UNIVERSITE PARIS I – PANTHEON – SORBONNE

U.F.R. de MATHEMATIQUES ET INFORMATIQUE

THESE DE DOCTORAT

présentée par

Madalina OLTEANU

le 13 décembre 2006

en vue de l'obtention du titre de

Docteur en Sciences

spécialité : Mathématiques Appliquées

Modèles à changements de régime : applications aux données financières

---

Directeurs de thèse :

Marie	COTTRELL	Professeur à l'Université Paris I
Joseph	RYNKIEWICZ	Maître de Conférences à l'Université Paris I

JURY :

Jean-Marc	AZAIS	Rapporteur
Marie	COTTRELL	Directeur
Elisabeth	GASSIAT	Examinateur
Patrice	GAUBERT	Examinateur
Marc	LAVIELLE	Examinateur
Joseph	RYNKIEWICZ	Co-directeur
Jian-Feng	YAO	Rapporteur

Me voilà à la fin ...

Je l'avoue, j'ai longtemps réfléchi à ce que j'allais dire dans ces quelques lignes et surtout à comment j'allais le dire. Je sais bien qu'à part quelques cas heureux, les matheux n'ont pas vraiment l'esprit littéraire et là j'ai bien peur de ne pas faire exception. J'essaierai donc de faire simple.

Je voudrais d'abord remercier une personne qui m'éblouie chaque jour par son énergie, sa générosité, sa disponibilité et la capacité incroyable de gérer les problèmes du labo, qu'il s'agisse de la recherche, des missions, des soutenances de thèse, des déménagements, du stock de café ou de l'organisation des pots. Un grand merci à ma directrice, Marie Cottrell.

Je remercie aussi mon deuxième directeur de thèse, Joseph Rynkiewicz, d'avoir encadré mon travail, ses conseils et son humour ont été essentiels à l'aboutissement de cette thèse. Son obstination aussi, sans laquelle je n'aurais jamais renoncé à Microsoft pour découvrir Linux.

Pour avoir relu ce document et pour m'avoir fait des commentaires et des remarques précieuses, je remercie également Xavier Guyon.

Mes remerciements s'adressent aussi à Jian-Feng Yao et Jean-Marc Azaïs pour avoir accepté de rapporter mon travail et pour leur participation à ce jury.

Je remercie Elisabeth Gassiat, Marc Lavielle et Patrice Gaubert de m'avoir fait l'honneur de faire partie de mon jury.

Le tour aux Samosiens maintenant... mais avant, une petite histoire. A la fin des années 80, Cioran avouait dans une interview que son arrivée en France lui a permis de découvrir deux choses : l'art de l'écriture et celui de la nourriture. Il habitait un hôtel du Cartier Latin et chaque matin il entendait son concierge discutant avec sa femme de ce qu'ils allaient manger à midi. Au début, il a cru qu'ils avaient des invités mais, le dialogue se répétant chaque jour, il a compris qu'il s'agissait d'un concept de vie... gastronomique. Je vous remercie tous de m'avoir fait découvrir une équipe conviviale et accueillante, mais aussi les habitudes gastronomiques et autres (je me rappelle bien les débats sur la Constitution Européenne) des français. Je n'oublie pas non plus les autres "émigrés" du labo, nos discussions sur les cartes de séjour, les permis de travail et les magnifiques gâteaux tunisiens.

Merci à mes amis d'ici et de là-bas. A ceux d'ici pour avoir accepté la petite roumaine malgré les "différences culturelles" et pour avoir su comment faire passer le mal du pays, tellement pesant parfois. A ceux de là-bas pour avoir gardé des liens très forts malgré la distance et de m'avoir rappelée toujours et encore d'où je venais et quelles étaient mes racines.

Je remercie mes parents de m'avoir fait confiance et de m'avoir suportée toujours et en toute occasion. Il y a vingt ans, ils m'achetaient des "Gazettes Mathématiques" pour remplacer les BD inexistantes. Pourtant, ils n'avaient pas songé une seconde que j'allais vouloir en faire un métier.

Pour son soutien inconditionnel et sa bonne humeur sans lesquels je n'aurais pas réussi à arriver au bout de ces derniers mois, merci à Andrei.

à ma grand-mère ...

Alta matematica...

Noi stim ca unu ori unu fac unu,
dar un inorog ori o para
nu stim cit face.
Stim ca cinci fara patru fac unu,
dar un nor fara o corabie
nu stim cit face.
Stim, noi stim ca opt
impartit la opt fac unu,
dar un munte impartit la o capra
nu stim cit face.
Stim ca unu plus unu fac doi,
dar eu si cu tine,
nu stim, vai, nu stim cit facem.

Ah, dar o plapuma
inmultita cu un iepure
face o roscovana, desigur,
o varza impartita la un steag
fac un porc,
un cal fara un tramvai
face un inger,
o conopida plus un ou,
face un astragal...

Numai tu si cu mine
inmultiti si impartiti
adunati si scazuti
raminem aceiasi...

Pieri din mintea mea!
Revino-mi in inima!

Nichita Stanescu

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Cadre de l'étude

Cours du pétrole, taux de cholestérol, radars automatiques, trente-cinq heures ou pollution à l'ozone, sujets communs des journaux télévisés qui font l'objet de questions sur... séries temporelles et modélisation. Le cours du pétrole augmente et les analystes cherchent des causes dans les guerres télégéniques et des solutions dans les énergies renouvelables. Le taux de cholestérol est élevé et les médecins n'arrêtent pas d'expliquer que faire du sport le diminue. Le gouvernement introduit des radars automatiques en espérant réduire le nombre de tués sur les routes et justifie, selon la couleur politique, la hausse ou la baisse du chômage par les trente-cinq heures. Ces réponses et justifications sont basées sur des valeurs observées au cours du temps, des séries chronologiques que l'on essaie d'expliquer, modéliser et prédire.

Pour la majorité de ces phénomènes, il existe souvent une dépendance temporelle entre les observations, ce qui mène à des modélisations de type autorégressif : on utilise le passé pour expliquer le présent et prédire le futur. Modéliser et prédire une série chronologique suppose, dans la plupart des cas, de faire des hypothèses sur son comportement. Puisqu'il s'agit de séries non-déterministes, il va falloir que la composante aléatoire "varie, mais pas trop". Cette condition se traduira par la "stationnarité", qui implique une certaine régularité du processus et permet de dériver ses propriétés asymptotiques.

Dans toute la suite, on se place dans un cadre paramétrique : trouver un modèle pour les observations revient à déterminer les paramètres d'une fonction dont la forme est connue "a priori" et qui vérifie

$$y_t = f\left(\varepsilon_t, y_{t-1}, ...\right)$$

où y_t représente l'observation à l'instant t et ε_t est le bruit aléatoire.

Depuis les années 20 (par exemple l'étude de Yule sur les taches solaires publiée en 1927) et jusqu'aux années 80, les modèles linéaires à bruit gaussien AR ont tenu la tête d'affiche. Ils modélisent le présent par une combinaison linéaire d'un nombre fini de retards plus un bruit :

$$y_t = a_0 + a_1 y_{t-1} + ... + a_p y_{t-p} + \varepsilon_t$$

Largement utilisés pendant ce temps, ils ont l'avantage d'une écriture simple qui permet le dévelopement d'une théorie statistique complète et de plus l'estimation des paramètres ne demande pas des capacités de calcul importantes. Cependant, depuis une vingtaine d'années, les limitations des modèles linéaires ont été soulignées, les données réelles ayant souvent des caractéristiques non-linéaires qui ne sont pas prises en compte. Dans les séries financières, par exemple, on observe des périodes très volatiles, suivies de périodes plus stables. Ce phénomène appelé *"volatility clustering"* ne sera donc pas expliqué par un modèle à volatilité constante dans le temps. Pour d'autres séries on peut avoir d'autres non-linéarités comme l'asymétrie dans les valeurs des crues d'une rivière ou la périodicité qui apparaît dans l'évolution du PIB suivant les périodes de récession ou de croissance économique.

Pour résoudre les problèmes évoqués, des modèles plus complexes ont été proposés à partir des années 80. Sans chercher l'exhaustivité, en voici quelques-uns des plus utilisés aujourd'hui : les modèles à volatilité conditionnelle hétéroscedastique de type ARCH, GARCH (Engle (1982), Bollerslev (1986)), les modèles à seuil ou linéaires par morceaux (Tong (1983)), les modèles à changements de régime (Hamilton (1989)) ou même les modèles provenant de l'intelligence artificielle comme les perceptrons multicouches qui ont la propriété pratique d'approximateurs universels (Hornik et alii (1989)).

Dans ce document, on s'intéresse en particulier aux séries ayant des changements de régimes ou des périodicités et aux modèles pouvant être employés dans ces cas. Pour avoir un aperçu du sujet et des résultats intéressants qu'un modèle à changements de régime peut apporter, mais aussi des problèmes statistiques qui peuvent apparaître, voici un exemple.

1.2 Périodisation du bimétallisme international 1821-1873

Avant que le standard Or ne se soit répandu à partir des années 1870, le système monétaire international reposait sur deux métaux précieux, l'or et l'argent, ce qui créa des opportunités d'arbitrage sur les taux de change et les payments internationaux dans les principaux centres financiers européens, Hambourg, Londres et Paris. Une période intéressante à étudier sur ces trois marchés commence en 1821, année à partir de laquelle les billets de la Banque d'Angleterre sont convertibles en or, et s'étend jusqu'à 1873 quand, dû à la position de leader économique et financier de l'Angleterre, le standard Or est adopté en Allemagne, France et Etats Unis.

Dans notre étude (Boyer-Xambeu et alii (2006)) on a voulu montrer que pendant cette période, on peut déjà parler d'un système monétaire européen opérant entre trois marchés et mettant en contact trois organisations différentes, le standard Or à Londres, Argent à Hambourg et le double standard Or-Argent à Paris. Ce système apparaît comme une alternative à l'hypothèse d'un seul système basé sur un seul pays puissant et la monnaie correspondante. Ce mode de fonctionnement des trois marchés est appelé "bimétallisme

international" et constitue, peut-être, le précédent historique d'un possible système de payment basé sur deux monnaies principales, l'euro et le dollar américain.

Les buts de la modélisation ont été, d'une part, de justifier l'hypothèse de l'existence de ce système monétaire réglé par les trois principales monnaies et, d'autre part, d'expliquer le rôle majeur des trois marchés, en particulier celui de Hambourg qui reposait sur l'Argent. En même temps, il fallait regarder l'influence de facteurs extérieurs comme les réformes institutionnelles ou les chocs monétaires et politiques. Un modèle à changements de régime a été choisi pour révéler l'éventuelle existence de sous-périodes distinctes, pendant lesquelles l'évolution des marchés serait différente.

La base de données a été construite à partir des principaux journaux de l'époque avec des valeurs manquantes dues soit à l'absence de certains journaux dans les archives, soit à l'arrêt des quotations à cause de crises financières ou politiques. Pour améliorer la qualité de la base et réduire le nombre de "trous", les données manquantes de la première catégorie ont été remplacées par des valeurs moyennes estimées à partir d'une classification non-supervisée réalisée avec une carte de Kohonen. Les séries dont on a disposé sont les rapports des taux de change entre les trois monnaies, les cours locaux des métaux précieux qui sont libres de l'intervention d'une banque centrale (l'Argent à Londres et l'Or à Hambourg), ainsi que l'évolution du rapport Or/Argent sur les trois marchés.

Le modèle estimé explique l'évolution du rapport Or/Argent à Paris ($poa\,(t)$) en fonction de ses valeurs passées ($poa\,(t-1)$ et $poa\,(t-2)$) et des autres variables disponibles (le rapport Or/Argent à Londres et Hambourg, $lgs\,(t)$ et, respectivement, $hoa\,(t)$ et le rapport entre les taux de change des trois monnaies $hpl\,(t)$). Il s'écrit sous la forme :

$$poa\,(t) = a^1_{x_t} poa\,(t-1) + a^2_{x_t} poa\,(t-2) + a^3_{x_t} lgs\,(t) + a^4_{x_t} hoa\,(t) + a^5_{x_t} hpl\,(t) + \sigma_{x_t}\varepsilon_t,$$

où x_t est une chaîne de Markov homogène à deux états.

On obtient deux régimes très stables ($\mathbb{P}\,(x_t = i \mid x_{t-1} = i) > 0,8$ pour $i = 1, 2$) : dans le premier, il y a une forte corrélation avec le même rapport à Hambourg et avec les valeurs passées à Paris, tandis que dans le deuxième, la corrélation avec Hambourg est toujours présente, mais beaucoup plus faible, et apparaît, avec un coefficient très important, le rapport entre les taux de change des trois monnaies. Ceci peut aussi être vu comme une séparation entre un régime dans lequel Paris et Hambourg fonctionnent indépendament de Londres et un deuxième régime dans lequel les trois marchés fonctionnent ensemble.

De plus, en identifiant "a posteriori" les périodes de temps qui correspondent à chaque régime à l'aide des probabilités conditionnelles estimées (dans la Figure 1.1, la probabilité conditionnelle d'être dans le premier régime est représentée en pointillé), on obtient que le deuxième régime correspond à des périodes de crise au Royaume-Uni : 1824-25, 1827-28 crises économiques et bancaires, 1832-33 supression de l'esclavage dans tout l'Empire Britannique et des privilèges de la Compagnie des Indes, 1855 guerre de Crimée, 1857 récession économique due à une surproduction d'or qui engendre la faillite des systèmes bancaires aux Etats-Unis et en Angleterre.

La dépendance forte du cours à Paris avec celui de Londres pendant ces périodes de crise et leur séparation nette par le modèle à deux régimes, ainsi que la corrélation avec Hambourg,

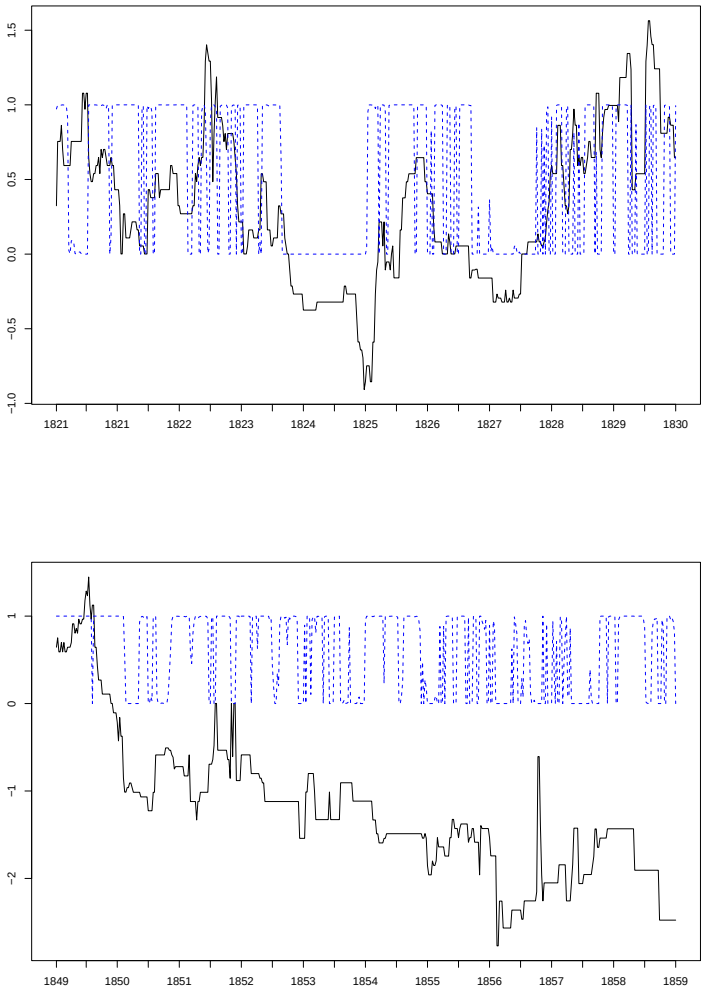

FIG. 1.1 – Le cours Or/Argent à Paris (en ligne continue) et les probabilités conditionnelles du premier régime (en pointillé)

qui se révèle plus importante pendant les périodes "calmes" au Royaume-Uni confirment l'hypothèse de l'existence d'un système monétaire européen basé sur trois marchés interdépendants et met en évidence un phénomène de contagion quand le marché anglais est en crise. Cet exemple, ainsi que d'autres présents dans la literature économique (Hamilton (1989), Hamilton et Raj (2002)) montrent combien les modèles à changements de régime sont importants dans la pratique et peuvent répondre à des questions diverses. Cependant, leur propriétés statistiques n'ont pas toujours été établies, surtout dans le cas où le nombre de régimes n'est pas fixé à l'avance.

1.3 Organisation de la thèse

Ce document s'organise autour du but suivant : comment trouver un bon modèle autorégressif pour les séries temporelles qui subissent des changements de comportement ? Les exemples les plus courants, et celui de la section 1.2 en fait partie, sont les séries provenant des marchés financiers où des périodes très volatiles alternent avec des périodes moins turbulentes. Dans un premier temps on décrit les modèles à changements de régime en général et on s'intéresse à une application à la quantification des chocs sur les marchés financiers. Le problème du choix du nombre de comportements différents ou de régimes apparaît alors naturellement. Deux approches seront proposées, la première est basée sur un critère de vraisemblance pénalisée, la deuxième revient à faire une classification non-supervisée des données.

Voici les résumés des différents chapitres :

Chapitre 2 : Modèles non-linéaires à changements de régime

Ce chapitre est dédié à un bref état de l'art sur les modèles autorégressifs non-linéaires. Parmi tous les modèles proposés pendant les vingt dernières années, on s'intéresse aux modèles à changements de régimes et à leur possibles applications dans la pratique pour détecter des comportements différents à l'intérieur d'une série de données chronologiques. Plus précisément, on étudie les cas où les changements de régime sont engendrés par une autre série non-observée qui peut être considérée comme une variable endogène cachée à valeurs dans un espace d'états fini et dont les passages d'un état vers un autre sont markoviens. Dans la suite de ce document on sera obligé parfois de se restreindre au cas particulier des changements d'état indépendants. Le nombre de régimes sera défini dans la suite par la dimension de l'espace d'états de la variable cachée.

Chapitre 3 : Caractérisation des crises financières à l'aide des modèles hybrides HMC-MLP

Les marchés financiers sont le lieu de violents mouvements qui font régulièrement l'actualité de la presse économique. Pour quantifier ces turbulences, un indicateur de crise *IMS* (*Index of Market Shocks*, Maillet et Michel (2003)) s'inspirant d'une analogie avec l'échelle de Richter et calculé à partir de la volatilité historique a été introduit récemment. Le but étant de caractériser les formes de ces turbulences -violentes et de courte durée versus étalées dans le temps et de durée plus longue- une modélisation à changements de régime a été proposée. On s'attend à pouvoir séparer un régime exceptionnel, qui régit les événements extrêmes, et un (ou plusieurs) régimes qui correspond(ent) aux évolutions "normales" des

marchés. Pour tenir compte de la forte non-linéarité de la volatilité et donc de l'IMS, le modèle choisi est doublement non-linéaire : d'une part on prend en compte des changements de régime markoviens et d'autre part on utilise des perceptrons multi-couches dans chaque régime. Ce type de modèle hybride a été abrégé en HMC-MLP, *Hidden Markov Chain - Multilayer Perceptron*. Dans ce chapitre, après la description des données et la construction de l'indice de crise, on présente l'algorithme d'estimation et d'ajustement aux données, la qualité du modèle en prévision et une étude de la séparation des états - crise versus accalmie. Finalement, le problème du choix du nombre de régimes est évoqué.

Chapitre 4 : Stationnarité et propriétés de dépendance faible des modèles autorégressifs à changements de régime markoviens

Pour estimer le nombre de régimes d'un modèle autorégressif, on travaille dans un cadre dépendant. Des résultats asymptotiques comme la loi des grands nombres et le théorème central limite ont été démontrés récemment sous des conditions de stationnarité et mélangeance, cette dernière notion assurant une indépendance asymptotique. Pour pouvoir démontrer les résultats énoncés au chapitre suivant, on revient sur ces deux propriétés et on rappelle sous quelles hypothèses, un modèle à changements de régime les vérifie.

Chapitre 5 : Estimation du nombre d'états pour les modèles autorégressifs à changements de régime

Estimer et tester le nombre d'états dans un modèle à changements de régime soulève le problème de la non-identifiabilité du modèle. Sous cette hypothèse, les conditions de régularité usuelles ne sont pas vérifiées et la théorie classique de la convergence de l'estimateur du maximum de vraisemblance ne peut pas être appliquée. Récemment, Keribin (2000) et Gassiat (2002) ont montré la consistence d'un estimateur de maximum de vraisemblance marginale pénalisée dans le cadre des modèles de mélange et, respectivement, des chaînes de Markov cachées. Le cas des modèles autorégressifs n'a pourtant pas été traité.

Dans un premier temps, on étend le résultat de Gassiat aux modèles autorégressifs à changements de régime indépendants. La consistence de l'estimateur du nombre de régimes est démontrée sous des conditions portant sur la stationnarité, la β-mélangeance et l'entropie à crochets de la classe des fonctions scores généralisés. Ces hypothèses sont ensuite vérifiées dans le cadre d'un bruit gaussien, le problème de la non-identifiabilité étant résolu par une reparamétrisation du modèle inspirée de Liu et Shao (2003). Des simulations sont proposées pour illustrer le résultat, ses propriétés de stabilité et sa vitesse de convergence. Dans la dernière partie, on démontre qu'une extension aux modèles autorégressifs à changements de régime markoviens n'est pas immédiate, la fonction de coût utilisée n'étant plus un contraste.

