pages 1 et 11.)

F. Hallé, R.A.A. Oldeman et P.B. Tomlinson. *Tropical trees and forests. An architectural analysis.* Springer Verlag, Berlin, 1978. (Cité page 12.)

P.J. Hanson, D.E. Todd et J.S. Amthor. A six-year study of sapling and large-tree growth and mortality responses to natural and induced variability in precipitation and throughfall. *Tree Physiology*, 21:345–358, 2001. (Cité page 2.)

D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338, 1977. (Cité page 66.)

C.R. Henderson, O. Kempthorne, S.R. Searle et C.M. von Krosigk. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15:192–218, 1959. (Cité pages 4 et 66.)

P. Heuret. *Analyse et modélisation de séquences d'évènements botaniques : application à la compréhension de la régularité d'expression des processus de croissance, de ramification et de floraison.* Thèse, 2002. (Cité pages 13 et 14.)

P. Heuret, D. Barthélémy, E. Nicolini et C. Atger. 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, 2000. (Cité pages 21, 22 et 171.)

F.L. Hulting et D.A. Harville. Some bayesian and non-bayesian procedures for the analysis of comparative experiments and for small area estimation : computational aspects, frequentist properties and relationships. *Journal of the American Statistical Society*, 86:557–568, 1991. (Cité page 146.)

F. Jelinek. Continuous speech recognition by statistical methods. Dans *Proceedings of the IEEE*, volume 64, pages 532–556, 1976. (Cité page 2.)

S. Kim et P. Smyth. Segmental hidden Markov models with random effects for waveform modelling. *Journal of Machine Learning Research*, 7:645–969, 2006. (Cité pages 5, 6, 27, 96, 99 et 181.)

E. Kuhn et M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM, Probabities and Statistics*, 8:115–131, 2004. (Cité page 36.)

V.G. Kulkarni. *Modeling and analysis of stochastic systems.* London : Chapman & amp ; Hall, 1995. (Cité page 42.)

K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 57(2):425–437, 1995a. (Cité pages 33 et 34.)

K. Lange. A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5(1):1–18, 1995b. (Cité pages 33 et 34.)

K. Lange. *Optimization.* Springer Texts in Statistics. New York, NY : Springer, 2004. (Cité pages 6, 33, 46, 60, 74 et 94.)

R.M. Lanner. Patterns of shoot development in Pinus and their relation to growth potential. Dans *Tree Physiology and yield improvement*, pages 223–243. Cannell MGR at Last FT eds, Academic Press, 1976. (Cité pages 20, 24 et 148.)

S.L. Lauritzen. *Graphical models.* Oxford Statistical Science Series. 17. Oxford : Oxford Univ. Press, 1998. (Cité page 43.)

C. Lavergne, M.J. Martinez et C. Trottier. Finite mixture models for exponential repeated data. Rapport technique 6119, INRIA, 2007. (Cité pages 90 et 117.)

M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85:1501–1510, 2005. (Cité page 185.)

E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736, 2005. (Cité page 185.)

J. Lecoeur et B. Ney. Change with time in potential radiation-use efficiency in field pea. *European Journal of Agronomy*, 19:91–105, 2003. (Cité page 174.)

F. Lefèvre, C. Pichot et J. Pinon. Intra- and interspecific inheritance of some components of the resistance to leaf rust (Melampsora larici-populina Kleb.). *Theoretical and Applied Genetics*, 88:501–507, 1994. (Cité page 19.)

B.G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143, 1992. (Cité pages 43 et 57.)

R.A. Levine et G. Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001. (Cité page 35.)

D.K. Li et D.W. Paty. Magnetic resonance imaging results of the PRISMS trial : a randomized, double-blind, placebo-controlled study of interferon-beta1a in relapsing-remitting multiple sclerosis. *Annals of Neurology*, 46:197–206, 1999. (Cité page 84.)

G. Lindgren. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 5:81–91, 1978. (Cité pages 5, 70 et 73.)

A.V. Lukashin et M. Borodovsky. GeneMark.hmm : new solutions for gene finding. *Nucleic Acids Research*, 26:1107–1115, 1998. (Cité page 3.)

I.L. MacDonald et W. Zucchini. *Hidden Markov and other models for discrete-valued time series.* Monographs on Statistics and Applied Probability. 70. London : Chapman & Hall, 1997. (Cité pages 43, 45 et 57.)

D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. (Cité pages 119 et 133.)

A. Martelli. *Python en concentré. Manuel de référence.* O'Reilly, Paris, 2004. (Cité page 178.)

M.J. Martinez. *Modèles Linéaires Généralisés à effets aléatoires : Contribution au choix de modèle et au modèle de mélange.* Thèse, 2006. (Cité pages 90 et 185.)

P. McCullagh et J.A. Nelder. *Generalized Linear Models. 2nd Edition.* Monographs on Statistics and Applied Probability. 37. London etc. : Chapman & Hall, 1989. (Cité pages 3, 58, 61 et 77.)

C.E. McCulloch. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89:330–335, 1994. (Cité pages 35 et 36.)