Chapitre 6 : Une méthode empirique pour calculer le nombre d'états dans un modèle à changements de régime

Les inconvénients de l'estimateur consistent construit au chapitre précédent sont évalués : si on veut rester sous des hypothèses assez faibles, on n'a pas la vitesse de convergence, ni la loi limite et l'extension du critère de vraisemblance marginale pénalisée des modèles à changements de régime indépendants à ceux à changements markoviens n'est pas possible. Cela nous a encouragé à chercher d'autres méthodes pour déterminer le nombre de régimes. Puisqu'on considère le cadre où dans chaque régime la fonction de régression est linéaire,

trouver le nombre de régimes revient à trouver le nombre d'hyperplans de régression qui caractérisent le modèle global. Cette dernière formulation peut être traitée comme un problème de classification non-supervisée et a l'avantage de ne pas tenir compte de la nature des changements de régime qui peuvent être indépendants, markoviens ou dépendant de seuils. On propose un nouvel algorithme où l'on combine les cartes de Kohonen et une classification hiérarchique pour laquelle on définit une nouvelle dispersion. La méthode est testée sur plusieurs jeux de données, simulées et réelles, et on discute ensuite ses avantages et ses limites.

Chapitre 2

Modèles non-linéaires à changements de régime

Parmi la large variété des modèles non-linéaires cités dans l'introduction, on s'intéresse à ceux qui subissent des changements de comportement ou de régime. Ce chapitre est consacré à un bref état de l'art sur le sujet. On commence par rappeler la définition des chaînes de Markov cachées et on continue avec leur généralisation dans un cadre autorégressif. On passe en revue les problèmes liés à l'estimation, les différentes méthodes proposées et les propriétés statistiques démontrées dans la littérature. Pour chaque classe de modèles, on présente la question restant parfois ouverte du calcul du nombre de régimes, en soulignant les contributions que ce document se propose d'apporter.

2.1 Chaînes de Markov cachées

Introduites dans la reconnaissance de la parole à la fin des années 60, les chaînes de Markov cachées sont les "pionniers" des modèles à changements de régime. De manière formelle, la définition est la suivante :

Définition 2.1 *: Une chaîne de Markov cachée est un processus bivarié $(X_t, Y_t)_{t\in\mathbb{Z}}$ tel que*

- *(X_t) est une chaîne de Markov homogène à valeurs dans un espace d'états fini de dimension fixée $E = \{e_1, ..., e_N\}$, $N \in \mathbb{N}^\star$, caractérisée par sa matrice de transition $A = (a_{ij})_{i,j=1,...,N}$, où*

$$a_{ij} = \mathbb{P}(X_{t+1} = e_j \mid X_t = e_i)$$

- *(Y_t) est une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$, $d \geq 1$, indépendantes conditionnellement à (X_t) et $\mathcal{L}(Y_t \mid X_t = e_i) \equiv \mathcal{L}_i$.*

Sans perte de généralité, on peut identifier l'espace des états E à la base canonique de $\mathbb{R}^N$.

Schématiquement, un modèle à chaîne de Markov cachée fonctionne comme dans la Figure 2.1 :

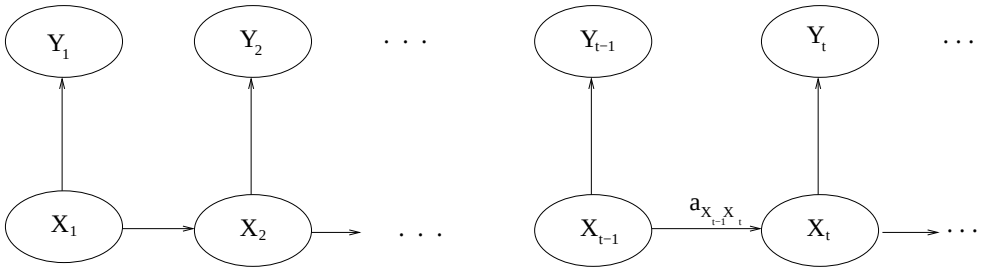

FIG. 2.1 – Diagramme d'un modèle à chaîne de Markov cachée

Puisque la chaîne de Markov (X_t) n'est pas observée directement, ce type de modèles reçoit l'adjectif "caché" et la remarque essentielle à faire est que l'inférence statistique s'étudie uniquement à partir du processus observé (Y_t), les probabilités conditionnelles d'appartenir à un régime ou à un autre (on appelle "régime" un état e_i de (X_t)) se calculant "a posteriori", une fois les paramètres du modèle estimés.

Dans toute la suite, on se place dans un cadre paramétrique, la loi de Y_t conditionnellement à $X_t = e_i$ dépend d'un paramètre θ_i qui est supposé à l'intérieur de Θ, sous-ensemble compact de $\mathbb{R}^k$, $k \geq 1$. Le paramètre global sera noté

$$\theta = (a_{ij}, i, j = 1, ..., N, \theta_i, i = 1, ..., N) \in]0, 1[^{N \times N} \times \Theta^N$$

Largement utilisés depuis leur introduction par Baum et Petrie (1966), ces modèles sont devenus des outils indispensables dans la reconnaisance de la parole (Rabiner (1989), Bourlard et Morgan (1994)), l'analyse des séquences biologiques et la bio-informatique (Durbin et alii. (1998)), mais des applications dans l'économétrie et l'analyse des séries temporelles (De Jong et Shephard (1995), Chan et Ledolter (1995), Mac Donald et Zucchini (1997), Kim et alii. (1998)) ont aussi donné des résultats intéressants.

Cependant, leurs propriétés statistiques n'ont commencé à être étudiées que récemment. Une première approche avait été proposée par Baum et Petrie (1966), mais dans un cadre restreint. En supposant le processus (X_t) stationnaire (la loi de probabilité initiale est égale à la loi invariante) et l'espace des valeurs de (Y_t) fini, ils ont montré la consistence et la normalité asymptotique de l'estimateur de maximum de vraisemblance de θ. Le cadre théorique a été élargi quelques dizaines d'années plus tard. En considérant un espace topologique général pour (Y_t) et sous l'hypothèse de stationnarité de (X_t), Leroux (1992) a démontré la consistence de l'estimateur de maximum de vraisemblance. La normalité asymptotique a été obtenue, sous l'hypothèse de consistance, d'abord de manière locale par Bickel et Ritov (1996) et puis globale par Bickel, Ritov et Ryden (1998).

Des résultats dans le cadre non-stationnaire ont été démontrés par Bakry et alii. (1997), qui étendent la preuve de Baum et Petrie (1966), et Le Gland et Mevel (2000a, 2000b) qui, indépendamment des travaux de Leroux et Bickel et alii., ont développé une technique différente pour démontrer la consistence et la normalité asymptotique dans le cas où (Y_t) prend des valeurs dans un espace séparable et complet.

Tous ces résultats ont été démontrés sous l'hypothèse d'un nombre de régimes fini, connu et fixé à l'avance. Travailler avec un nombre inconnu soulève des problèmes de non-identifiabilité du modèle et les conditions de régularité usuelles ne sont plus remplies. Récemment, Gassiat (2002) a proposé un estimateur de maximum de vraisemblance marginale pénalisée pour le nombre de régimes d'une chaîne de Markov cachée pour lequel, avec une condition de stationnarité sur la chaîne de Markov étendue (X_t, Y_t), elle a démontré la consistence faible. Cependant, les questions sur la vitesse de convergence et l'existence d'une loi limite restent ouvertes.

Ces dernières années, des généralisations aux chaînes de Markov à espace d'états infini ont été proposées. La normalité asymptotique a été prouvée parallèlement par Jensen et Petersen (1999) et Douc et Matias (2001). Les premiers se placent dans un cadre stationnaire et utilisent les techiques de Baum et Petrie (1966) et Bickel, Ritov et Ryden (1998) pour obtenir le résultat sous l'hypothèse de consistence de l'estimateur, tandis que les derniers étendent la méthode de Le Gland et Mevel (2000a, 2000b) et démontrent la consistence, la normalité asymptotique et l'efficacité sans conditions de stationnarité.

La problèmatique abordée dans ce document ne s'adresse donc pas aux modèles généraux du dernier paragraphe. Dans les modèles qui seront considérés par la suite, le processus observé (Y_t) est défini sur un espace continu et la chaîne de Markov (X_t) est à nombre d'états fini, comme dans la Définition 2.1.

2.2 Modèles autorégressifs localement linéaires à changements de régime markoviens

En faisant intervenir des retards dans l'expression de Y_t, les modèles autorégressifs à changements de régime markoviens deviennent une extension naturelle des modèles à chaîne de Markov cachée. Dans ce cas, le processus observé Y_t ne dépend plus uniquement de X_t, mais aussi de son passé $Y_{t-1}, ..., Y_{t-p}$ avec $p \geq 1$ fixé. Pour des raisons d'écriture plus simple, on a choisi de considérer le même nombre de retards dans tous les régimes. Prendre des nombres de retards différents dans chaque régime est immédiat, il suffit d'annuler des coefficients. Dans ce cadre, le modèle est autorégressif, linéaire par morceaux et ses paramètres peuvent être vus comme des réalisations de la chaîne de Markov (X_t).

On introduit la définition suivante pour décrire cette classe de modèles :

Définition 2.2 *: On définit un modèle autorégressif linéaire par morceaux et à changements de régime markoviens* ***HMC-AR(p)*** *par un processus bivarié $(X_t, Y_t)_{t\in\mathbb{Z}}$ tel que*

- *(X_t) est une chaîne de Markov homogène à valeurs dans un espace d'états fini $E = \{e_1, ..., e_N\}$, $N \in \mathbb{N}^\star$, caractérisée par sa matrice de transition $A = (a_{ij})_{i,j=1,...,N}$, où*

$$a_{ij} = \mathbb{P}(X_{t+1} = e_j \mid X_t = e_i)$$

- *$Y_t = F_{X_t}(Y_{t-1}, ..., Y_{t-p}) + \sigma_{X_t}\varepsilon_t$, où, pour tout $i = 1, ..., N$,*

$$F_{e_i}(Y_{t-1}, ..., Y_{t-p}) = a_0^{e_i} + a_1^{e_i} Y_{t-1} + ... + a_p^{e_i} Y_{t-p}$$

est une fonction linéaire à coefficients dans un compact de $\mathbb{R}^d$, $\sigma_{e_i} \in \{\sigma_1, ..., \sigma_N\} \subset \left(\mathbb{R}_+^\star\right)^N$ *et* (ε_t) *est une suite i.i.d.*

Cette extension des chaînes de Markov cachées a été introduite par Hamilton (1989) pour modéliser la série du Produit National Brut aux Etats-Unis. Le caractère périodique de cette série dû à des périodes de croissance ou de récession a motivé ce choix, les changements d'une période à l'autre n'étant pas, de plus, observés directement. Hamilton (1989) a construit un modèle basé sur deux régimes, le premier associé à un "taux de croissance positif" et le deuxième à un "taux de croissance négatif", en espérant pouvoir déduire un critère objectif pour définir et mesurer les récessions économiques. Une justification théorique du choix de deux régimes contre toute autre possibilité n'a cependant pas été fournie.

Pour l'estimation des paramètres, Hamilton (1989) utilise une méthode de type filtre prédictif pour calculer la vraisemblance et l'estimateur de maximum de vraisemblance. Dans un article ultérieur (Hamilton (1990)), il propose un algorithme de type EM pour approcher l'estimateur de maximum de vraisemblance. Parallèlement, des procédures récursives d'estimation ont été aussi proposées (Millnert (1986), Holst et alii. (1994)). Depuis, en s'appuyant sur la validation empirique de l'algorithme EM, les modèles linéaires à changements de régime ont été largement employés, surtout en économétrie où ils ont fourni des résultats intéressants en séparation des données et prévision. Cependant, aucun résultat théorique sur l'estimation n'avait été démontré jusqu'à très récemment. Francq et Roussignol (1998) et Krishnamurthy et Ryden (1998) ont prouvé la consistence de l'estimateur de maximum de vraisemblance sous l'hypothèse de stationnarité pour (X_t). Dans ces résultats, le nombre de régimes est fixé à l'avance, la construction d'un estimateur pour le nombre de régimes posant les mêmes problèmes de non-identifiabilité du modèle qu'on vient de citer dans la section précédente et sur lesquels on reviendra dans les chapitres 4 et 5. Au chapitre 5, on étendra le résultat de Gassiat (2002) au cadre des modèles autorégressifs à changements markoviens.

Même si cela reste en dehors de notre sujet, il faut citer le résultat très général démontré dans Douc, Moulines et Ryden (2004) qui permet à (X_t) de prendre des valeurs dans un espace compact infini. Sous des conditions de régularité standard et sans l'hypothèse de stationnarité de (X_t), ils ont prouvé la consistence et la normalité asymptotique de l'estimateur du maximum de vraisemblance.

2.3 Modèles hybrides avec des perceptrons multicouches et changements de régime markoviens

La généralisation des modèles précédents aux modèles plus complexes où les fonctions autorégressives dans chaque régime peuvent avoir n'importe quelle forme non-linéaire est immédiate. Dans cette section on présente brièvement une classe de modèles non-linéaires à changements de régime pour lesquels chaque régime est caractérisé par un perceptron multi-couches (MLP - *Multilayer perceptron*). Rappelons que dans le langage statistique, un perceptron multi-couches (MLP) est une fonction paramétrique non-linéaire s'inspirant des neurones biologiques. Pour que l'exposé soit complet, on commence par introduire la

définition du neurone formel et du perceptron multi-couches et ensuite on donne la définition des modèles hybrides HMC-MLP (*Hidden Markov Chain - Multilayer Perceptron*).

Le neurone formel ou le perceptron simple, introduit par Cullogh et Pitts (1943) et Rosenblatt (1962), est une approximation très schématique du neurone biologique (voir la Figure 2.2).

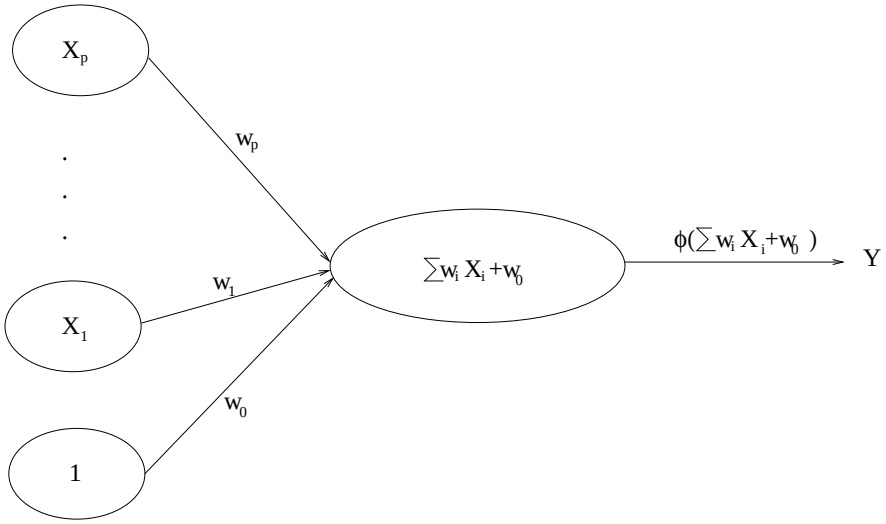

FIG. 2.2 – Schéma d'un neurone formel

La description formelle est la suivante :

Définition 2.3 *: Si $X_1, ..., X_p$ sont les variables d'entrée et Y est la variable de sortie, un neurone formel est défini par la fonction*

$$Y = \phi\left(\sum_{i=1}^{p} w_i X_i + w_0\right)$$

où $w_0, w_1, ..., w_p$ sont les paramètres (appelés aussi poids) et $\phi : \mathbb{R} \to \mathbb{R}$ est une fonction d'activation sigmoïde :

$$\phi(x) = th(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Dans l'analogie avec le neurone biologique, $w_1, ..., w_p$ sont les poids synaptiques, $-w_0$ la valeur seuil au delà de laquelle le neurone s'active, $\sum_{i=1}^{p} w_i X_i + w_0$ le potentiel, tandis que $\phi\left(\sum_{i=1}^{p} w_i X_i + w_0\right)$ représente la sortie de l'axone.

Dans les applications, le neurone formel fonctionne très bien lorsque sa tâche est de séparer deux ensembles linéairement séparables, mais il fait défaut dans la modélisation d'une fonction non-linéaire simple comme le "ou exclusif" (XOR). Pour résoudre ce type de problèmes plus complexe et non-linéaires, des perceptrons multi-couches (MLP) ont été

introduits (LeCun (1985), Rumelhart et McClelland (1986)), ainsi qu'un nouvel algorithme d'apprentissage basé sur la minimisation d'une fonction de coût. L'extension est naturelle, un MLP n'étant autre qu'un réseau de neurones formels. Dans la Figure 2.3, on présente le schéma d'un perceptron à une couche cachée, avec deux variables en entrée, deux unités cachées et une sortie linéaire.

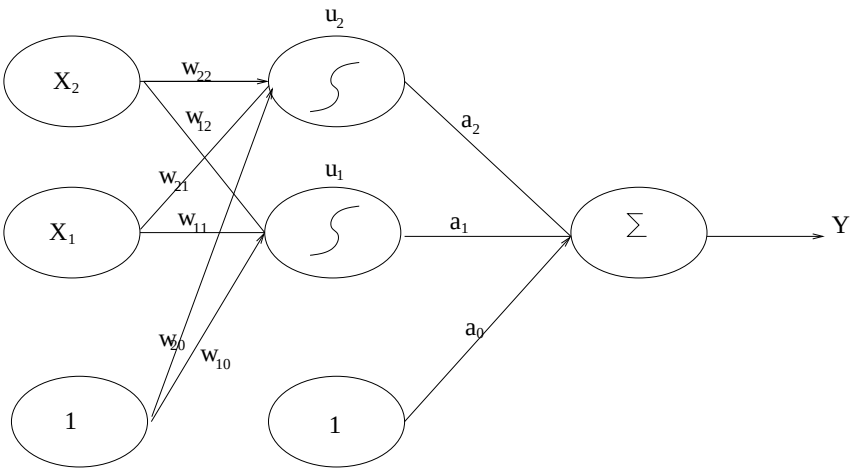

FIG. 2.3 – Schéma d'un MLP à une couche cachée et une sortie linéaire

Chaque unité de la couche cachée, ainsi que l'unité de sortie sont des neurones formels (appelés aussi perceptrons simples). De manière générale, un MLP se définit comme suit :

Définition 2.4 *: Si $X_1, ..., X_p$ sont les variables d'entrée et $Y_1, ..., Y_s$ les variables de sortie, un perceptron à une couche cachée est défini par la fonction*

$$(Y_i)_{i=1,...,s} = \left(\sum_{j=1}^{c} a_{ij} \phi \left(\sum_{k=1}^{p} w_{jk} X_k + w_{j0} \right) + a_{i0} \right)_{i=1,...,s}$$

où c représente le nombre d'unités cachées, $w_{j0}, w_{j1}, ..., w_{jp}$ les poids de l'unité cachée u_j, $j = 1, ..., s$, $a_{i0}, a_{i1}, ..., a_{ic}$ les poids de l'unité de sortie i, $i = 1, ..., s$, et $\phi : \mathbb{R} \to \mathbb{R}$ est une fonction d'activation sigmoïde. L'extension de la définition à un nombre quelconque de couches cachées est immédiate.

Les propriétés fondamentales du perceptron multi-couches l'ont rendu très populaire dans la modélisation non-linéaire. On pourrait citer celle d'approximateur universel, valable pour toute fonction continue à support compact, l'identifiabilité du modèle et la facilité des calculs numériques par des équations de propagation et rétro-propagation.

Combiner des MLP et des chaînes de Markov cachées pour créer des modèles hybrides HMC-MLP *(Hidden Markov Chain - Multilayer Perceptron)* pourrait donc fournir des résultats intéressants dans la modélisation des séries fortement non-linéaires. Une définition formelle d'un HMC-MLP est la suivante :

Définition 2.5 *: Un modèle hybride* ***HMC-MLP(p)*** *est un processus bivarié $(X_t, Y_t)_{t \in \mathbb{Z}}$ tel que*

- (X_t) *est une chaîne de Markov homogène à valeurs dans un espace d'états fini* $E = \{e_1, ..., e_N\}$, $N \in \mathbb{N}^\star$, *caractérisée par sa matrice de transition* $A = (a_{ij})_{i,j=1,...,N}$, *où*

$$a_{ij} = \mathbb{P}(X_{t+1} = e_j \mid X_t = e_i)$$

- $Y_t = F_{X_t}(Y_{t-1}, ..., Y_{t-p}) + \sigma_{X_t}\varepsilon_t$, *où, pour tout* $i = 1, ..., N$, $F_i(Y_{t-1}, ..., Y_{t-p})$ *est un perceptron multi-couches à valeurs dans* $\mathbb{R}$ *paramétré par le vecteur des poids* W_i, $\sigma_i \in \{\sigma_1, ..., \sigma_N\} \in \left(\mathbb{R}_+^\star\right)^N$ *et* (ε_t) *est une suite i.i.d. gaussienne centrée et réduite.*

Ces modèles, ainsi que leurs propriétés statistiques de consistence et normalité asymptotique de l'estimateur de maximum de vraisemblance ont été étudiés dans Rynkiewicz (2000). Il les a aussi appliqués à deux séries de données : la série laser complète de "Santa Fe time series prediction and analysis competition" (Weigend et Gershenfeld (1993)) et les pics de pollution à l'ozone en région parisienne. Pour la première série, les résultats sont comparables avec ceux déjà existants en littérature (Weigend et alii.(1995)), mais ils améliorent la modélisation par une réduction importante du nombre des paramètres et une bonne segmentation de la série. Sur la deuxième série, il y a de bonnes valeurs pour les erreurs de prévision et ce malgré le fait qu'il y a des données manquantes. Dans le chapitre 3 de cette thèse, on utilise ces modèles pour étudier et éventuellement prévoir des renversements de tendances sur une série quantifiant les chocs sur les marchés financiers.

2.4 Modéles autorégressifs localement linéaires et à changements de régime indépendants

Cette dernière section traite d'un cas particulier de modèles autorégressifs à changements de régime. L'idée centrale de ce document étant la construction d'un estimateur consistent pour le nombre de régimes, le caractère markovien des changements est parfois une condition trop faible et on n'arrive pas toujours à déduire la convergence. D'où la nécessité, au Chapitre 5, de se restreindre au cas particulier des changements de régime indépendants.

Les modèles autorégressifs à changements de régime indépendants, appelés aussi mélanges de modèles autorégressifs (*"mixture autoregressive models"*), sont donc un cas particulier de la Définition 2.2. Dans ce cas, (X_t) est un processus i.i.d. à valeurs dans un espace d'états fini $E = \{e_1, ..., e_N\}$, $N \in \mathbb{N}^\star$. Leur définition a été formalisée par Wong et Li (2000) qui ont motivé ces modèles par leur propriétés de prendre en compte les distributions conditionnelles multi-modales et l'hétéroscédasticité d'une série chronologique. Cependant, leur article porte uniquement sur des conditions de stationnarité faible et le calcul des estimateurs du maximum de vraisemblance via un algorithme de type EM. La consistence et la normalité asymptotique des estimateurs se déduisent immédiatement du cadre général markovien traité dans la section 2.2.

En ce qui concerne le problème de sélection du nombre de régimes, Wong et Li (2000) ont montré de manière empirique le bon comportement des critères pénalisés (BIC, AIC), mais sans en démontrer la convergence. Dans le Chapitre 5, on démontre la consistence faible de l'estimateur du maximum de vraisemblance pénalisée pour le nombre de régimes, en utilisant des techniques de processus empiriques pour des données dépendantes.

Chapitre 3

Caractérisation des crises financières à l'aide des modèles hybrides HMC-MLP

De violentes turbulences des cours secouent souvent les marchés financiers et un indice de crise - appelé IMS, *Index of Market Shocks* (voir Maillet et Michel (2002)) - a été récemment introduit pour tenter de quantifier les turbulences des marchés se produisant à l'occasion de ces crises financières.

L'objet de ce chapitre est de fournir, à l'aide de plusieurs modélisations, une description du comportement de l'indicateur IMS, calculé sur le marché international, en essayant de caractériser la présence de régimes dans la série. Des modèles plus simples sont étudiés au début (autorégressifs linéaires, perceptrons multi-couches), la justification de l'utilisation des modèles hybrides étant ici plutôt empirique. Une justification théorique du point de vue de la sélection de modèles sera présentée aux chapitres suivants.

Les modèles à changements de régime ont déjà été employés dans la modélisation de la volatilité conditionnelle des rentabilités boursières (Hamilton (1994)), ainsi que des crises bancaires et financières du siècle dernier (Coe (2002)). Par ailleurs, les modèles hybrides combinant une chaîne de Markov cachée à des perceptrons multicouches ont été utilisés dans l'étude des pics de pollution (Rynkiewicz (2000)), partageant *a priori* quelques similitudes avec les phénomènes de crise observés sur les marchés financiers.

Ce travail a été présenté dans plusieurs colloques pour des séries de l'indicateur IMS calculées sur les marchés américain et français (ACSEG 2003, ESANN 2004, AEA 2004) et s'est concrétisé dans un article paru dans la Revue d'Economie Politique (Maillet, Olteanu et Rynkiewicz (2004)).

3.1 Motivation de l'étude

Les marchés financiers sont le lieu de brusques mouvements qui font régulièrement l'actualité de la presse économique. Qu'il s'agisse de la Grande Dépression des années 30, du

krach d'octobre 1987, de la chute brutale des cours suite aux attaques terroristes du 11 septembre 2001, de la lente dégradation de forte ampleur du second semestre 2002, les marchés financiers sont passés par de fortes turbulences depuis leur création.

Néanmoins, caractériser la forme de ces turbulences - violentes et de courte durée *versus* étalées dans le temps et de durée plus longue - et fournir une mesure objective de leurs conséquences financières et réelles restent un exercice délicat qui se résume souvent à une analyse *post-mortem* au final peu utile. Récemment, un indicateur de crise appelé IMS, *Index of Market Shocks* (Maillet et Michel (2002)) a été proposé pour traduire de manière simple la turbulence des marchés. Fondé sur l'approche originale de Zumbach et alii. (2000a et 2000b) sur le marché des changes, l'indicateur IMS est une mesure de risque multi-échelle calculée à partir de la volatilité historique qui s'inspire d'une analogie avec l'échelle de Richter, couramment utilisée pour quantifier l'importance des tremblements de terre. Cependant, il reste à définir précisément une crise financière au sens de la mesure introduite. Maillet et Michel (2002) ont proposé une définition d'une crise financière : l'indicateur IMS a franchi à la hausse la valeur arbitraire de 3, ce qui correspond approximativement à la valeur du dernier décile de l'indicateur calculé sur les échantillons français et américains étudiés. Cette valeur seuil a un caractère *ad hoc* certain et une caractérisation endogène plus fine reste à fournir. C'est le principal objet du présent travail.

Les travaux pionniers d'Engle (1982) et de Bollerslev (1986) sur la volatilité conditionnelle hétéroscédastique ont été suivi par de nombreux développements. Un courant de la littérature sur les modèles ARCH, GARCH fait apparaître l'existence de régimes ou de seuils (SWARCH (Hamilton et Susmel (1994), TARCH (Rabemananjara et Zakoian (1993)) avec des transitions ou des franchissements plus ou moins doux (ANST-GARCH (Anderson et alii. (1999)). La volatilité - mesure traditionnelle du risque financier - se prête ainsi bien à ce type de modélisation. Par ailleurs, le paroxysme du risque est atteint pendant les crises financières et c'est aussi un modèle à changements de régime qui a été choisi (Coe (2002)) pour étudier l'histoire mouvementée du monde bancaire américain au siècle dernier. Enfin, certaines crises financières ont de particulier le fait que leurs amplitudes sont hors de proportion par rapport aux événements rythmant la vie quotidienne des marchés. Une telle démesure s'accommode donc facilement d'une représentation à plusieurs régimes : un régime exceptionnel qui régit les événements extrêmes et un régime (ou plusieurs) qui correspond(ent) aux évolutions "normales " des marchés. On retrouve ici *a priori* quelques similitudes avec d'autres phénomènes de "pics", comme la pollution à l'ozone (Rynkiewicz (2000)), qui nous ont encouragés à utiliser des modèles à changements de régimes fortement non-linéaires de type hybride HMC-MLP définis dans la section 2.3.

La distinction endogène des épisodes de turbulence des marchés est importante à plusieurs titres. Une caractérisation fine des crises et du risque systématique permettrait un meilleur contrôle de leurs évolutions : les acteurs du marché et les autorités de tutelle sachant mieux distinguer les crises, les mesures préventives et curatoires pourraient être mieux proportionnées. Une mesure plus fine du risque systématique peut aussi jouer un rôle pour une meilleure optimisation des choix micro-économiques d'allocation stratégique et tactique. Si la relation de long terme entre risque et rendement est effectivement inversée pendant les périodes de crise financière, les conclusions à en tirer en termes d'optimisation de portefeuille ne sont pas négligeables. Enfin, plusieurs auteurs, par exemple Sornette et alii (2003), ont mis en avant la distinction entre crises endogènes et crises exogènes.

Une caractérisation par états des crises financières peut servir de première étape à une classification - endogène *versus* exogène - des chocs de marché. Il est en effet intuitif de considérer que certaines crises sont purement imprévisibles (par exemple, la crise de septembre 2001 liée aux attaques terroristes), tandis que d'autres épisodes de turbulences pourraient avoir été précédés de signes larvés de krach imminent.

Le reste de ce chapitre est organisé comme suit : on continue par une description de l'indicateur IMS (définition et propriétés) et des données qui seront utilisées. Ensuite, la série est modélisée par des modèles simples (autorégressifs et perceptrons multi-couches) pour finir avec des modèles plus complexes comme les HMC-MLP. L'ajustement aux données, les qualités en prévision et l'étude de la séparation des états - crises *versus* accalmies - sont ensuite présentés. Le "meilleur" modèle est choisi de manière empirique, plusieurs critères étant proposés. La conclusion résume les résultats obtenus et propose de nouvelles pistes de recherche concernant l'étude de la variabilité du risque systématique.

3.2 Définition et propriétés de l'indicateur de crise IMS

Les mesures de risque traditionnelles comme la volatilité historique et le spectre multifractal sont souvent inadaptées pour quantifier les turbulences des marchés financiers puisqu'elles ne tiennent pas compte de la variabilité de l'étendue et de la durée des périodes de crise. De plus, le niveau de la volatilité, par exemple, dépend de la fréquence des données et de l'horizon choisi (variations intra-journalières, variation entre deux cours de clôture, tendances à long terme). Il fallait donc définir un indicateur des chocs sur les marchés qui prenne en compte l'hétérogénéité des intervenants, en intégrant leur différents horizons d'observation et de décision.

Une première approche a été proposée par Zumbach et alii. (2000a et 2000b) sur le marché des changes. En faisant une analogie entre les fluctuations sur le marché et les mouvements continus de la surface terrestre et en remarquant les similarités entre l'énérgie mécanique et la volatilité des prix, ils ont introduit une échelle logarithmique de l'agrégation pondérée de mesures de volatilité intégrant les différents horizons des agents économiques. Cette définition est assez intuitive si on garde l'analogie avec la géophysique : l'intensité d'un tremblement de terre suit une loi exponentielle et l'échelle de Richter n'est autre qu'une transformation logarithmique de l'énergie totale dissipée.

Maillet et Michel (2003) ont généralisé cette définition aux marchés financiers. Leur indice des chocs de marchés, appelé IMS (*Index of Market Shocks*) est une transformation logarithmique des volatilités des cours. Pour calculer une observation de l'IMS, on a besoin des variances des variations des cours de l'échelle la plus fine à la plus grossière. Par exemple, si on s'intéresse à l'indice CAC40 en haute fréquence, on pourrait calculer la valeur journalière de l'IMS comme une transformation pondérée d'une trentaine de variances allant des observations faites toutes les dix minutes jusqu'au cours de clôture. Les avantages de cette définition sont, d'une part, l'absence de référence à la valeur précédente de l'indicateur et, d'autre part, l'absence du biais de saisonnalité. Il reste cependant un inconvénient majeur qui est l'impossibilité de calculer cet indice en temps réel.

Schématiquement, l'IMS se construit de la manière suivante :

– On récupère les valeurs du cours d'un indice de marché (CAC40, Dow Jones etc.) ou d'un titre en haute, moyenne ou basse fréquence et on calcule les volatilités instantanées associées à un pas de temps sur plusieurs fréquences ou échelles d'observation. Par exemple, si on s'intéresse à la série IMS journalière associée au CAC40 pour lequel on dispose des valeurs enregistrées toutes les dix minutes, on calcule des volatilités journalières pour des fréquences qui peuvent être de dix minutes, une demi-heure, une heure jusqu'à la fréquence maximale qui est la valeur entre deux cours de clôture. A l'instant t (par exemple, au jour t), la volatilité associée à chaque fréquence s'écrit :

$$\sigma_t\left(\Delta t_\sigma, \Delta t_r\right) = \left[\eta \sum_{i=1}^{I} \left(p_{t_i} - p_{t_i+\Delta t_r}\right)^2\right]^{\frac{1}{2}}$$

où Δt_σ représente le pas de temps (par exemple un jour), I est le nombre d'observations entre les instants $t-1$ et t ou encore le nombre d'observations dans l'intervalle Δt_σ et $\Delta t_r = \frac{\Delta t_\sigma}{I}$ est l'incrément de temps entrant dans le calcul de chaque volatilité ou encore la fréquence (par exemple dix minutes ou une heure). Les autres notations sont $\eta = \frac{N}{\Delta t_\sigma}$, un facteur qui sert à standardiser les volatilités avec N le nombre de jours ouvrés de l'année et p_{t_i} la séquence des prix avec $t_i = t - i\Delta t_r$.

A chaque instant t, on pose $o_t = \left(\sigma_t^1, ..., \sigma_t^L\right)$, où L est le nombre de fréquences considérées. Si chaque instant est une journée, o_t représente le vecteur des volatilités journalières calculées sur des fréquences allant de dix minutes à une journée.

– Les composantes du vecteur o_t étant corrélées avec un corrélation qui dépend des fréquences choisies, on les transforme en les rendant indépendantes :

1. puisque les variances sont log-normales (voir Andersen et alii.(2001)), on applique une transformation logarithmique au vecteur o_t
2. à partir des valeurs transformées, des facteurs non-corrélés sont extraits par une analyse en composantes principales et puisqu'il s'agit de variables normales, les facteurs sont de plus indépendants. Le nouveau vecteur s'écrit donc

$$\widehat{o}_t = \left(\lambda_{t,1}, ..., \lambda_{t,K}\right)$$

où K est le nombre de facteurs retenus après l'ACP.

– Avec des variables qui sont maintenant normales et indépendantes, on peut définir, en analogie avec l'échelle de Richter, l'indicateur IMS à l'instant t :

$$IMS_t = -\sum_{k=1}^{K} \left[\omega_k log_2\left(1 - F\left(\lambda_{t,k}\right)\right)\right]$$

où ω_k est la contribution du k-ième facteur à la variance totale et $F\left(\cdot\right)$ est la fonction de repartition de la loi normale.

Maillet et Michel (2003) ont montré que l'IMS est relativement robuste aux hypothèses utilisées pour sa construction, notamment la log-normalité des variances, ce qui autorise des comparaisons historiques et des mises en perspective des différentes crises.

3.3 Description des données

L'étude des récentes crises financières repose dans ce texte sur le calcul d'un IMS journalier (fin de journée), fondé sur l'indice englobant des pays développés et des pays émergents MSCI (*Morgan Stanley Capital International*) World Equities Index. On dispose d'une base de données en haute fréquence comprenant les valeurs enregistrées toutes les dix minutes sur la période 25 juin 1998 - 25 juillet 2002 et en appliquant l'algorithme de calcul de l'IMS décrit à la section précédente (on a calculé 19 variances journalières, la plus haute fréquence étant de dix minutes, la plus basse la différences entre les cours de clôture journaliers). On obtient finallement une série de 1066 observations de l'IMS.

Cette période est intéressante à étudier puisqu'on y retrouve les dernières crises ayant marqué l'économie mondiale comme la crise russe en août 1998, brésilienne en janvier 1999, la bulle informatique en 2000-2001 et les attaques terroristes du 11 septembre 2001. Les valeurs de l'indice MSCI et de l'indicateur de crises IMS sont représentées dans la Figure 3.1. On reconnaît la période correspondant à la bulle informatique, ainsi que les valeurs élevées de l'IMS pendant les autres crises.

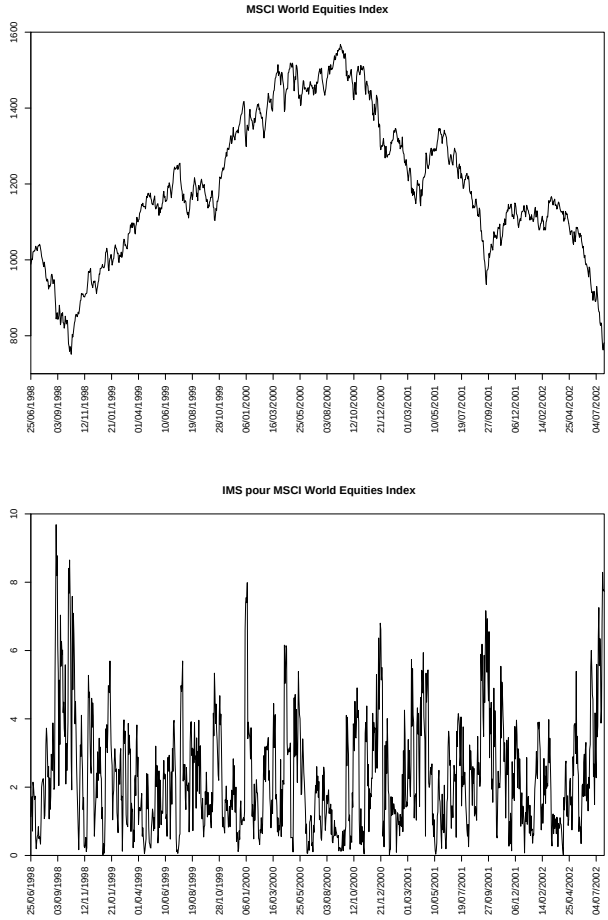

FIG. 3.1 – MSCI World Equity Index et l'IMS correspondant

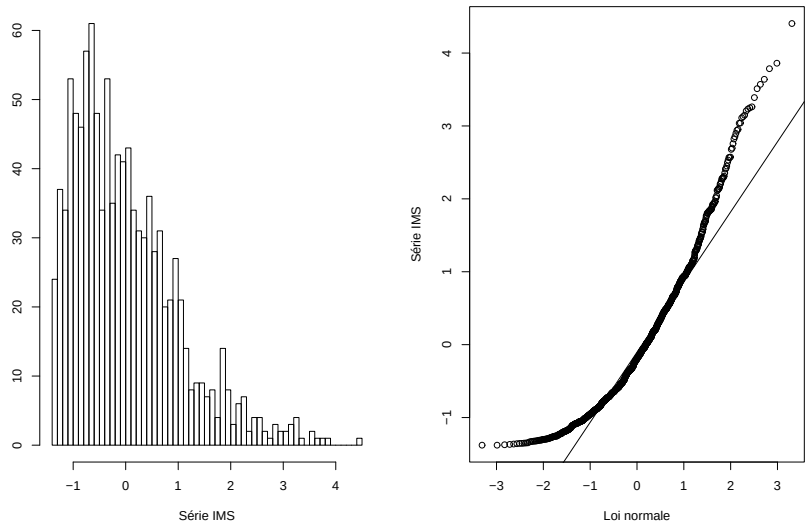

FIG. 3.2 – Ecart par rapport à une loi gaussienne de la série IMS

Le caractère gaussien des données est étudié de manière empirique en traçant l'histogramme et le diagramme quantile-quantile de la serié IMS (Fig. 3.2). L'hypothèse de normalité est donc rejetée à cause de l'écart important par rapport à une loi gaussienne.

La série IMS est fortement autocorrélée et cela est dû à la redondance des informations prises en compte dans son calcul, mais aussi à une autorégressivité généralement constatée des variances conditionnelles. Cette hypothèse est renforcée par la représentation des versions empiriques de la fonction d'autocorrélation et de la fonction d'autocorrélation partielle dans la Figure 3.3 qui montrent un possible caractère non-linéaire de la série.

Si $x_1, ..., x_n$ sont des observations d'une série chronologique, on rappelle que la fonction d'autocorrélation empirique est définie par :

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)},\ -n < h < n$$

où $\widehat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}\left(x_{i+|h|} - \overline{x}\right)(x_i - \overline{x})$ et $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ représentent la covariance et, respectivement, la moyenne empirique.

L'autocorrélation partielle empirique est définie par :

$$\begin{aligned}\widehat{\alpha}(0) &= 1\\ \widehat{\alpha}(h) &= \widehat{\phi}_{hh},\ h \geq 1\end{aligned}$$

où $\widehat{\phi}_{hh}$ est la dernière composante de $\widehat{\phi}_h = \widehat{\Gamma}_h^{-1}\widehat{\gamma}_h$ avec $\widehat{\Gamma}_h = [\widehat{\gamma}(i-j)]_{i,j=1}^{h}$ et $\widehat{\gamma}_h = (\widehat{\gamma}(1), ..., \widehat{\gamma}(h))^T$.

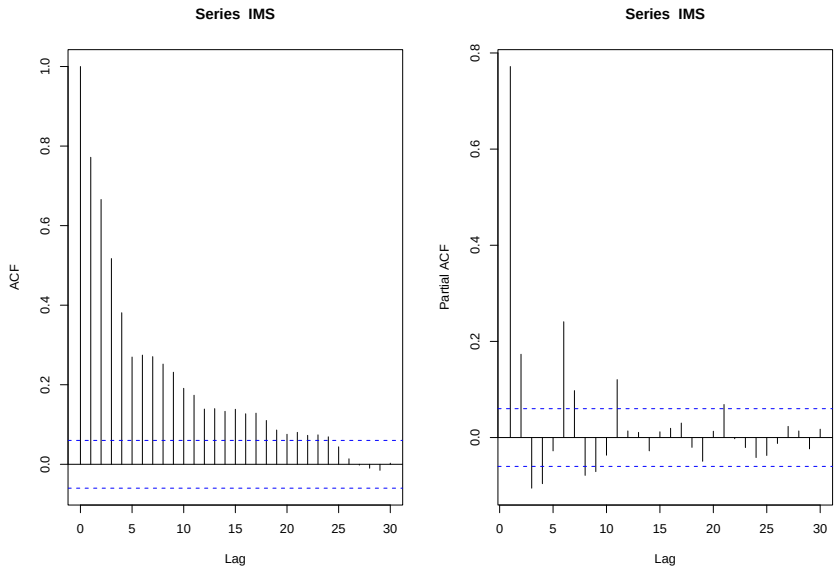

FIG. 3.3 – IMS - autocorrélation et autocorrélation partielle empiriques pour $h = 0, ..., 30$

3.4 Etude préliminaire

Nous travaillons sur la série centrée et réduite. Cette précaution est plutôt d'ordre numérique, pour éviter les éventuels phénomènes de saturation dans les fonctions d'activation des perceptrons. Pour comparer les modèles étudiés, on a choisi de diviser la série de données en un ensemble d'estimation ou d'apprentissage (800 observations) et un ensemble de validation (266 observations).