C.E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170, 1997. (Cité pages 35, 36 et 117.)

C.E. McCulloch, S.R. Searle et J.M. Neuhaus. *Generalized, Linear, and Mixed models. 2nd Edition.* Hoboken, NJ : John Wiley & amp ; Sons, 2008. (Cité pages 4, 63 et 184.)

G.J. McLachlan. On Aitken's method and other approaches for accelerating convergence of the EM algorithm. Dans *Proceedings of the A.C. Aitken Centenary Conference*, pages 201–209, University of Otago, August 1995. Dunedin : University of Otago Press. (Cité page 33.)

G.J. McLachlan et T. Krishnan. *The EM algorithm and extensions. 2nd Edition.* Wiley Series in Probability and Statistics. New York, NY : John Wiley & Sons, 2008. (Cité pages 5, 30, 32, 34, 36 et 186.)

Q. Meng, C.J. Cieszewski, M. Madden et B. Borders. A linear mixed-effects model of biomass and volume of trees using Landsat ETM+ images. *Forest Ecology and Management*, 244:93–101, 2007. (Cité page 181.)

X.L. Meng et D.B. Rubin. Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm. *Journal of the American Statistical Association*, 86:899–909, 1991. (Cité page 37.)

X.L. Meng et D. van Dyk. The EM algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59(3):511–567, 1997. (Cité pages 29 et 37.)

C. Meredieu. *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, 1998. (Cité pages 20 et 91.)

S. Mutke, J. Gordo, J. Climent et L. Gil. Shoot growth and phenology modelling of grafted Stone pine (*Pinus pinea L.*) in Inner Spain. *Annals of Forest Science*, 60:527–537, 2003. (Cité page 174.)

R.M. Neal et G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. Dans *Jordan, Michael I. (ed.), Learning in graphical models. Proceedings of the NATO ASI*, pages 355–368, Ettore Maiorana Centre, Erice, Italy, September 27 - October 7 1998. Dordrecht : Kluwer Academic Publishers. NATO ASI Series. Series D. Behavioural and Social Sciences 89. (Cité pages 31, 37 et 110.)

J.A. Nelder et R.W.M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society, Series B*, 56(1):61–69, 1972. (Cité pages 3 et 58.)

E. Nicolini. Architecture et gradient morphologénétiques chez de jeunes hêtres en milieu forestier. *Canadian Journal of Botany*, 76:1232–1244, 1998. (Cité page 160.)

Y. Ninomiya et A. Yoshimoto. Statistical Method for Detecting Structural Change in the Growth Process. *Biometrics*, 64(1):46–53, 2008. (Cité page 1.)

A. Nougarède et J. Rembur. Le point végétatif en tant que modèle pour l'étude du cycle cellulaire et de ses points de contrôle. *Bulletin de la Société botanique Française*, 132: 9–34, 1985. (Cité page 12.)

H.D. Patterson et R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971. (Cité page 65.)

F. Picard, E. Lebarbier, E. Budinská et S. Robin. Joint segmentation of multivariate Gaussian processes using mixed linear models. Research Report no. 5, Statistics for systmes biology group, 2007. (Cité pages 99 et 124.)

PRISMS Study Group. Randomised double-blind placebo-controlled study of interferon beta-1a in relapsing/remitting multiple sclerosis. *The Lancet*, 352:1498–1504, 1998. (Cité page 84.)

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989. (Cité page 38.)

C.R. Rao et J. Kleffe. *Estimation of variance components and applications.* North-Holland Series in Statistics and Probability, Vol. 3. Amsterdam etc. : North-Holland, 1988. (Cité page 67.)

A. Raynal-Roques. *La botanique redécouverte.* BELIN-INRA, 1994. (Cité pages 14 et 18.)

F. Rijmen, E.H. Ip, S. Rapp et E.G. Shaw. Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumours. *Journal of the Royal Statistical Society, Series A*, 171(3):739–753, 2008. (Cité pages 4, 7, 27, 96, 98, 180 et 184.)

F. Rijmen, K. Vansteelandt et P. De Boeck. Latent class models for diary method data : parameter estimation by local computations. *Psychometrica*, doi :10.1007/s11336-007-9001-8, 2007. (Cité page 98.)

C.P. Robert et G. Casella. *Monte Carlo statistical methods. 2nd Edition.* Springer Texts in Statistics. New York, NY : Springer, 2004. (Cité pages 34, 36, 103, 104, 105, 116, 120, 127, 130, 133 et 178.)

M. Russell. A segmental HMM for speech pattern matching. Dans *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 499–502, Minneapolis, Avril 1993. (Cité pages 5, 43 et 70.)

S. Sabatier, D. Barthelemy et I. Ducousso. Periods of organogenesis in mono- and bicyclic annual shoots of *Juglans regia* L. (Juglandaceae). *Annals of Botany*, 92:231–238, 2003. (Cité page 23.)

S.R. Searle, G. Casella et C.E. McCulloch. *Variance components.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York, NY : Wiley, 1992. (Cité pages 4, 62 et 67.)