Dans un premier temps, on a estimé un modèle linéaire, pour vérifier ensuite si un modèle non-linéaire de type MLP ou HMC-MLP améliore les résultats.

3.4.1 Modèle linéaire

Le modèle linéaire optimal au sens du critère BIC (Schwarz (1978)) est un AR(11) dont les résultats sont reportés dans le tableau 3.1 et les Figures 3.4 et 3.5 :

$$IMS_t = \sum_{i=1}^{11} a_i IMS_{t-i} + \varepsilon_t$$

Le tableau 3.1 contient les valeurs estimées des coefficients et leurs écarts-type calculés sur l'ensemble d'apprentissage.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
coeff.	0.65	0.33	-0.05	-0.16	-0.25	0.22	0.17	0.05	-0.08	-0.14	0.14
ecart-type	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03

TAB. 3.1 – Coefficients du modèle AR(11)

Dans la Figure 3.4 on représente les caractéristiques des résidus. Les deux premiers graphiques contiennent les fonctions d'autocorrélation des résidus et des valeurs absolues des résidus, les deux derniers sont le diagramme quantile-quantile de la fonction de répartition empirique par rapport à une loi normale et l'histogramme des résidus. On obtient donc des résidus non-corrélés et non-gaussiens, avec une queue épaisse à gauche.

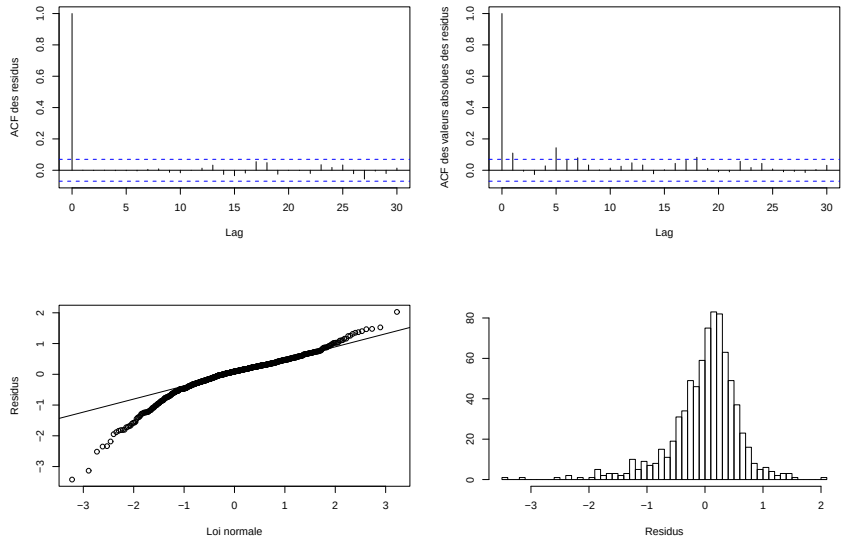

FIG. 3.4 – Les propriétés de corrélation et normalité pour les résidus du modèle AR(11)

Les erreurs sur les deux séries (apprentissage et validation) sont reportées dans le tableau 3.2. Elles ont été calculées comme les sommes des résidus au carré divisées par le nombre d'observations.

Données	MSE
Ensemble d'apprentissage (800 obs.)	0.344
Ensemble de validation (266 obs.)	0.349

TAB. 3.2 – Erreurs du modèle AR(11) sur les deux bases

Les résultats en prévision sur l'ensemble de validation sont présentés dans la Figure 3.5, les vraies valeurs étant en continu, les sorties du modèle linéaire en pointillé. On remarque

sur ce graphique que les valeurs moyennes sont plutôt sur-estimées, tandis que les pics sont le plus souvent sous-évalués.

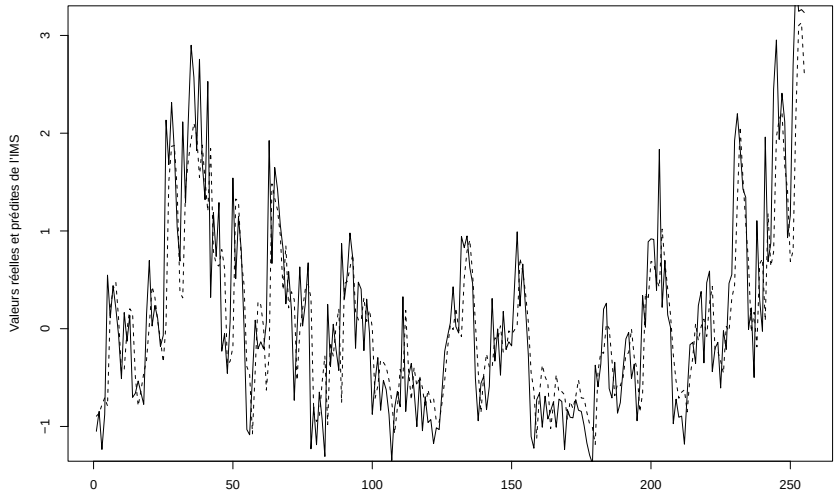

FIG. 3.5 – Prévision sur l'ensemble de validation pour le modèle AR(11)

3.4.2 Modèle MLP

Le modèle non-linéaire de type MLP a été estimé en utilisant le logiciel REGRESS développé par Rynkiewicz (2000), avec le nombre de retards fixé égal à celui qui a été obtenu dans le cas linéaire par le critère BIC, 11. Pour réduire le temps de calcul, on a choisi d'étudier uniquement les perceptrons à une couche cachée et avec un nombre d'unités cachées variant de 1 à 10. Le meilleur modèle au sens du critère BIC (voir la Figure 3.6) retient une unité cachée, les coefficients des retards 3, 8 et 9 et la constante étant nuls.

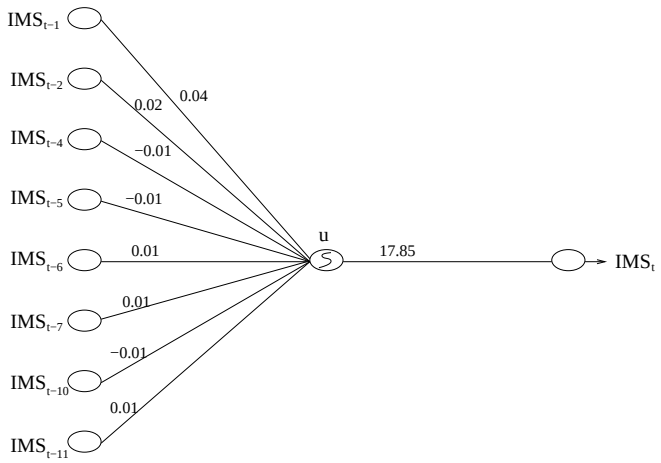

FIG. 3.6 – Modèle MLP à 11 retards

Notons que les résidus (Figure 3.7) sont non-corrélés et non-gaussiens, avec une queue épaisse à gauche.

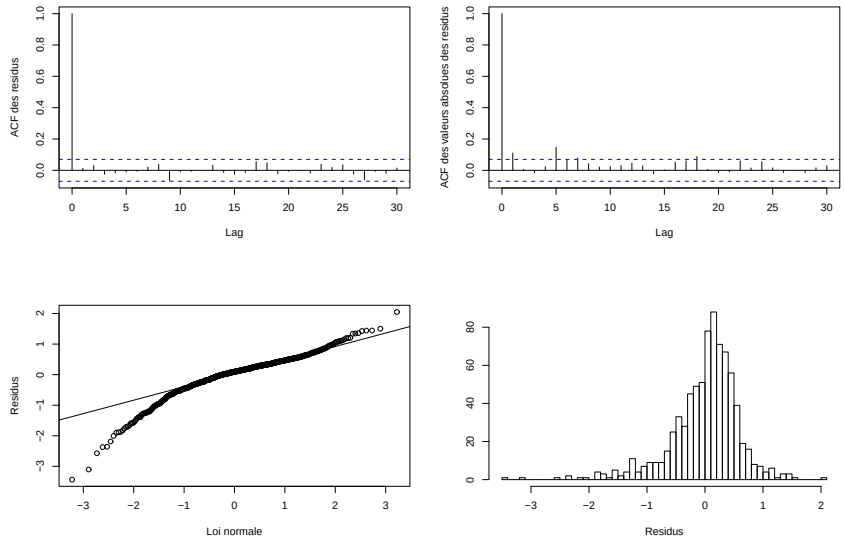

FIG. 3.7 – Les propriétés de corrélation et normalité pour les résidus du modèle MLP à 11 retards

En termes d'erreur quadratique, même si le modèle linéaire semble mieux adapté aux données (0.344 *versus* 0.347 sur l'ensemble d'apprentissage), le MLP est légèrement meilleur en prévision (0.349 *versus* 0.346 sur l'ensemble de validation).

La performance en prévision sur l'ensemble de validation est présentée dans la Figure 3.8, les vraies valeurs étant en continu, les sorties du perceptron en pointillé. On remarque

Données	MSE
Ensemble d'apprentissage (800 obs.)	0.346712
Ensemble de validation (266 obs.)	0.345865

TAB. 3.3 – Erreurs du modèle MLP à 11 retards sur les deux bases

sur ce graphique que les valeurs moyennes sont plutôt bien prédites, tandis que les pics sont le plus souvent sous-évalués. Le perceptron fait donc légèrement mieux que le modèle linéaire, mais n'arrive toujours pas à capter les pics, d'où l'intérêt d'introduire un modèle à plusieurs régimes, dont un spécifique aux pics.

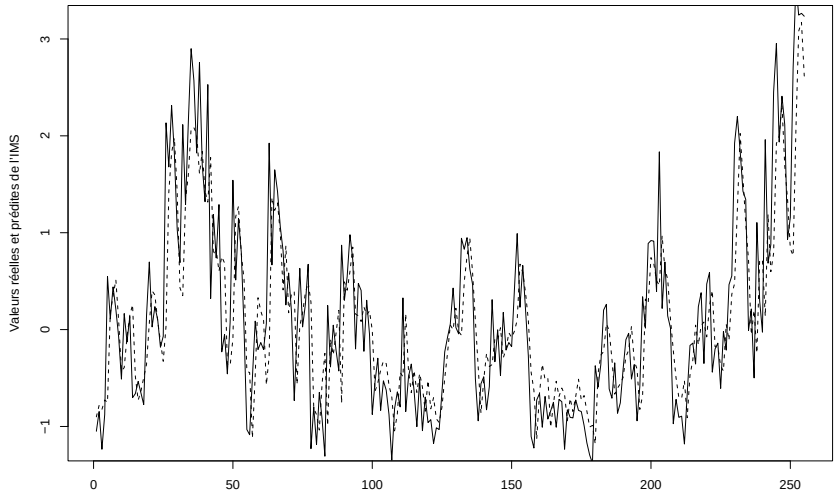

FIG. 3.8 – Prévision sur l'ensemble de validation pour le MLP à 11 retards

3.5 Estimation par un modèle hybride HMC-MLP

Une étude visuelle de la série ne permet pas de décider si le vrai modèle est à plusieurs régimes. La décision reste cependant empirique. Après l'utilisation des modèles plus simples (linéaire, MLP) qui sous-estiment les pics, on s'attend à ce qu'un modèle à deux ou plusieurs régimes arrive à capter ce phénomène.

3.5.1 Le modèle et la procédure d'estimation

Dans la suite de ce chapitre, on utilise des modèles à deux ou à trois régimes. Pour chaque architecture étudiée, le nombre de régimes est donc supposé connu et fixé égal à deux ou à trois. Estimer des modèles pour lesquels le nombre de régimes n'est pas fixé "a priori"

mène à des problèmes de non-identifiabilité quand le vrai nombre de régimes est surestimé. Ce problème sera abordé largement aux chapitres suivants et on proposera un critère d'identification du modèle consistant.

En ce qui concerne les autres paramètres, le nombre de retards a été fixé égal à celui sélectionné par le critère BIC dans le cas linéaire (onze), tandis que le nombre de couches cachées varie entre 0 (modèle linéaire) et 1 (MLP) et le nombre d'unités cachées sur une couche peut aller de 1 à 3. Le choix de fixer le nombre de retards a été dicté principalement par le temps de calcul. En effet, même avec ces contraintes, pour un modèle à trois régimes il y a vingt architectures différentes à estimer et chaque estimation dure plusieurs heures sur un processeur 512 MHz.

Les paramètres de chaque configuration possible sont estimés en maximisant la log- vraisemblance via un algorithme de type EM (Rabiner (1989)). Pour éviter les problèmes de mauvaises initialisations, ainsi que les minimas locaux, on a fait plusieurs initialisations différentes. Dans le cas de la série IMS, 10 initialisations ont suffi pour avoir des résultats satisfaisants.

A défaut d'une justification théorique de la consistance d'un critère d'information de type BIC pour ces modèles, le critère d'identification a été choisi de manière empirique. Plusieurs méthodologies ont été envisagées, comme par exemple choisir le modèle qui donne la meilleure prévision au sens des moindres carrés ou encore choisir le modèle qui maximise la trace de la matrice de transition de la chaîne de Markov cachée. Les meilleurs résultats au sens de la modélisation des pics ont été obtenus avec le deuxième critère et sont ceux qui sont présentés dans la suite. La justification intuitive pour le choix de ce critère sur la trace est le fait qu'on obtient ainsi le modèle avec la meilleure segmentation empirique de la série.

3.5.2 Les résultats

3.5.2.1 Modèle hybride HMC-MLP à deux régimes

Le modèle HMC-MLP à deux régimes (Figure 3.9) comporte une composante linéaire si on est dans le premier régime ou un MLP à deux unités cachées si on est dans le deuxième.

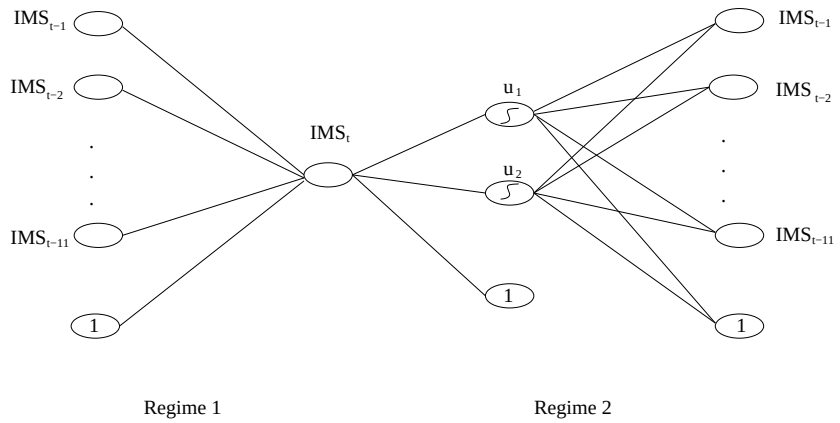

FIG. 3.9 – Modèle HMC-MLP à deux régimes

La matrice de transition estimée $\hat{A}$ pour la chaîne de Markov cachée est :

$$\hat{A} = \begin{pmatrix} 0.899 & 0.125 \\ 0.101 & 0.875 \end{pmatrix}$$

Les résidus (Figure 3.10) sont non-corrélés et non-gaussiens avec un queue épaisse à gauche.

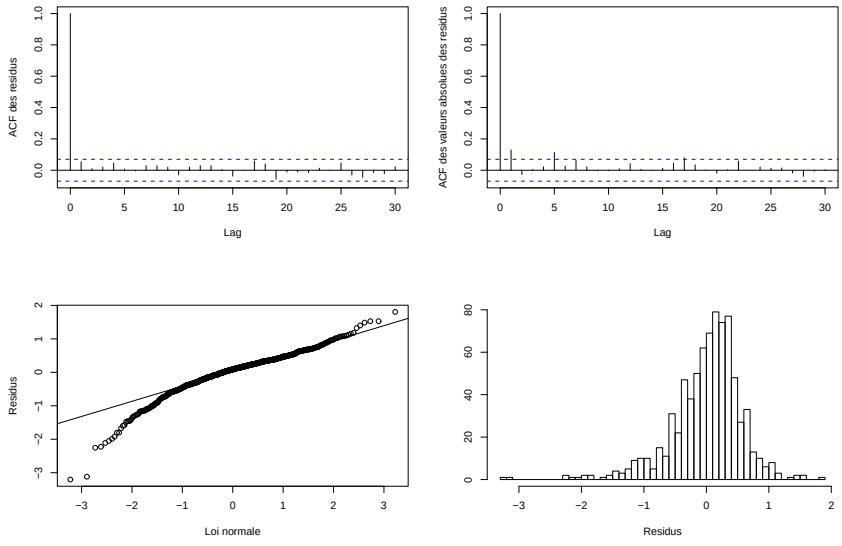

FIG. 3.10 – Les propriétés de corrélation et normalité pour les résidus du modèle HMC-MLP à deux régimes

Les erreurs quadratiques sur les deux séries (apprentissage et validation) sont résumées dans le Tableau 3.4. Le modèle semble mieux ajusté aux données que le modèle linéaire

Données	MSE
Ensemble d'apprentissage (800 obs.)	0.316263
Ensemble de validation (266 obs.)	0.361532

TAB. 3.4 – Erreurs du modèle HMC-MLP à deux régimes

ou le perceptron, mais sur l'ensemble de validation l'erreur est beaucoup plus importante, suggérant un phénomène de sur-apprentissage. Cela correspond en général à la situation où le modèle est "trop bien" ajusté aux données d'apprentissage et perd en conséquence son pouvoir prédictif. La cause usuelle d'un sur-apprentissage est un mauvais dimensionnement. Ici, cela pourrait se produire à cause du nombre de retards considéré. En effet, il n'y a pas de raison pour que le bon nombre de retards dans le cas linéaire soit toujours le bon dans le cas d'un modèle à changements de régimes, mais là on est malheureusement confronté à l'absence de critères théorique de sélection, ainsi qu'aux limitations imposées par le temps de calcul.

La Figure 3.11 montre les prévisions du modèle à deux régimes sur la série de validation (valeurs réelles en continu, valeurs prédites en pointillé), ainsi que les probabilités conditionnelles du premier régime en sachant toutes les données (filtrage "bilatéral"). On remarque que la probabilité d'être dans le premier régime est grande quand les valeurs de l'indicateur IMS sont faibles. Il semblerait donc qu'il s'agit d'un régime linéaire pour les périodes plus calmes et d'un régime non-linéaire, contrôlé par un perceptron, pour les moments de forte volatilité sur le marché.

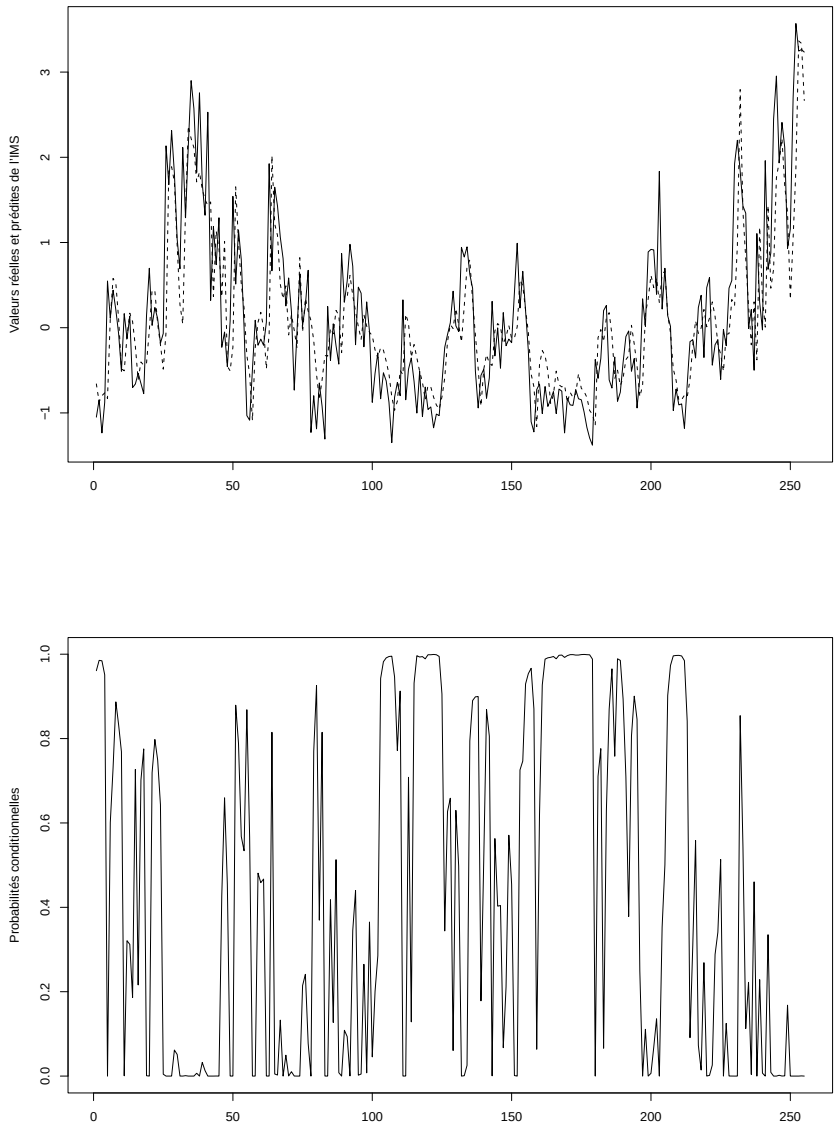

FIG. 3.11 – Prévision et probabilités conditionnelles du premier régime sur l'ensemble de validation pour le modèle HMC-MLP à deux régimes

3.5.2.2 Modèle hybride HMC-MLP à trois régimes

Le modèle HMC-MLP à trois régimes (Figure 3.12) est composé de trois perceptrons, le premier a une unité cachée et les deux autres deux unités cachées.

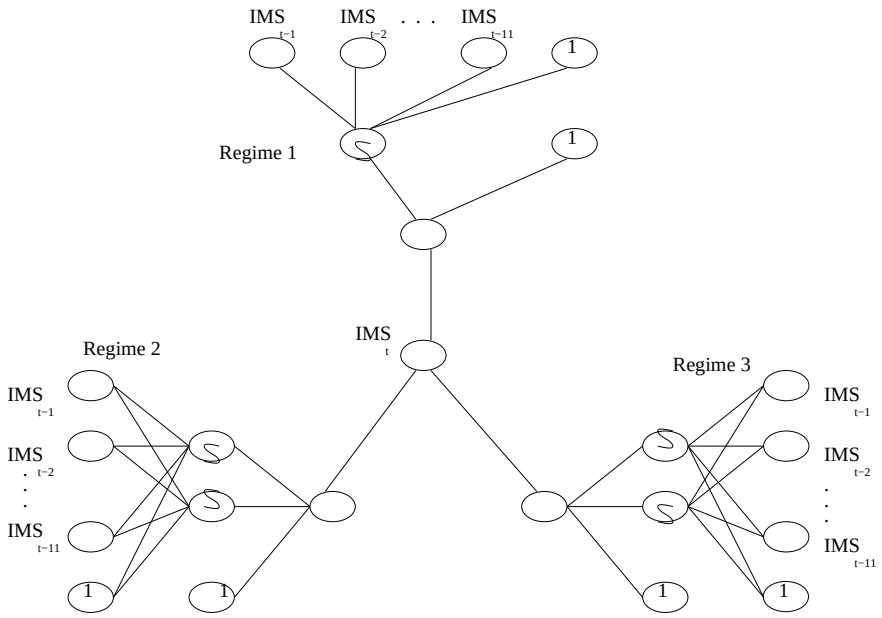

FIG. 3.12 – Modèle HMC-MLP à trois régimes

La matrice de transition estimée $\hat{A}$ pour la chaîne de Markov cachée est :

$$\hat{A} = \begin{pmatrix} 0.626 & 0.049 & 0.190 \\ 0.068 & 0.826 & 0.147 \\ 0.306 & 0.125 & 0.663 \end{pmatrix}$$

Les résidus (Figure 3.13) sont eux aussi non-corrélés et non-gaussiens, avec une queue épaisse à gauche.

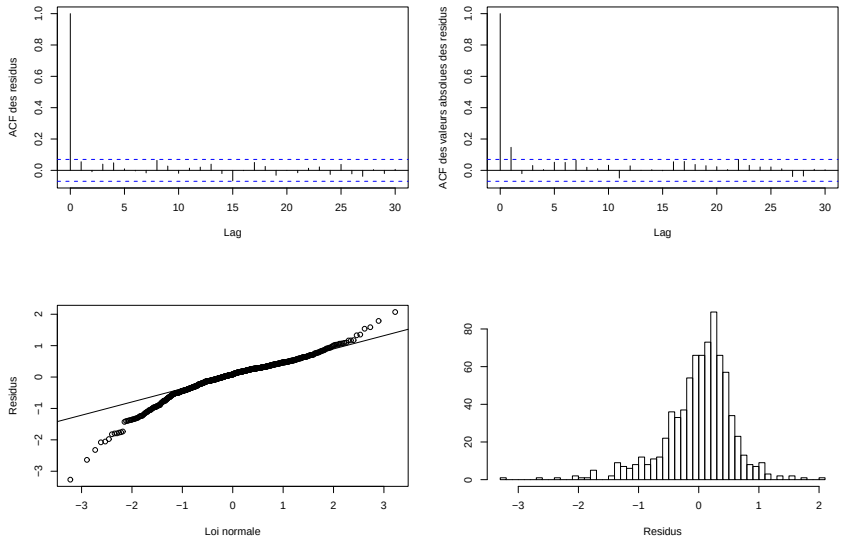

FIG. 3.13 – Les propriétés de corrélation et normalité pour les résidus du modèle HMC-MLP à trois régimes

Au niveau des erreurs quadratiques résumées dans le Tableau 3.5, ce modèle est le mieux ajusté aux données, mais se comporte très mal en prévision. L'erreur élevée sur l'ensemble de validation, ainsi que le nombre important de paramètres du modèle montrent qu'on est clairement dans un cas de sur-apprentissage.

Données	MSE
Ensemble d'apprentissage (800 obs.)	0.314859
Ensemble de validation (266 obs.)	0.379248

TAB. 3.5 – Erreurs du modèle HMC-MLP à trois régimes

La figure 3.14 présente les prévisions du modèle HMC-MLP à trois régimes sur la série de validation (valeurs réelles en continu, valeurs prédites en pointillé), ainsi que les probabilités conditionnelles des deux premiers régimes (le premier en continu, le deuxième en pointillé) en sachant toutes les données. Ici, la corrélation des régimes avec le niveau de l'indicateur IMS est difficile à établir et à interpréter à l'oeil nu. On pourrait faire des analyses supplémentaires pour identifier à quelles valeurs de l'IMS correspond le premier régime, le deuxième etc., mais le but de l'étude étant d'arriver à une caractérisation des périodes de crise, le modèle à deux régimes semble être suffisant.

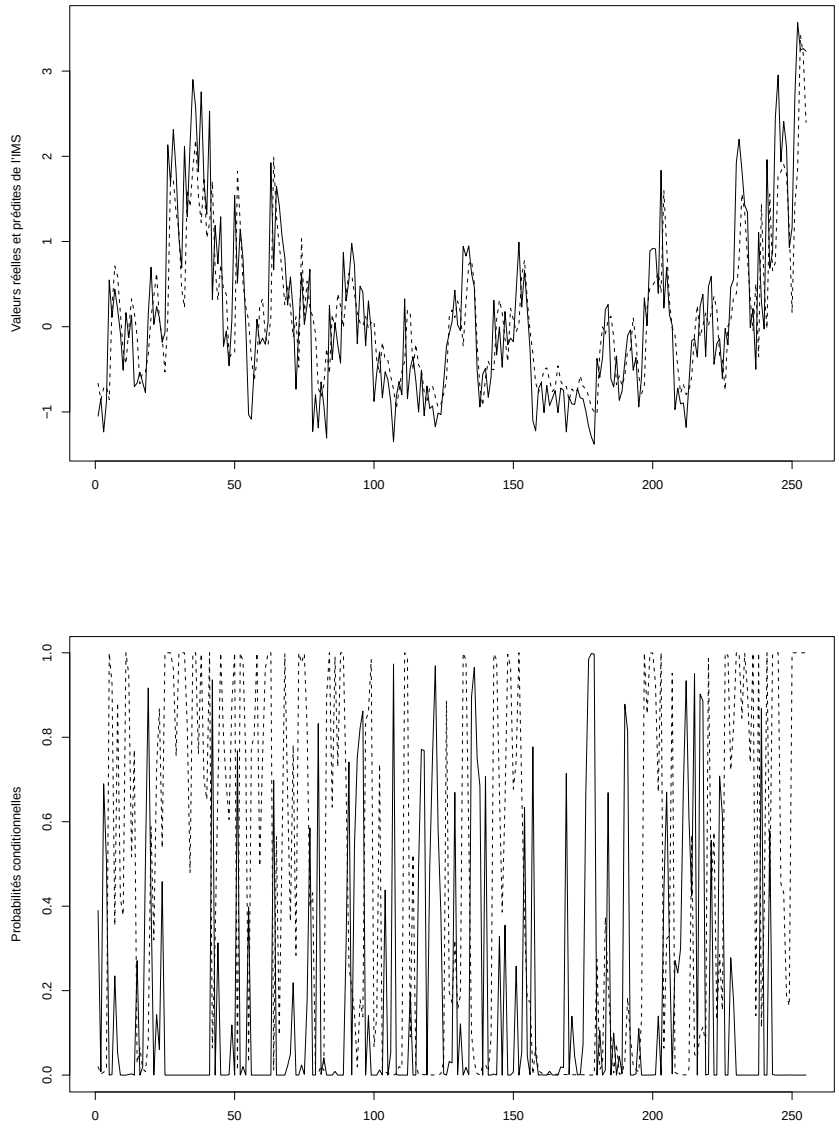

FIG. 3.14 – Prévision et probabilités conditionnelles des deux premiers régimes sur l'ensemble de validation pour le modèle HMC-MLP à trois régimes

3.5.3 Caractérisation des crises financières à l'aide du modèle HMC-MLP à deux régimes

L'utilisation des modèles hybrides n'a pas permis d'améliorer l'erreur de prédiction de la série IMS. Ceci est assez intuitif compte tenu du caractère imprévisible de certaines crises financières. L'étude des probabilités conditionnelles est toutefois intéressante puisque l'on obtient une séparation entre deux états relatifs à deux comportements différentes de l'indice et du marché. Le second régime semble correspondre aux périodes de crises et de fortes turbulences, tandis que le premier s'adapte aux périodes moins agitées.

Quand on s'intéresse plus précisément à certaines périodes troublées de l'histoire récente des marchés financiers, cette intuition semble nettement se confirmer. A titre d'exemple, on représente dans la Figure 3.15 l'evolution de l'indicateur IMS (en pointillé) et les probabilités conditionnelles du second régime (en continu) pendant les crises russe (août-septembre 1998), brésilienne (janvier 1999) et celle qui a suivi les attaques terroristes du 11 septembre (septembre-octobre 2001). On constate sur les graphiques que les turbulences durent plusieurs semaines ininterrompues, pendant lesquelles les probabilités conditionnelles restent très proches de la valeur limite unitaire.

3.6 Conclusion

L'étude de l'indicateur IMS à l'aide d'un modèle hybride HMC-MLP ne permet pas d'aboutir à de meilleurs résultats en termes de prévision qu'un modèle linéaire ou non-linéaire plus simple de type MLP. Ce résultat est conforme à l'hypothèse d'efficience faible des marchés financiers et correspond à la difficulté de prévoir les variations des cours, spécifiquement quand de grands événements surviennent sur les marchés. Néanmoins, cette modélisation fournit une séparation des états du marché représentée par deux régimes, l'un caractérisant les crises et l'autre les périodes stables. Quand on représente l'evolution des probabilités conditionnelles pour le second régime, on est capable d'extraire précisément les dates correspondant aux décisions ou aux événements générant de larges fluctuations de marché, et plus généralement d'évaluer la durée totale d'une crise. Cette modélisation permet aussi d'obtenir une caractérisation des crises sans recourir à des éléments *ad hoc* d'une définition arbitraire. C'est ici en effet le modèle, et plus précisément l'état de la chaîne de Markov cachée à un instant donné, qui autorise la classification de tel ou tel événement de marché comme crise majeure sur les marchés financiers.

D'un point de vue économétrique, la prochaine étape de ce travail aurait été une classification des périodes de marchés, en faisant intervenir l'indicateur IMS, les probabilités conditionnelles des différents régimes et les rentabilités associées aux différents régimes de turbulence, dans le but de caractériser plus en avant la nature des crises financières.

Cependant, on a choisi de s'intéresser au problème de non-identifiabilité et du choix du nombre de régimes. La convergence des critères de vraisemblance pénalisée a déjà été démontrée pour les mélanges de lois et pour les chaînes de Markov cachées (Keribin (2000), Gassiat (2002)). Au Chapitre 5, on étend ces résultats au cadre des mélanges de processus autorégressifs, mais on rappelle d'abord au Chapitre 4 les principales notions dont on a besoin plus tard. Puisqu'on travaille dans un cadre autorégressif et donc dépendant, on passe

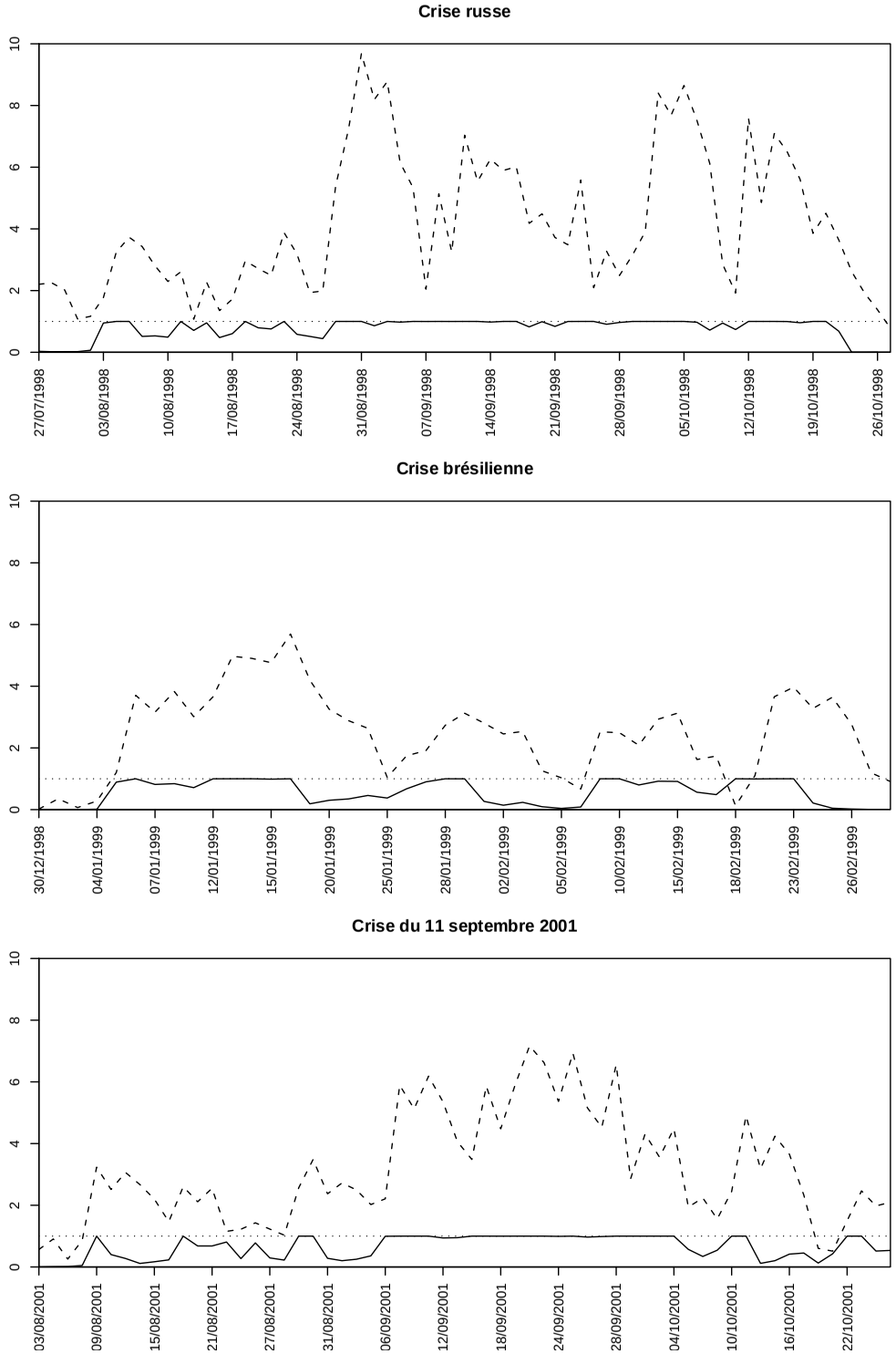

FIG. 3.15 – Les principales crises caractérisées par un modèle HMC-MLP à deux régimes : la série IMS est en pointillé, la probabilité conditionnelle d'être dans le premier régime en ligne continue

en revue rapidement les différents types de dépendance et les théorèmes limites associés, ainsi que les conditions de stationnarité et mélangeance pour les modèles autorégressifs à changements de régime.

Chapitre 4

Stationnarité et propriétés de dépendance faible des modèles autorégressifs à changements de régime markoviens

Très utilisés dans la modélisation des séries temporelles depuis les années 90, les modèles autorégressifs à changements de régime n'ont été étudiés du point de vue théorique que récemment. Les applications nombreuses en économétrie manquaient d'un support théorique concernant leurs propriétés de régularité et stabilité comme la stationnarité, l'ergodicité ou la dépendance faible. De plus, ces propriétés sont nécessaires pour pouvoir étudier le nombre de régimes, construire un estimateur de ce nombre et établir les propriétés asymptotiques.

Ce chapitre sera donc dédié à un état de l'art non-exaustif sur les conditions suffisantes de stationnarité, ergodicité et dépendance faible des modèles autorégressifs à changements de régime markoviens, suivis de quelques résultats asymptotiques développés dans ce cadre et qui sont employés par la suite.

4.1 Conditions suffisantes de stationnarité pour les modèles autorégressifs à changements de régime markoviens

Dans l'introduction, on a mentionné la propriété de "stationnarité" des séries temporelles, en la définissant de manière très intuitive comme une propriété de régularité du processus. Plus précisément, si $(Y_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires définies sur un espace de probabilité $(\Omega, \mathcal{K}, \mathbb{P})$, on définit les notions suivantes :

Définition 4.1 *: La série $(Y_t)_{t\in\mathbb{Z}}$ est stationnaire au sens stricte si la loi du vecteur $(Y_t, Y_{t+1}, ..., Y_{t+h})$ est indépendante de t, pour tout $h \in \mathbb{Z}$.*

Définition 4.2 *: La série $(Y_t)_{t\in\mathbb{Z}}$ est stationnaire au sens faible si $\mathbb{E}Y_t$ et $Cov\,(Y_t, Y_{t+h})$ existent, sont finies et indépendantes de t, pour tout $h \in \mathbb{Z}$.*

Tout processus strictement stationnaire qui admet des moments d'ordre deux finis est donc faiblement stationnaire.

Voyons maintenant sous quelles hypothèses les modèles autorégressifs à changements de régime markoviens vérifient les conditions de stationnarité stricte.

Dans cette section, on se place dans un cadre général qui comprend les modèles localement linéaires et les modèles hybrides avec des perceptrons multicouches décrits au Chapitre 2. On considère donc un processus bivarié $(X_t, Y_t)_{t\in\mathbb{Z}}$ qui vérifie le modèle suivant, noté dans la suite ***HMC-NAR(p)*** :

$$Y_t = F_{X_t}\left(Y_{t-1}, ..., Y_{t-p}\right) + \varepsilon_t\left(X_t\right) \text{ où,}$$

- (X_t) est une chaîne de Markov homogène, apériodique et irréductible, à valeurs dans un espace d'états fini $E = \{e_1, ..., e_N\}$, où $N \in \mathbb{N}^*$ et caractérisée par sa matrice de transition $A = (\pi_{ij})_{i,j=1,...,N}$ et sa mesure invariante $\mu = (\mu_i)_{i=1,...,N}$
- $\{F_{e_i},\ i = 1, ..., N\}$ est une famille de fonctions nonlinéaires autorégressives
- pour tout $i = 1, ..., N$, $(\varepsilon_t(e_i))_{t\in\mathbb{Z}}$ est un bruit i.i.d., les suites $(\varepsilon_t(e_i))_{1\leq i\leq N}$ étant indépendantes.

Introduisons maintenant les hypothèses suivantes, dans le cas particulier $p = 1$:

Hypothèse (HS)

(HS1) Les fonctions de régression F_{e_i} sont continues et sous-linéaires, c'est-à-dire que pour tout $i = 1, ..., N$ il existe $(a_i, b_i) \in \mathbb{R}_+^2$ tels que $|F_{e_i}(y)| \leq a_i |y| + b_i$, $y \in \mathbb{R}$

(HS2) Pour tout $i = 1, ..., N$, le bruit $\varepsilon_t(e_i)$ admet une densité strictement positive par rapport à la mesure de Lebesgue et il existe $s \geq 1$ tel que $\mathbb{E}|\varepsilon_1(e_i)|^s < \infty$

(HS3) Le rayon spectral de la matrice Q_s vérifie la rélation $\rho(Q_s) < 1$, où

$$Q_s = \begin{pmatrix} (a_1)^s \pi_{11} & \cdots & (a_N)^s \pi_{1N} \\ \vdots & \ddots & \vdots \\ (a_1)^s \pi_{N1} & \cdots & (a_N)^s \pi_{NN} \end{pmatrix}$$

Pour le modèle ***HMC-NAR(1)*** et sous l'hypothèse **(HS)**, Yao et Attali (2000) ont montré l'existence et l'unicité d'une solution strictement stationnaire et géométriquement ergodique, ainsi que l'existence des moments jusqu'à l'ordre s pour la mesure stationnaire de Y_t.