V. Segura, C. Cilas et E. Costes. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects : mixed linear modelling of repeated spatial and temporal measures. *New Phytologist*, 178(2):302–314, 2008. (Cité page 181.)

J.Q. Shi et S.Y. Lee. Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Series B*, (62):77–87, 2000. (Cité pages 36, 103 et 111.)

P. Smyth, D. Heckerman et M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997. (Cité pages 7, 44 et 98.)

D.M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society, Series B*, 46:257–267, 1984. (Cité pages 33 et 76.)

C. Trottier. *Estimation dans les modèles linéaires généralisés à effets aléatoires.* Thèse, 1998. (Cité page 66.)

R. Turner. Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis*, doi :10.1016/j.csda.2008.01.029, 2008. (Cité page 46.)

T.R. Turner, M.A. Cameron et P.J. Thomson. Hidden Markov chains in generalized linear models. *Canadian Journal of Statistics*, 26(1):107–125, 1998. (Cité pages 4, 6, 27, 70, 73 et 180.)

F.C.C. Uzoh et W.W. Oliver. Individual tree height increment model for managed even-aged stands of ponderosa pine throughout the western United States using linear mixed effects models. *Forest Ecology and Management*, 221:147–154, 2006. (Cité page 181.)

G. Verbeke et G. Molenberghs. *Linear mixed models for longitudinal data.* Springer Series in Statistics. Berlin : Springer, 2000. (Cité pages 4 et 63.)

C. Véra. *Modèles linéaires mixtes multiphasiques pour l'analyse de données longitudinales - Application à la croissance des plantes.* Thèse, 2004. (Cité pages 1, 4, 6, 20, 21, 69, 73, 96, 97, 98 et 103.)

P. Wang et M.L. Puterman. Markov Poisson regression models for discrete time series. *Journal of Appled Statistics*, 26(7):855–882, 1999. (Cité pages 4, 6, 27, 70, 73, 74 et 181.)

G.C.G. Wei et M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor's man data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990. (Cité pages 34, 35 et 36.)

N.R. Zhang et D.O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007. (Cité page 185.)

Résumé : Ce travail est consacré à l'étude des combinaisons markoviennes et semi-markoviennes de modèles de régression, i.e. des mélanges finis de modèles de régression avec dépendances (semi-)markoviennes. Cette famille de modèles statistiques permet l'analyse de données structurées en phases successives asynchrones entre individus, influencées par des covariables pouvant varier dans le temps et présentant une hétérogénéité inter-individuelle. L'algorithme d'inférence proposé pour les combinaisons (semi-)markoviennes de modèles linéaires généralisés est un algorithme du gradient EM. Pour les combinaisons (semi-)markoviennes de modèles linéaires mixtes, nous proposons des algorithmes de type MCEM où l'étape E se décompose en deux étapes de restauration conditionnelle : une pour les séquences d'états sachant les effets aléatoires (et les données observées) et une pour les effets aléatoires sachant les séquences d'états (et les données observées). Différentes méthodes de restauration conditionnelle sont présentées. Nous étudions deux types d'effets aléatoires : des effets aléatoires individuels et des effets aléatoires temporels. L'intérêt de cette famille de modèles est illustré par l'analyse de la croissance d'arbres forestiers en fonctions de facteurs climatiques. Ces modèles nous permettent d'identifier et de caractériser les trois principales composantes de la croissance (la composante ontogénique, la composante environnementale et la composante individuelle). Nous montrons que le poids de chaque composante varie en fonction de l'espèce et des interventions sylvicoles.

Mots-clés : Chaîne de Markov cachée, semi-chaîne de Markov cachée, modèle linéaire généralisé, modèle linéaire mixte, algorithme MCEM, composantes de la croissance d'arbres.

Title : Markov and semi-Markov switching regression models. Application to forest tree growth.

Abstract : This work focuses on Markov and semi-Markov switching regression models, i.e. finite mixtures of regression models with (semi-)Markovian dependencies. These statistical models enable to analyse data structured as a succession of stationary phases that are asynchronous between individuals, influenced by time-varying covariates and which present inter-individual heterogeneity. The proposed inference algorithm for (semi-)Markov switching generalized linear models is a gradient EM algorithm. For (semi-)Markov switching linear mixed models, we propose MCEM-like algorithms whose E-step decomposes into two conditional restoration steps : one for the random effects given the state sequences (and the observed data) and one for the state sequences given the random effects (and the observed data). Various conditional restoration steps are presented. We study two types of random effects : individual-wise random effects and environmental random effects. The relevance of these models is illustrated by the analysis of forest tree growth influenced by climatic covariates. These models allow us to identify and characterize the three main growth components (ontogenetic component, environmental component and individual component). We show that the weight of each component varies according to species and silvicultural interventions.

Keywords : Hidden Markov chain, hidden semi-Markov chain, generalized linear model, linear mixed model, MCEM algorithm, tree growth components.

Discipline : Mathématiques appliquées et applications des mathématiques.

Laboratoire : CIRAD, UMR Développement et Amélioration des Plantes et INRIA, Équipe-projet Virtual Plants, TA A-96/02, Av. Agropolis, 34398 Montpellier Cedex 5.