On rapelle qu'une chaîne de Markov (Y_t) est dite géométriquement ergodique s'il existe $0 < \eta < 1$ et $a(y) > 0$ une fonction borélienne qui vérifie $\int a(y)\,\mu(dy) < \infty$ tels que

$$\forall\, n \geq 1,\ \left\|\mathbb{P}^n(y, \cdot) - \mu\right\|_{Var} \leq a(y)\,\eta^n$$

où $\|\cdot\|_{Var}$ représente la variation totale.

On peut donc remarquer que tout modèle à changements de régime markoviens et fonctions de régression linéaires qui est stationnaire dans chaque régime, c'est-à-dire pour lequel

$|a_i| < 1$, $i = 1, ..., N$, est globalement stationnaire puisque dans ce cas la matrice Q_s devient sous-stochastique et l'hypothèse **(HS3)** est immédiatement vérifiée.

Dans le cas particulier des modèles à changements de régime indépendants, les probabilités de transition deviennent

$$\pi_{ij} = \mathbb{P}(X_t = j \mid X_{t-1} = i) = \mathbb{P}(X_t = j) = \pi_j$$

et l'hypothèse **(HS3)** s'écrit :

$$\sum_{i=1}^{N} \pi_i (a_i)^s < 1$$

La généralisation au cas $p \in \mathbb{N}^\star$ est immédiate en reécrivant le modèle. Si on considère maintenant le processus multivarié de dimension p :

$$Z_t = Y_{t-p+1}^t = (Y_t, Y_{t-1}, ..., Y_{t-p+1})^T,$$

(Z_t) peut s'exprimer comme un processus vectoriel autorégressif d'ordre 1, sous la forme

$$Z_t = \begin{pmatrix} F_{X_t}(Y_{t-1}, ..., Y_{t-p}) \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix} + \begin{pmatrix} \varepsilon_t(X_t) \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \widetilde{F}_{X_t}(Z_{t-1}) + \widetilde{\varepsilon}_t(X_t)$$

Le nouveau processus $(X_t, Z_t)_{t\in\mathbb{Z}}$ est une chaîne de Markov et admet une unique solution strictement stationnaire et géométriquement ergodique avec des moments d'ordre $s \geq 1$ sous les hypothèses **(HS')** identiques aux hypothèses **(HS)** sauf en ce qui concerne la première condition qui est remplacée par

(HS1') les fonctions de régression $\widetilde{F}_{e_i}$ sont continues et sous-linéaires, c'est-à-dire que pour tout $i = 1, ..., N$ il existe $(a_i, b_i) \in \mathbb{R}_+^2$ tels que $\|F_{e_i}(y_1, ..., y_p)\| \leq a_i \|(y_1, ..., y_p)\| + b_i$, $(y_1, ..., y_p) \in \mathbb{R}^p$, où $\|\cdot\|$ est une norme quelconque.

4.2 Rappels sur les propriétés de dépendance faible des processus autorégressifs

La stationnarité au sens strict n'est pas suffisante pour obtenir des résultats asymptotiques dans le cadre des séries temporelles et donc des suites de variables aléatoires dépendantes. Il est nécessaire d'introduire une mesure de la dépendance et de la quantifier. L'idée est de se rapprocher d'une indépendance asymptotique en considérant des suites de variables aléatoires dont le passé et le futur lointains sont indépendants. Cette indépendance asymptotique appelée aussi "dépendance faible" est mesurée par des coefficients de mélange et les suites ayant cette propriété sont appelées mélangeantes ou faiblement dépendantes.

4.2.1 Coefficients de mélange et processus absolument réguliers

Dans ce document, on utilise les coefficients de mélange introduits par Volkonskii et Rozanov (1959) et appelés coefficients de β-mélange ou absolument réguliers. Ce choix est justifié d'un côté par la large variété d'exemples qui vérifient l'hypothèse de régularité absolue et d'un autre côté par l'existence de résultats asymptotiques dans ce cadre, notamment un théorème limite central uniforme. Pour un état de l'art sur les autres coefficients définis dans la littérature et leurs propriétés, se rapporter à Doukhan (1995) et Bradley (2005).

Dans la suite, soit $(\Omega, \mathcal{K}, \mathbb{P})$ un espace de probabilité. Un coefficient de β-mélange sera défini en général comme une mesure de dépendance entre deux tribus :

Définition 4.3 *: Soient $\mathcal{A}, \mathcal{B} \subset \mathcal{K}$ deux σ-algèbres. Le coefficient de β-mélange entre $\mathcal{A}$ et $\mathcal{B}$ est :*

$$\beta(\mathcal{A}, \mathcal{B}) = sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\,\mathbb{P}(B_j)|$$

où le sup est considéré sur toutes les partitions finies de Ω, $\{A_1, ..., A_I\}$ $\mathcal{A}$-mesurable et $\{B_1, ..., B_J\}$ $\mathcal{B}$-mesurable.

On considère maintenant $(Y_t)_{t\in\mathbb{Z}}$, suite de variables aléatoires définies sur $(\Omega, \mathcal{K}, \mathbb{P})$ non nécessairement stationnaire. En notant

$$\mathcal{F}_k^l = \sigma(Y_t,\ k \leq t \leq l)$$

la tribu engendrée par les variables Y_t, $k \leq t \leq l$, on peut introduire la définition suivante :

Définition 4.4 *: Pour chaque $n \geq 1$, on définit les coefficients de β-mélange de la suite $(Y_t)_{t\in\mathbb{Z}}$ par*

$$\beta(n) = sup_{k\in\mathbb{Z}} \beta\left(\mathcal{F}_{-\infty}^k, \mathcal{F}_{k+n}^{\infty}\right)$$

La suite $(Y_t)_{t\in\mathbb{Z}}$ est absolument régulière ou β-mélangeante si $\beta(n) \to 0$ quand $n \to \infty$.

Si, de plus, la suite $(Y_t)_{t\in\mathbb{Z}}$ est strictement stationnaire, alors l'écriture des coefficients de β-mélange se reduit à :

$$\beta(n) = \beta\left(\mathcal{F}_{-\infty}^0, \mathcal{F}_n^{\infty}\right)$$

On a vu à la section précédente que les processus autorégressifs à changements de régime markoviens et fonctions de régression nonlinéaires pouvaient s'écrire comme des chaînes de Markov. Pour cela, on s'intéresse dans la suite aux conditions qui assurent la propriété de régularité absolue aux chaînes de Markov à espace d'états continu.

Soit $(Y_t)_{t\in\mathbb{Z}}$ une chaîne de Markov à valeurs dans un espace continu et non nécessairement stationnaire. Alors, par la propriété de Markov, les coefficients de β-mélange s'écrivent :

$$\beta(n) = sup_{k\in\mathbb{Z}}\beta(\sigma(Y_k), \sigma(Y_{k+n}))$$

Si de plus $(Y_t)_{t\in\mathbb{Z}}$ est strictement stationnaire, on a :

$$\beta(n) = \beta(\sigma(Y_0), \sigma(Y_n))$$

Avec ces définitions, on peut maintenant citer le résultat suivant (Doukhan (1995)) :

Théorème 4.1 *: Soit $(Y_t)_{t\in\mathbb{Z}}$ une chaîne de Markov strictement stationnaire et soit μ sa mesure invariante. Si Y_t est aussi géométriquement ergodique, alors Y_t est β-mélangeante et il existe $0 < \eta < 1$ tel que $\beta(n) = \mathcal{O}(\eta^n)$.*

Remarquons donc que les modèles autorégressifs à changements de régime markoviens qui vérifient les hypothèses **(HS')** de la section précédente seront en particulier β-mélangeants.

4.2.2 Résultats asymptotiques pour les processus absolument réguliers

Au chapitre suivant, on s'intéresse à un estimateur du nombre de régimes pour un modèle autorégressif de type HMC-NAR(p) et pour obtenir sa consistance, on a besoin d'utiliser des théorèmes limite classiques dans le cadre des processus stationnaires et faiblement dépendants.

Soit $(Y_t)_{t\in\mathbb{Z}}$ un processus aléatoire qui vérifie les hypothèses de stationnarité et β-mélangeance.

Sous la condition $\mathbb{E}|Y_t| < \infty$, la loi forte des grands nombres se déduit du théorème ergodique.

En ce qui concerne un théorème limite central, on travaille dans la suite avec des processus indexés par des classes de fonctions et on a donc besoin d'un résultat fonctionnel uniforme. Le résultat qui suit est tiré de Doukhan, Massart et Rio (1995). Pour l'énoncer, on va d'abord rappeler quelques notions sur les processus empiriques et l'entropie à crochets et ensuite on introduit une nouvelle norme fonctionnelle et l'espace associé.

4.2.2.1 Rappels sur les processus empiriques

Construits à partir d'un échantillon aléatoire, les processus empiriques sont des processus stochastiques indexés par des fonctions permettant d'étendre les théorèmes limite sur les variables aléatoires à des fonctions mesurables sur un espace arbitraire quelconque. On ne s'intéresse pas ici au cas plus général des événements non nécessairement mesurables, pour cela se rapporter au livre de Van der Vaart (2000) duquel cette sous-section, ainsi que la suivante sont inspirées.

Rappelons, par exemple, que si $X_1, ..., X_n$ est un tirage i.i.d. de variables aléatoires de fonction de répartition F, alors la fonction de répartition empirique est

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i\leq t\}},\ t\in\mathbb{R}$$

Plus généralement, si $X_1, ..., X_n$ est un tirage de fonctions mesurables indépendantes de loi P, définies sur un espace de probabilité $(\Omega, \mathcal{K}, \mathbb{P})$ et à valeurs dans un espace arbitraire $\mathcal{X}$, on définit la *mesure empirique*

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

où δ est la mesure de Dirac. Si $\mathcal{F}$ est une classe de fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$, on introduit le *processus empirique* indexé par $\mathcal{F}$:

$$\left\{ \mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \mid f \in \mathcal{F} \right\}$$

Avec cette définition, la fonction de répartition empirique peut être représentée comme un processus empirique pour lequel $\mathcal{X} = \mathbb{R}$ et $\mathcal{F} = \left\{ 1_{]-\infty,t]},\ t \in \mathbb{R} \right\}$.

On introduit maintenant deux classes de fonctions qui sont importantes dans la suite et qui permettent des généralisations en version uniforme de la loi des grands nombres et du théorème limite central :

Définition 4.5 *: Une classe de fonctions mesurables $\mathcal{F}$ est P-Glivenko-Cantelli si*

$$sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to 0 \; p.s.$$

où $Pf = \int_{\mathcal{X}} f(x) P(dx)$.

Définition 4.6 *: Une classe de fonctions mesurables $\mathcal{F}$ est P-Donsker si*

$$\mathbb{G}_n \rightsquigarrow G \; dans \; l^{\infty}(\mathcal{F})$$

où $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, $\rightsquigarrow$ est une notation pour la convergence faible, $l^{\infty}(\mathcal{F})$ est l'ensemble des fonctions bornées définies sur $\mathcal{F}$ à valeurs dans $\mathbb{R}$ et G est un processus gaussien indexé par $\mathcal{F}$, de moyenne nulle et covariance $\mathbb{E}(f(X) g(X)) - \mathbb{E}f(X) \mathbb{E}g(X)$, $\forall f, g \in \mathcal{F}$.

4.2.2.2 Rappels sur l'entropie à crochets

Une classe de fonctions sera Glivenko-Cantelli ou Donsker en fonction de sa complexité ou son entropie. Plusieurs conditions suffisantes ont été établies dans la littérature, suivant les différentes définitions de l'entropie. Ici on considère uniquement la notion d'entropie à crochets.

Soit $\mathcal{F}$ une classe de fonctions mesurables et pour tout $r \geq 1$, $L_r(P)$ est l'espace des fonctions mesurables vérifiant

$$\|g\|_{r,P} = \left[\int_{\mathcal{X}} |g(x)|^r dP(x) \right]^{1/r} < \infty.$$

Pour tout $\varepsilon > 0$, on définit un ε-crochet dans $L_r(P)$ par un couple de fonctions $l, u \in L_r(P)$ telles que $\mathbb{P}\{l(X) \leq u(X)\} = 1$ et $\|l - u\|_{r,P} < \varepsilon$.

Une fonction $f \in \mathcal{F}$ appartient au crochet $[l, u]$ si $\mathbb{P}\{l(X) \leq f(X) \leq u(X)\} = 1$.

L'ε-entropie à crochets de $\mathcal{F}$ par rapport à la norme de $L_r(P)$, notée $\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, L_r(P))$ se définit alors comme le logarithme du nombre minimum de ε-crochets nécessaires pour couvrir $\mathcal{F}$ noté $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_r(P))$.

Deux théorèmes établissent si $\mathcal{F}$ est Glivenko-Cantelli ou Donsker en fonction de

$\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, L_r(P))$ (Van der Vaart (2000)) :

Théorème 4.2 *: Soit $\mathcal{F}$ une classe de fonctions mesurables. Si on suppose que*

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$$

pour chaque $\varepsilon > 0$, alors $\mathcal{F}$ est P-Glivenko-Cantelli.

Théorème 4.3 *: Soit $\mathcal{F}$ une classe de fonctions mesurables. Si on suppose que*

$$\int_0^\infty \sqrt{\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon < \infty,$$

alors $\mathcal{F}$ est P-Donsker.

4.2.2.3 Un théorème limite central uniforme pour les séries stationnaires absolument régulières

Une généralisation de la propriété de Donsker pour des séries d'observations dépendantes et absolument régulières a été propsée par Doukhan, Massart et Rio (1995).

Tout d'abord, ils ont introduit une nouvelle norme notée $\|\cdot\|_{2,\beta}$ et construite à partir des coefficients de β-mélange. Pour la définir, introduisons quelques notations.

Soit $(Y_t)_{t\in\mathbb{Z}}$ un processus strictement stationnaire de loi marginale P, défini sur un espace de probabilité $(\Omega, \mathcal{K}, \mathbb{P})$ et à valeurs dans un espace polonais $\mathcal{X}$. On suppose que (Y_t) est absolument régulier et que la série des coefficients de β-mélange est sommable.

On définit la fonction de β-mélange de (Y_t) par l'extension cadlag de β_n, $\beta(u) = \beta_{[u]}$ avec la convention $\beta_0 = 1$.

Pour toute fonction non-décroissante ψ, on note par ψ^{-1} son inverse :

$$\psi^{-1}(u) = inf\{t,\ \psi(t) \leq u\}$$

Finalement, pour chaque fonction $f \in L_1(P)$, soit Q_f la fonction quantile de $|f(Y_0)|$ définie comme l'inverse de $t \to \mathbb{P}(|f(Y_0)| > t)$.

On considère maintenant l'espace $\mathcal{L}_{2,\beta}(P)$ qui comprend toutes les fonctions mesurables vérifiant :

$$\|f\|_{2,\beta} = \sqrt{\int_0^1 \beta^{-1}(u)\left[Q_f(u)\right]^2 du} < \infty$$

Le résultat de Doukhan, Massart et Rio (1995) est alors le suivant :

Théorème 4.4 *: Soit $(Y_t)_{t\in\mathbb{Z}}$ une suite strictement stationnaire et β-mélangeante de variables aléatoires (ou vecteurs aléatoires) de loi marginale P. On suppose que la série des coefficients de β-mélange est sommable et on considère une famille de fonctions $\mathcal{F}$ telle que $\mathcal{F} \subset \mathcal{L}_{2,\beta}(P)$. Si l'entropie à crochets de $\mathcal{F}$ par rapport à la norme $\|\cdot\|_{2,\beta}$ vérifie la condition d'intégrabilité*

$$\int_0^1 \sqrt{\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, L_{2,\beta}(P))} d\varepsilon < \infty$$

alors :

(i) La série $\sum_{t\in\mathbb{Z}} Cov\left(f(Y_0), f(Y_t)\right)$ est absolument convergente sur $\mathcal{F}$ vers une forme quadratique non-négative $\Gamma(f,f)$ et

$$\left(\Gamma(f,f)\right)^{1/2} = \|f\|_\Gamma \leq \|f\|_{2,\beta}$$

(ii) Il existe une suite $\left(Z^{(n)}\right)_{n>0}$ de processus gaussiens indexés par $\mathcal{F}$ avec la fonction de covariance Γ et des trajectoires p.s. uniformement continues telle que

$$sup_{f\in\mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(Y_i) - Pf\right) - Z^{(n)}(f) \right| \to_P 0$$

quand $n \to \infty$.

On a donc établi des conditions suffisantes de régularité et stabilité des modèles autorégressifs à changements de régime markoviens. Sous les hypothèses **(HS')**, on a vu que les modèles ***HMC-NAR(p)*** étaient strictement stationnaires, géométriquement ergodiques et en particulier géométriquement β-mélangeants. Sous ces mêmes hypothèses, on peut maintenant traiter au chapitre suivant le problème d'estimation du nombre de régimes.

Chapitre 5

Estimation du nombre d'états pour les modèles autorégressifs à changements de régime

Ce chapitre est consacré à l'étude d'un estimateur de maximum de vraisemblance pénalisée pour le nombre d'états d'un modèle autorégressif à changements de régime. On suppose dans la suite qu'un n-échantillon $(Y_1, ..., Y_n)$ d'une série temporelle à valeurs réelles $(Y_t)_{t\in\mathbb{Z}}$ a été observé. On suppose aussi que Y_t dépend de Y_{t-1} et d'un processus caché à valeurs discrètes (X_t), ce dernier pouvant être soit i.i.d. (modèles autorégressifs à changements de régime indépendants), soit une chaîne de Markov homogène (modèles autorégressifs à changements de régime markoviens). Le processus caché peut prendre un nombre fini de valeurs, chacune d'elle définissant ce qu'on appelle un "régime" ou un "état".

Les résultats théoriques concernant ces modèles et cités au Chapitre 2 ont été obtenus dans le cadre particulier d'un nombre de régimes supposé connu à l'avance. Si le nombre de régimes n'est pas fixé, tester quelle est sa vraie valeur soulève un problème de non-identifiabilité sous l'hypothèse alternative. Dans ce cas, sous l'hypothèse nulle, la matrice de l'information de Fisher est dégénérée, les conditions de régularité usuelles ne sont pas remplies et la théorie sur la convergence du rapport de vraisemblance vers une distribution du chi-deux ne s'applique pas.

Cependant, plusieurs méthodes pour estimer le cardinal de l'espace d'états de (X_t) ont été proposées dans le cadre particulier des mélanges : des techniques nonparamétriques dans Henna (1985), Roeder (1994) ou Izenman et Sommer (1998), des méthodes de moments dans Lindsay (1983) ou Dacunha-Castelle et Gassiat (1997) et des méthodes de maximum de vraisemblance pénalisée dans Leroux (1992b), Keribin (2000) et Gassiat (2002). D'ailleurs, Gassiat (2002) a montré la convergence en probabilité vers le vrai nombre de régimes pour un estimateur de vraisemblance pénalisée dans le cas où le processus $(Y_t, X_t)_{t\in\mathbb{Z}}$ est une chaîne de Markov cachée (voir la Définition 2.1). On va étendre ce résultat au cadre des modèles autorégressifs non-linéaires à changements de régime, avec des modifications étant dues aux problèmes de dépendance qui apparaissent entre les observations.

5.1 Le modèle

Dans un premier temps, on étudie les modèles autorégressifs à changements de régime *indépendants*. On verra plus tard si les résultats qu'on va montrer peuvent s'étendre au cadre plus général des changements de régime markoviens. Dans toute la suite, on considère que le nombre de retards est connu et égal à un, la généralisation à un nombre de retards fini quelconque étant immédiate.

Soit donc le processus $(X_t, Y_t)_{t\in\mathbb{Z}}$ qui vérifie le vrai modèle :

$$Y_t = F^0_{X_t}(Y_{t-1}) + \varepsilon_t(X_t) \qquad (5.1)$$

où

- (X_t) est une suite non-observée de variables aléatoires i.i.d., prenant des valeurs dans un espace fini $\{1, ..., p_0\}$ et ayant pour distribution la loi discrète $\pi^0 = \left(\pi^0_i\right)_{i=1,...,p_0}$;
- pour chaque $i = 1, ..., p_0$, $F^0_i(y)$ est une fonction nonlinéaire autorégressive qui dépend du paramètre $\theta^0_i \in E$, avec E sous-ensemble compact de $\mathbb{R}^d$;
- pour chaque $i = 1, ..., p_0$, $(\varepsilon_t(i))_{t\in\mathbb{Z}}$ est un bruit i.i.d. et les suites $(\varepsilon_t(i))_{1\leq i\leq N}$ sont indépendantes.

Introduisons maintenant les hypothèses suivantes :

Hypothèses (HS)

(HS1) Les fonctions de régression sont continues et sous-linéaires, c'est-à-dire que pour tout $i = 1, ..., p_0$, il existe $\left(a^0_i, b^0_i\right) \in \mathbb{R}^2_+$ tels que $\left|F^0_i(y)\right| \leq a^0_i |y| + b^0_i$, $y \in \mathbb{R}$

(HS2) Pour tout $i = 1, ..., p_0$, le bruit $(\varepsilon_t(i))$ admet une densité f^0_i strictement positive par rapport à la mesure de Lebesgue, qui dépend du paramètre θ^0_i et il existe $s \geq 1$ tel que $\mathbb{E}|\varepsilon_1(i)|^s < \infty$

(HS3) La loi π^0 des X_t vérifie la rélation :

$$\sum_{i=1}^{N} \pi^0_i \left(a^0_i\right)^s < 1$$

Avec les hypothèses **(HS)** et suivant le résultat de Yao et Attali (2000) décrit à la section 4.1, le modèle (5.1) admet une unique solution, strictement stationnaire et géométriquement ergodique, donc en particulier géométriquement β-mélangeante, dont la mesure stationnaire a des moments finis jusqu'à l'ordre s.

5.2 Construction du critère de vraisemblance pénalisée

On considère maintenant qu'on a observé un n-échantillon $\{y_1, ..., y_n\}$ de la série $(Y_t)_{t\in\mathbb{Z}}$. Pour chaque observation y_t, la vraie densité conditionnelle sachant la valeur précédente y_{t-1} et marginale par rapport à (X_t) s'écrit :

$$f\left(y_t \mid y_{t-1}\right) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_t - F_i^0\left(y_{t-1}\right)\right)$$

Remarquons que les changements de régime étant indépendants, (Y_t) est une chaîne de Markov et $f\left(y_t \mid y_{t-1}\right)$ n'est autre que sa densité de transition. Dans ce cas, il s'agit donc d'un processus markovien à transition de type mélange de populations.

Le but étant de construire un estimateur consistent du nombre d'états p_0 du processus caché $(X_t)_{t\in\mathbb{Z}}$, on considère la famille de tous les mélanges de densités conditionnelles jusqu'à un nombre maximum de régimes $P > 0$. Soit donc

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p$$

$$\mathcal{G}_p = \left\{ g \mid \ g\left(y_2 \mid y_1\right) = \sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right),\ \pi_i \geq 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

où, pour chaque $i = 1, ..., p$,

- F_i est une fonction nonlinéaire autorégressive qui dépend du paramètre $\theta_i \in E$
- f_i est une densité strictement positive par rapport à la mesure de Lebesgue, qui dépend du paramètre θ_i .

On fait de plus une hypothèse sur la compacité de l'espace des paramètres :

(HC) Pour tout $i = 1, ..., p$, $(\pi_i, \theta_i) \in \Gamma \times E$, avec $\Gamma \times E \subset [0,1] \times \mathbb{R}^d$ ensemble compact.

Pour chaque $g \in \mathcal{G}_P$, on définit le nombre de régimes de g par :

$$p\left(g\right) = min\left\{p \in \{1, ..., P\},\ g \in \mathcal{G}_p\right\}$$

Avec cette notation, le vrai nombre de régimes est $p_0 = p\left(f\right)$.

On introduit maintenant $\widehat{p}$, l'estimateur de maximum de vraisemblance pénalisée pour le vrai nombre de régimes p_0 défini comme l'argument qui maximise

$$T_n\left(p\right) = sup_{g\in\mathcal{G}_p} l_n\left(g\right) - a_n\left(p\right) \qquad (5.2)$$

pour $p \in \{1, ..., P\}$ et où

$$l_n\left(g\right) = \sum_{t=2}^{n} log\, g\left(y_t \mid y_{t-1}\right)$$

est la log-vraisemblance marginale par rapport à $X_1, ..., X_n$ et $a_n\left(p\right)$ est un terme de pénalité qui reste à définir.

Nous prouvons maintenant dans la section 5.3 que $\widehat{p}$ est un estimateur consistent de p_0.

5.3 Consistance de l'estimateur du nombre de régimes dans le cadre des changements de régime indépendants

Avant d'énoncer le résultat principal, on a besoin pour le démontrer de l'inégalité suivante, qui est une extension au cas multivarié du résultat de Gassiat (2002) et dont la preuve qui est identique est donc omise :

Proposition 5.1 *: Soit $\mathcal{G} \subset \mathcal{G}_P$ une famille paramétrique de densités conditionnelles contenant le vrai modèle f. Pour chaque $g \in \mathcal{G}$, on considère la fonction score généralisé*

$$s_g(y_1, y_2) = \frac{\frac{g(y_2|y_1)}{f(y_2|y_1)} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}}$$

où μ est la mesure stationnaire de (Y_{t-1}, Y_t). Alors on a l'inégalité :

$$sup_{g \in \mathcal{G}} \left(l_n(g) - l_n(f) \right) \leq \frac{1}{2} sup_{g \in \mathcal{G}} \frac{\left(\sum_{t=2}^{n} s_g(y_{t-1}, y_t) \right)^2}{\sum_{t=2}^{n} (s_g)_-^2 (y_{t-1}, y_t)}$$

avec la notation $(s_g)_-(y_{t-1}, y_t) = min(0,\ s_g(y_{t-1}, y_t))$.

A l'aide de cette inégalité, on peut énoncer et prouver le résultat suivant :

Théorème 5.1 *: On considère le modèle autorégressif $(X_t, Y_t)_{t \in \mathbb{Z}}$ à changements de régime indépendants défini par (5.1) et le critère de vraisemblance pénalisée $T_n(p)$ introduit dans (5.2). Faisons aussi les hypothèses suivantes :*

(A1) *$a_n(\cdot)$ est une fonction croissante de p, $a_n(p_1) - a_n(p_2) \to \infty$ quand $n \to \infty$ pour tout $p_1 > p_2$ et $\frac{a_n(p)}{n} \to 0$ quand $n \to \infty$ pour tout $p \in \mathbb{N}$*

(A2) *le modèle (X_t, Y_t) vérifie une propriété d'identifiabilité faible :*

$$\sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0 \left(y_t - F_i^0(y_{t-1}) \right) \Leftrightarrow \sum_{i=1}^{p} \pi_i \delta_{\theta_i} = \sum_{i=1}^{p_0} \pi_i^0 \delta_{\theta_i^0}$$

où, pour tout $x \in E$, δ_x est la mesure de Dirac concentrée en x.

(A3) *pour tout $i = 1, ..., p$, la paramétrisation $\theta_i \to f_i(y_2 - F_i(y_1))$ est continue pour chaque $(y_1, y_2) \in \mathbb{R}^2$ et il existe une fonction $m(y_1, y_2)$, intégrable par rapport à la mesure stationnaire de (Y_t, Y_{t-1}), telle que $\forall\, g \in \mathcal{G}$, $|log(g)| < m$*

(A4) *le modèle $(X_t, Y_t)_{t \in \mathbb{Z}}$ vérifie l'hypothèse de stationnarité **(HS)** et la famille de fonctions scores généralisés*

$$\mathcal{S} = \left\{ s_g,\ s_g(y_1, y_2) = \frac{\frac{g(y_2|y_1)}{f(y_2|y_1)} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}},\ g \in \mathcal{G},\ \left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \neq 0 \right\}$$

est dans $\mathcal{L}^2(\mu)$ *et pour chaque* $\varepsilon > 0$ *:*

$$\mathcal{H}_{[]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) = \mathcal{O}(|log\,\varepsilon|)$$

Alors, sous les hypothèses ***(A1)-(A4)*** *et* ***(HC)****,* $\hat{p} \to p_0$ *en probabilité quand* $n \to \infty$*, où* $\hat{p}$ *est l'estimateur qui maximise* $T_n(p)$ *pour* $p \in \{1, ..., P\}$*.*

Preuve *:*

La preuve de ce théorème est une légère adaptation de celui de Gassiat (2002).

On montre dans un premier temps que $\hat{p}$ ne surestime pas la vraie valeur p_0 :

$$\mathbb{P}(\hat{p} > p_0) \leq \sum_{p=p_0+1}^{P} \mathbb{P}(T_n(p) > T_n(p_0)) =$$

$$= \sum_{p=p_0+1}^{P} \mathbb{P}\left(sup_{g\in\mathcal{G}_p}(l_n(g) - l_n(f)) - a_n(p) > sup_{g\in\mathcal{G}_{p_0}}(l_n(g) - l_n(f)) - a_n(p_0)\right) \leq$$

$$\leq \sum_{p=p_0+1}^{P} \mathbb{P}\left(\frac{1}{2} sup_{g\in\mathcal{G}_p} \frac{\left(\sum_{t=2}^{n} s_g(Y_{t-1}, Y_t)\right)^2}{\sum_{t=2}^{n} (s_g)_-^2(Y_{t-1}, Y_t)} > a_n(p) - a_n(p_0)\right)$$

en appliquant l'inégalité de la Proposition 5.1 et en remarquant $sup_{g\in\mathcal{G}_{p_0}}(l_n(g) - l_n(f)) \geq 0$.

Sous l'hypothèse **(HS)**, il existe une unique solution strictement stationnaire (Y_t) qui est aussi géométriquement ergodique et en particulier geométriquement β-mélangeante. Alors, en remarquant que

$$\beta_n^{(Y_{t-1}, Y_t)} = \beta_{n-1}^{Y_t},$$

la suite bivariée (Y_{t-1}, Y_t) est aussi strictement stationnaire, ergodique et géométriquement β-mélangeante.

Doukhan, Massart et Rio (1995) ont donné des conditions suffisantes basées uniquement sur la norme L_2 pour que la condition sur l'entropie à crochets $\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, L_{2,\beta}(P))$ du Théorème 4.4 soit vérifiée. Ainsi, si les coefficients de β-mélange convergent géométriquement vers 0, il suffit que

$$\mathcal{H}_{[]}(\varepsilon, \mathcal{S}, L_2(P)) = \mathcal{O}(|log\,\varepsilon|)$$

pour qu'on ait la convergence uniforme du Théorème 4.4.

Le processus (Y_{t-1}, Y_t) étant géométriquement β-mélangeant, cette remarque et l'hypothèse **(A4)** permettent d'appliquer le théorème 4.4 et on aura que

$$\left\{ \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} s_g \left(Y_{t-1}, Y_t\right) \mid \ g \in \mathcal{G}_p \right\}$$

est une suite uniformément tendue et vérifie un théorème limite central uniforme, alors

$$sup_{g \in \mathcal{G}_p} \frac{1}{n-1} \left(\sum_{t=2}^{n} s_g \left(Y_{t-1}, Y_t\right) \right)^2 = \mathcal{O}_{\mathbb{P}}(1)$$

En même temps, $\mathcal{S} \subset \mathcal{L}^2(\mu)$ et en utilisant l'hypothèse **(A4)** une nouvelle fois, $\mathcal{S}_-^2 = \left\{ (s_g)_-^2 \, , \, g \in \mathcal{G}_p \right\}$ est Glivenko-Cantelli. Et finallement, puisque (Y_{t-1}, Y_t) est strictement stationnaire et ergodique, on aura la convergence uniforme en probabilité :

$$inf_{g \in \mathcal{G}_p} \frac{1}{n-1} \sum_{t=2}^{n} (s_g)_-^2 \left(Y_{t-1}, Y_t\right) \to inf_{g \in \mathcal{G}_p} \left\| (s_g)_- \right\|_{L^2(\mu)}^2$$

quand $n \to \infty$.

Pour finir la première partie, il reste à montrer que

$$inf_{g \in \mathcal{G}_p} \left\| (s_g)_- \right\|_{L^2(\mu)} > 0$$

Si on suppose, au contraire, que $inf_{g \in \mathcal{G}_p} \left\| (s_g)_- \right\|_{L^2(\mu)} = 0$, alors il existe une suite de fonctions $(s_{g_n})_{n \geq 1}$ avec $g_n \in \mathcal{G}_p$ et telle que $\left\| (s_{g_n})_- \right\|_{L^2(\mu)} \to 0$ quand $n \to \infty$. La convergence en L_2 de $(s_{g_n})_-$ implique aussi la convergence en L_1, ainsi que p.s. pour une sous-suite $s_{g_{n,k}}$ de s_{g_n}. Puisque $\int s_{g_n} d\mu = 0$ et $s_{g_n} = (s_{g_n})_- + (s_{g_n})_+$, où $(s_{g_n})_+ = max(0, s_{g_n})$, on obtient que $\int (s_{g_n})_+ \, d\mu = -\int (s_{g_n})_- \, d\mu = \int \left| (s_{g_n})_- \right| d\mu$ et donc $(s_{g_n})_+$ tend vers 0 en L_1 et p.s. pour une sous-suite $s_{g_{n,k'}}$. L'hypothèse **(A4)** assure l'existence d'une fonction dominante de $\mathcal{S}$ de carré intégrable et on a finalement qu'il existe une sous-suite de s_{g_n} convergente vers 0 p.s. et en L_2, ce qui contredit le fait que $\int s_g^2 d\mu = 1$, pour tout $g \in \mathcal{G}_p$.

On obtient alors que

$$sup_{g \in \mathcal{G}_p} \frac{\left(\sum_{t=2}^{n} s_g \left(Y_{t-1}, Y_t\right) \right)^2}{\sum_{t=2}^{n} (s_g)_-^2 \left(Y_{t-1}, Y_t\right)} = \mathcal{O}_{\mathbb{P}}(1)$$

En utilisant la suite uniformément bornée en probabilité au dessus et l'hypothèse **(A1)**, on a

$$\mathbb{P}\left(\hat{p} > p_0\right) \to 0$$

quand $n \to \infty$.

Maintenant il reste à démontrer que $\hat{p}$ ne sous-estime pas p_0 :

$$\mathbb{P}(\hat{p} < p_0) \leq \sum_{p=1}^{p_0-1} \mathbb{P}(T_n(p) > T_n(p_0)) =$$

$$= \sum_{p=1}^{p_0-1} \mathbb{P}\left(sup_{g\in\mathcal{G}_p}(l_n(g) - l_n(f)) - a_n(p) > sup_{g\in\mathcal{G}_{p_0}}(l_n(g) - l_n(f)) - a_n(p_0)\right) \leq$$

$$\leq \sum_{p=1}^{p_0-1} \mathbb{P}\left(\frac{sup_{g\in\mathcal{G}_p}(l_n(g) - l_n(f))}{n-1} > \frac{a_n(p) - a_n(p_0)}{n-1}\right)$$

en utilisant le fait que $sup_{g\in\mathcal{G}_{p_0}}(l_n(g) - l_n(f)) \geq 0$.

Sous l'hypothèse **(A3)** la classe de fonctions $\left\{log\frac{g}{f},\ g \in \mathcal{G}_p\right\}$ est Glivenko-Cantelli et on a la convergence uniforme en probabilité suivante :

$$\frac{1}{n-1} sup_{g\in\mathcal{G}_p}(l_n(g) - l_n(f)) \to sup_{g\in\mathcal{G}_p} \int log\frac{g}{f} d\mu$$

D'un autre côté, $p < p_0$ et en utilisant l'hypothèse **(A2)**, la limite est négative. Ceci avec l'hypothèse **(A1)** donne finalement

$$\mathbb{P}(\hat{p} < p_0) \to 0$$

quand $n \to \infty$ et la preuve est complète.

■

5.4 Application pour les modèles à bruit gaussien

5.4.1 Le modèle

Dans cette section, on s'intéresse aux applications du théorème précédent. Il s'agit de savoir si les hypothèses **(A1)-(A4)** sont vérifiées dans le cadre d'un modèle où les fonctions de régression sont linéaires et le bruit est gaussien.

On considère le processus $(X_t, Y_t)_{t\in\mathbb{Z}}$ qui vérifie le vrai modèle :

$$Y_t = F^0_{X_t}(Y_{t-1}) + \sigma^0_{X_t}\varepsilon_t \qquad (5.3)$$

où

- (X_t) est une suite non-observée de variables aléatoires i.i.d., prenant des valeurs dans un espace fini $\{1, ..., p_0\}$ et ayant pour distribution la loi discrète $\pi^0 = \left(\pi_i^0\right)_{i=1,...,p_0}$
- Pour chaque $i = 1, ..., p_0$, $F_i^0(y) = a_i^0 y + b_i^0$ est une fonction autorégressive qui dépend du paramètre $\theta_i^0 = \left(a_i^0, b_i^0, \sigma_i^0\right) \in E$, avec $E \subset \mathbb{R}^2 \times \mathbb{R}_+^\star$ ensemble compact
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit i.i.d. suivant une densité gaussienne centrée et réduite.

5.4.1.1 Stationnarité et ergodicité du modèle

Le processus vérifiant le vrai modèle (5.3) est stationnaire et ergodique si les hypothèses **(HS)** énoncées à la section 5.1 sont vérifiées.

Mais, comme les fonctions de régression sont linéaires et que le bruit est gaussien, la seule condition à vérifier est **(HS3)**. Pour qu'elle soit remplie, il suffit de supposer que chaque régime est stationnaire, c'est-à-dire de faire l'hypothèse supplémentaire :

Hypothèses (HS') : Pour chaque $i \in \{1, ..., p_0\}$, $\left|a_i^0\right| < 1$.

Avec cette hypotèse, on a le résultat suivant, qui garantit la stabilité du modèle, ainsi que l'existence d'un moment exponentiel pour Y_t qui servira dans la suite :

Proposition 5.2 *: Sous l'hypothèse* ***(HS')****, le processus $(X_t, Y_t)_{t\in\mathbb{Z}}$ vérifiant le modèle (5.3) est strictement stationnaire, géométriquement ergodique et, en particulier, géométriquement β-mélangeant. De plus, il existe $\delta > 0$ tel que $\mathbb{E}\left(e^{\delta Y_t^2}\right) < \infty$.*

Preuve :

La première partie du résultat est une conséquence du Théorème 2 de Yao et Attali (2000).

Pour démontrer l'existence d'un moment exponentiel, faisons d'abord quelques notations. Soit donc $\sigma = max_{i=1,...,p_0}\sigma_i^0$, $\rho = max_{i=1,...,p_0}\left|a_i^0\right| < 1$ et $b = max_{i=1,...,p_0}\left|b_i^0\right|$. Alors, pour tout $s \in \mathbb{N}^\star$, on a :

$$|Y_t|^{2s} = \left|F_{X_t}^0(Y_{t-1}) + \sigma_{X_t}^0\varepsilon_t\right|^{2s} \leq (\rho|Y_{t-1}| + b + \sigma|\varepsilon_t|)^{2s} \leq ... \leq$$

$$\leq \left(b + \sigma|\varepsilon_t| + \sum_{k=1}^{\infty}\rho^k(b + \sigma|\varepsilon_{t-k}|)\right)^{2s} = \left(\sum_{k=0}^{\infty}\rho^k(b + \sigma|\varepsilon_{t-k}|)\right)^{2s}$$

En passant à l'espérance par rapport à la loi stationnaire de Y_t :

$$\mathbb{E}\left(|Y_t|^{2s}\right)^{\frac{1}{2s}} \leq \mathbb{E}\left(\left(\sum_{k=0}^{\infty}\rho^k(b + \sigma|\varepsilon_{t-k}|)\right)^{2s}\right)^{\frac{1}{2s}} \leq \sum_{k=0}^{\infty}\rho^k\left(b + \sigma\mathbb{E}\left(|\varepsilon_{t-k}|^{2s}\right)^{\frac{1}{2s}}\right)$$

et comme $\rho < 1$ et que la norme L_2 est dominée par la norme L_{2s}, on obtient :

$$\mathbb{E}\left(|Y_t|^{2s}\right)^{\frac{1}{2s}} \leq \frac{b + \sigma\mathbb{E}\left(|\varepsilon_t|^{2s}\right)^{\frac{1}{2s}}}{1-\rho} \leq \frac{b+\sigma}{1-\rho}\mathbb{E}\left(|\varepsilon_t|^{2s}\right)^{\frac{1}{2s}}$$

Le moment exponentiel peut s'écrire

$$\mathbb{E}\left(e^{\delta Y_t^2}\right) = \sum_{k=0}^{\infty} \frac{\mathbb{E}|Y_t|^{2k}}{k!}\delta^k \leq \sum_{k=0}^{\infty} \frac{\mathbb{E}|\varepsilon_t|^{2k}}{k!}\left[\delta\left(\frac{b+\sigma}{1-\rho}\right)^2\right]^k$$

Le dernier terme étant la fonction génératrice des moments d'une loi $\chi^2(1)$, il est fini pour tout δ tel que $0 < \delta < \frac{1}{2}\left(\frac{1-\rho}{b+\sigma}\right)^2$.

■

5.4.1.2 Construction de l'estimateur de maximum de vraisemblance pénalisée

Pour estimer le vrai nombre de régimes du modèle p_0, on suppose qu'on peut fixer une borne supérieure $P > 0$ et considérer la famille de toutes les densités possibles pour Y_t , conditionnellement à Y_{t-1} et marginales en (X_t) :

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p$$

$$\mathcal{G}_p = \left\{ g \mid g(y_2 \mid y_1) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)),\ \pi_i \geq \eta > 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

où, pour chaque $i \in \{1, ..., p\}$, $F_i(y) = a_i y + b_i$ et $f_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$, avec $\theta_i = (a_i, b_i, \sigma_i) \in E$, $E \subset \mathbb{R}^2 \times \mathbb{R}_+^\star$ compact. La condition $\pi_i \geq \eta > 0$ est assez naturelle et, de plus, garantit la consistence des estimateurs de maximum de vraismblance quand p est fixé.

On définit alors $\hat{p}$ comme l'argument qui maximise le critère de vraisemblance pénalisée $T_n(p)$, où

$$T_n(p) = sup_{g \in \mathcal{G}_p} l_n(g) - a_n(p)$$

avec $l_n(g)$ et $a_n(p)$ définis dans la section 5.2.

Sous les hypothèses du théorème 5.1, $\hat{p}$ converge en probabilité vers p_0 quand $n \to \infty$. Avec un bon choix de la pénalité, l'hypothèse **(A1)** est vérifiée. Il suffit, par exemple, de considérer pour $a_n(p)$ le terme de pénalité du critère BIC. L'hypothèse **(A2)** est vérifiée si les densités sont gaussiennes selon le résultat de Teicher (1963), tandis que **(A3)** est immédiatement vraie dans le cas gaussien.

L'hypothèse clé à vérifier est donc **(A4)**. Pour le montrer, soit $\mathcal{S}$ la classe des fonctions score généralisés :

$$\mathcal{S} = \left\{ s_g,\ s_g\left(y_1, y_2\right) = \frac{\frac{g(y_2|y_1)}{f(y_2|y_1)} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}},\ g \in \mathcal{G}_P,\ \left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \neq 0 \right\}$$

où $f\left(y_2 \mid y_1\right) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0\left(y_1\right)\right)$ est la vraie densité conditionnelle marginale en X_t et μ est la mesure stationnaire de (Y_t, Y_{t-1}).

On vérifie dans la suite que cette classe est bien définie et qu'elle vérifie la condition sur l'entropie à crochets de l'hypothèse **(A4)**.

5.4.2 Fonctions scores généralisés et vérification de la propriété de Donsker

5.4.2.1 Existence des fonctions scores généralisés

Pour que le quotient des densités soit intégrable dans le cadre gaussien, on montre qu'il faut que les modèles possibles considérés ne soient pas trop différents du vrai modèle.

On commence par étudier le cas le plus simple d'un vrai régime unique contre deux régimes possibles, $p_0 = 1$ et $p = 2$. Dans ce cas, la vraie densité conditionnelle est

$$f\left(y_2 \mid y_1\right) = f^0\left(y_2 - F^0\left(y_1\right)\right)$$

et les densités possibles ont la forme suivante :

$$g\left(y_2 \mid y_1\right) = \pi f_1\left(y_2 - F_1\left(y_1\right)\right) + \left(1 - \pi\right) f_2\left(y_2 - F_2\left(y_1\right)\right)$$

Avec ces notations, on peut alors montrer le résultat qui suit :

Proposition 5.3 *: Une condition suffisante pour vérifier l'inégalité* $\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} < \infty$ *est que* $\sigma_i^2 < 2\left(\sigma^0\right)^2$ *et* $\left|a_i - a^0\right| < \sqrt{\delta\left(2\left(\sigma^0\right)^2 - \sigma_i^2\right)}$ *pour* $i = 1, 2$ *et* $\delta > 0$ *tel que* $\mathbb{E}\left(e^{\delta Y_t^2}\right) < \infty$.

Preuve :

Si μ^1 est la mesure stationnaire de Y_t, on a

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}^2 = \int_{y_1, y_2 \in \mathbb{R}} \left(\frac{g\left(y_2 \mid y_1\right)}{f\left(y_2 \mid y_1\right)} - 1 \right)^2 f\left(y_2 \mid y_1\right) dy_2 d\mu^1\left(y_1\right) =$$

$$= \int_{y_1, y_2 \in \mathbb{R}} \frac{g^2\left(y_2 \mid y_1\right)}{f\left(y_2 \mid y_1\right)} dy_2 d\mu^1\left(y_1\right) - 1$$

et en remplaçant par $g(y_2 \mid y_1) = \pi f_1(y_2 - F_1(y_1)) + (1-\pi) f_2(y_2 - F_2(y_1))$ et utilisant l'inégalité

$$2 f_1(y_2 - F_1(y_1)) f_2(y_2 - F_2(y_1)) \leq f_1^2(y_2 - F_1(y_1)) + f_2^2(y_2 - F_2(y_1)),$$

on obtient :

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}^2 \leq \int_{y_1, y_2 \in \mathbb{R}} \frac{\pi f_1^2(y_2 - F_1(y_1)) + (1-\pi) f_2^2(y_2 - F_2(y_1))}{f(y_2 \mid y_1)} dy_2 d\mu^1(y_1)$$

Le dernier terme est fini si, par exemple,

$$\begin{cases} \int_{y_1, y_2 \in \mathbb{R}} \frac{f_1^2(y_2 - F_1(y_1))}{f(y_2|y_1)} dy_2 d\mu^1(y_1) < \infty \\ \int_{y_1, y_2 \in \mathbb{R}} \frac{f_2^2(y_2 - F_2(y_1))}{f(y_2|y_1)} dy_2 d\mu^1(y_1) < \infty \end{cases}$$

Si on remplace maintenant f_1, f_2, f par des densités gaussiennes centrées d'écarts-type σ_1, σ_2 et, respectivement, σ^0 et F_1, F_2, F^0 par des fonctions linéaires, pour $i = 1, 2$ les intégrales au-dessus deviennent égales à

$$\int_{y_1 \in \mathbb{R}} \left(\int_{y_2 \in \mathbb{R}} \frac{\sigma^0}{2\sigma_i^2} e^{-\left(\frac{1}{\sigma_i^2} - \frac{1}{2(\sigma^0)^2}\right)(y_2 - m(y_1))^2} dy_2 \right) e^{\frac{\left(F_i(y_1) - F^0(y_1)\right)^2}{2(\sigma^0)^2 - \sigma_i^2}} d\mu^1(y_1)$$

où $m(y_1) = \frac{2\left(\sigma^0\right)^2 F_i(y_1) - \sigma_i^2 F^0(y_1)}{2(\sigma^0)^2 - \sigma_i^2}$.

Pour que l'intégrale en y_2 soit finie, il suffit d'avoir $\sigma_i^2 < 2\left(\sigma^0\right)^2$ et, en utilisant l'existence d'un moment exponentiel fini pour Y_t, l'intégrale en y_1 est à son tour finie si $\frac{\left(a_i - a^0\right)^2}{2(\sigma^0)^2 - \sigma_i^2} < \delta$.

■

La généralisation au cas $p, p_0 \in \mathbb{N}$ quelconques se déduit aisément du cas précédent :

Proposition 5.4 *: Une condition suffisante pour vérifier l'inégalité* $\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} < \infty$ *est que pour chaque* $i \in \{1, ..., p\}$*, il existe* $k \in \{1, ..., p_0\}$ *tel que* $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$ *et* $\left|a_i - a_k^0\right| < \sqrt{\delta\left(2\left(\sigma_k^0\right)^2 - \sigma_i^2\right)}$ *avec* $\delta > 0$ *et* $\mathbb{E}\left(e^{\delta Y_t^2}\right) < \infty$.

Preuve :

Dans le cas général, la vraie densité conditionnelle est notée

$$f(y_2 \mid y_1) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0(y_1)\right)$$

et les densités possibles ont la forme suivante

$$g\left(y_2 \mid y_1\right) = \sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right)$$

Alors, en notant μ^1 la mesure stationnaire de Y_t, la norme de la fonction score généralisé s'écrit :

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}^2 = \int_{y_1, y_2 \in \mathbb{R}} \frac{g^2\left(y_2 \mid y_1\right)}{f\left(y_2 \mid y_1\right)} dy_2 d\mu^1\left(y_1\right) - 1 =$$

$$= \int_{y_1, y_2 \in \mathbb{R}} \frac{\left(\sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right)\right)^2}{\sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0\left(y_1\right)\right)} dy_2 d\mu^1\left(y_1\right) - 1$$

En utilisant l'inégalité $\left(\sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right)\right)^2 \leq \sum_{i=1}^{p} \pi_i f_i^2\left(y_2 - F_i\left(y_1\right)\right)$, l'intégrale est finie si, pour chaque $i = 1, ..., p$,

$$\int_{y_1, y_2 \in \mathbb{R}} \frac{f_i^2\left(y_2 - F_i\left(y_1\right)\right)}{\sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0\left(y_1\right)\right)} dy_2 d\mu^1\left(y_1\right) < \infty$$

D'un autre côté, $\sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0\left(y_1\right)\right) \geq \pi_k^0 f_k^0\left(y_2 - F_k^0\left(y_1\right)\right)$ pour tout $k = 1, ..., p_0$ et donc la norme est finie si pour chaque $i \in \{1, ..., p\}$, il existe $k \in \{1, ..., p_0\}$ tel que

$$\int_{y_1, y_2 \in \mathbb{R}} \frac{f_i^2\left(y_2 - F_i\left(y_1\right)\right)}{f_k^0\left(y_2 - F_k^0\left(y_1\right)\right)} dy_2 d\mu^1\left(y_1\right) < \infty$$

Finalement, par la Proposition 5.3, cette dernière intégrale est finie si $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$ et $\frac{\left(a_i - a_k^0\right)^2}{2\left(\sigma_k^0\right)^2 - \sigma_i^2} < \delta$.

■

Par ailleurs, puisque l'hypothèse d'identifiabilité faible **(A2)** est vérifiée pour les mélanges de lois gaussiennes (Teicher (1963)) et comme on a supposé $\pi_i \geq \eta > 0$ pour tout $i \in \{1, ..., p\}$, les estimateurs de maximum de vraisemblance $\hat{\theta}_n = \left(\hat{\theta}_{1,n}, ..., \hat{\theta}_{p,n}\right)$ sont consistants. La preuve de ce résultat dans le cas plus général des changements de régime markoviens se trouve, par exemple, dans Krishnamurthy et Rydén (1998). Mais la consistance des estimateurs est suffisante pour que les conditions de la proposition 5.4 soient remplies pour n suffisament grand.

L'existence des fonctions scores généralisés étant prouvée, il reste à montrer que pour tout $\varepsilon > 0$, $\mathcal{H}_{[]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_2\right) = \mathcal{O}\left(|log\, \varepsilon|\right)$.

5.4.2.2 Vérification de la propriété de Donsker pour la classe des fonctions scores généralisés $\mathcal{S}$

Pour chaque $g \in \mathcal{G}_P$, soit $p = p(g)$ le nombre de régimes de g. On considère le paramètre global $\Phi = (\theta, \pi)$ où $\theta = (\theta_1, ..., \theta_p)$, $\pi = (\pi_1, ..., \pi_p)$ et la fonction score généralisé associée :

$$s_\Phi := s_g = \frac{\frac{g}{f} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}}$$

En général, vérifier des conditions sur l'entropie à crochets d'une classe de fonctions indexées par un paramètre est assez immédiat sous de bonnes conditions de régularité, selon le résultat suivant tiré de Van der Vaart (2000) :

Proposition 5.5 *: Soit $X_1, ..., X_n$ un n-échantillon aléatoire de loi P et soit $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ une classe de fonctions mesurables indexée par l'ensemble borné $\Theta \subset \mathbb{R}^d$. S'il existe une fonction mesurable m telle que $P|m|^2 < \infty$ et*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x) \|\theta_1 - \theta_2\|$$

pour tout $\theta_1, \theta_2 \in \Theta$, alors il existe aussi une constante K qui dépend uniquement de d et de Θ, telle que

$$\mathcal{N}_{[]}\left(\varepsilon \|m\|_{2,P}, \mathcal{F}, \|\cdot\|_{2,P}\right) \leq K \left(\frac{diam\,\Theta}{\varepsilon}\right)^d$$

pour tout $0 < \varepsilon < diam\,\Theta$.

En ce qui concerne la classe de fonctions $\mathcal{S}$, des problèmes de régularité apparaissent lorsque $\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \to 0$ et le résultat énnoncé ne peut pas être appliqué directement. Dans ce cas, l'idée consiste à diviser $\mathcal{S}$ en deux sous-ensembles, $\mathcal{S}_0$ et $\mathcal{S} \setminus \mathcal{S}_0$, tels que $\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \to 0$ sur $\mathcal{S}_0$. Il existe donc $\delta > 0$ tel que $\mathcal{S}_0 = \{s_g, g \in \mathcal{F}_0\}$ où $\mathcal{F}_0 = \left\{ g \in \mathcal{G}_P, \left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} < \delta, \left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \neq 0 \right\}$.

Sur $\mathcal{S} \setminus \mathcal{S}_0$, il est immédiat de voir que

$$\left\| \frac{\frac{g_1}{f} - 1}{\left\| \frac{g_1}{f} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{f} - 1}{\left\| \frac{g_2}{f} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq 2 \frac{\left\| \frac{g_1}{f} - \frac{g_2}{f} \right\|_{L^2(\mu)}}{\left\| \frac{g_1}{f} - 1 \right\|_{L^2(\mu)}}$$

pour tous les $g_1, g_2 \in \mathcal{G}_P \setminus \mathcal{F}_0$ et donc

$$\left\| \frac{\frac{g_1}{f} - 1}{\left\| \frac{g_1}{f} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{f} - 1}{\left\| \frac{g_2}{f} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq \frac{2}{\delta} \left\| \frac{g_1}{f} - \frac{g_2}{f} \right\|_{L^2(\mu)}$$

D'un autre côté, les conditions énoncées à la Proposition 5.4 pour assurer l'existence des fonctions score généralisés garantissent aussi que $\frac{g}{f}$ admet des dérivées partielles d'ordre un de carré intégrable. Ceci permet d'appliquer la Proposition 5.5 et on aura finalement que

$$\mathcal{N}_{[]}\left(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\delta\varepsilon}\right)^{4P}$$

Il reste donc à prouver que $\mathcal{S}_0$ est Donsker. Pour cela, l'idée est de reparamétriser le modèle de façon à ce qu'une partie des nouveaux paramétres soit identifiable et le reste regroupe toute la non-identifiabilité avec pour but d'arriver à obtenir un développement de Taylor autour de la vraie valeur du paramètre identifiable. Dans la suite, on utilise une modification de la méthode proposée par Liu et Shao (2003).

Remarquons que puisque $g \to f$ sur $\mathcal{F}_0$, on est dans le cas $p \geq p_0$. Lorsque $\frac{g}{f} - 1 = 0$, l'hypothèse d'identifiabilité faible **(A2)**, ainsi que la condition $\pi_i \geq \eta > 0$ pour tout $i = 1, ..., P$, impliquent l'existence d'un vecteur $\mathbf{t} = (t_i)_{0 \leq i \leq p_0}$ qui, modulo une permutation, permet de réécrire le paramètre Φ comme suit :

$$\theta_{t_{i-1}+1} = ... = \theta_{t_i} = \theta_i^0 \text{ et } \textstyle\sum_{j=t_{i-1}+1}^{t_i} \pi_j = \pi_i^0, \text{ pour tout } i = 1, ..., P.$$

Dans le cas général, on va alors définir $s = (s_i)_{1 \leq i \leq p_0}$ et $q = (q_j)_{1 \leq j \leq p}$ tels que, pour chaque $i \in \{1, ..., p_0\}$ et pour chaque $j \in \{t_{i-1} + 1, ..., t_i\}$,

$$s_i = \sum_{j=t_{i-1}+1}^{t_i} \pi_j - \pi_i^0, \; q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l}$$

En remarquant que par construction on a $s_{p_0} = -\sum_{i=1}^{p_0-1} s_i$, on considère alors la nouvelle paramétrisation $\Phi_{\mathbf{t}} = (\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ où

$$\phi_{\mathbf{t}} = \left((\theta_j)_{1 \leq j \leq p}, (s_i)_{1 \leq i \leq p_0 - 1} \right) \text{ et } \psi_{\mathbf{t}} = (q_j)_{1 \leq j \leq p}$$

Avec cette reparamétrisation, $\phi_{\mathbf{t}}$ contient toute la partie identifiable du modèle, alors que $\psi_{\mathbf{t}}$ contient les paramètres non-identifiables. Pour le vrai modèle $g = f$, on a donc

$$\phi_{\mathbf{t}}^0 = (\underbrace{\theta_1^0, ..., \theta_1^0}_{t_1} \; , ..., \; \underbrace{\theta_{p_0}^0, ..., \theta_{p_0}^0}_{t_{p_0} - t_{p_0-1}}, \; \underbrace{0, ..., 0}_{p_0 - 1})^T$$

Maintenant, on s'intéresse à l'existence d'un développement de Taylor à l'ordre deux de $\frac{g}{f}-1$ autour de $\phi_{\mathbf{t}}^0$. Pour cela, on introduit d'abord quelques notations qui permettent de faciliter l'écriture. Soit donc

$$g_j(y_1,y_2)=g_{\theta_j}(y_2\mid y_1)=\frac{f_j(y_2-F_j(y_1))}{\sum_{i=1}^{p_0}\pi_i^0 f_i^0\left(y_2-F_i^0(y_1)\right)}-1$$

Avec les paramètres nouveaux définis, le quotient des densités s'écrit alors

$$\frac{g}{f}-1=\sum_{i=1}^{p_0}\left(s_i+\pi_i^0\right)\sum_{j=t_{i-1}+1}^{t_i}q_jg_j$$

et comme $s_{p_0}=-\sum_{i=1}^{p_0-1}s_i$, on a

$$\frac{g}{f}-1=\sum_{i=1}^{p_0-1}\left(s_i+\pi_i^0\right)\sum_{j=t_{i-1}+1}^{t_i}q_jg_j+\left(\pi_{p_0}^0-\sum_{i=1}^{p_0-1}s_i\right)\sum_{j=t_{p_0-1}+1}^{t_{p_0}}q_jg_j \tag{5.4}$$

En remarquant aussi que lorsque $\phi_{\mathbf{t}}=\phi_{\mathbf{t}}^0$, le rapport $\frac{g}{f}$ ne varie pas en fonction de $\psi_{\mathbf{t}}$, nous allons étudier la variation de ce rapport pour $\phi_{\mathbf{t}}$ dans un voisinage de $\phi_{\mathbf{t}}^0$ et pour $\psi_{\mathbf{t}}$ fixé. On utilise les notations suivantes pour les $\phi_{\mathbf{t}}$-dérivées partielles de la fonction quotient calculées en $\phi_{\mathbf{t}}^0$:

$$g_j':=\frac{\partial g_j}{\partial\theta_j}\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right),\ g_j'':=\frac{\partial^2 g_j}{\partial\theta_j^2}\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right),\ g_j'''':=\frac{\partial^3 g_j}{\partial\theta_j^3}\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)$$

Avec ces notations, on a la proposition suivante :

Proposition 5.6 : *Pour chaque $\psi_{\mathbf{t}}$ fixé, il existe un développement de Taylor à l'ordre deux autour de $\phi_{\mathbf{t}}^0$:*

$$\frac{g}{f}-1=\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)^T g'_{\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)}+\frac{1}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)^T g''_{\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)+o\left(D\left(\phi_{\mathbf{t}},\psi_{\mathbf{t}}\right)\right)$$

où $D\left(\phi_{\mathbf{t}},\psi_{\mathbf{t}}\right)=\left\|\frac{g}{f}-1\right\|_{L^2(\mu)}$,

$$\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)^T g'_{\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)}=\sum_{i=1}^{p_0}\pi_i^0\left(\sum_{j=t_{i-1}+1}^{t_i}q_j\theta_j-\theta_i^0\right)^T g_i'+\sum_{i=1}^{p_0}s_i g_{\theta_i^0} \tag{5.5}$$

et

$$\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} g''_{\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)=\sum_{i=1}^{p_0}\left[2s_i\left(\sum_{j=t_{i-1}+1}^{t_i} q_j\theta_j-\theta_i^0\right)^{T} g_i'+\right.$$

$$\left.+\pi_i^0\sum_{j=t_{i-1}+1}^{t_i} q_j\left(\theta_j-\theta_i^0\right)^{T} g_i''\left(\theta_j-\theta_i^0\right)\right] \qquad (5.6)$$

De plus,

$$\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} g'_{\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)}+\frac{1}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} g''_{\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)=0\Leftrightarrow\phi_{\mathbf{t}}=\phi_{\mathbf{t}}^{0} \qquad (5.7)$$

Preuve :

Le premier terme se calcule facilement en remarquant que le $\phi_{\mathbf{t}}$-gradient de $\frac{g}{f}-1$ calculé au point $\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)$ s'écrit :

- pour $i\in\{1,...,p_0\}$ et $j\in\{t_{i-1}+1,...,t_i\}$, $\frac{\partial\left(\frac{g}{f}-1\right)}{\partial\theta_j}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=\pi_i^0 q_j g_i'$
- pour $i\in\{1,...,p_0-1\}$,

$$\frac{\partial\left(\frac{g}{f}-1\right)}{\partial s_i}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=\sum_{j=t_{i-1}+1}^{t_i} q_j g_{\theta_i^0}-\sum_{j=t_{p_0-1}+1}^{t_{p_0}} q_j g_{\theta_{p_0}^0}=g_{\theta_i^0}-g_{\theta_{p_0}^0}$$

Le terme d'ordre deux s'obtient directement, une fois que la matrice hessienne est calculée en $\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)$:

- $\frac{\partial^2\left(\frac{g}{f}-1\right)}{\partial\theta_j^2}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=\pi_i^0 q_j g_i''$, pour $i=1,...,p_0$ et $j=t_{i-1}+1,...,t_i$
- $\frac{\partial^2\left(\frac{g}{f}-1\right)}{\partial\theta_j\partial\theta_l}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=0$, pour $j,l=1,...,p$ avec $j\neq l$
- $\frac{\partial^2\left(\frac{g}{f}-1\right)}{\partial s_i\partial s_k}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=0$, pour $i,k=1,...,p_0-1$
- $\frac{\partial^2\left(\frac{g}{f}-1\right)}{\partial s_i\partial\theta_j}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=q_j g_i'$, pour $i=1,...,p_0-1$ et $j=t_{i-1}+1,...,t_i$
- $\frac{\partial^2\left(\frac{g}{f}-1\right)}{\partial s_i\partial\theta_j}\left(\phi_{\mathbf{t}}^{0},\psi_{\mathbf{t}}\right)=-q_j g_{p_0}'$, pour $i=1,...,p_0-1$ et $j=t_{p_0-1}+1,...,t_{p_0}$
- les autres dérivées croisées de s_i et θ_j sont nulles

Montrons maintenant que le reste est $o\left(D\left(\phi_{\mathbf{t}},\psi_{\mathbf{t}}\right)\right)$. Pour cela, il suffit que les $\phi_{\mathbf{t}}$-dérivées partielles à l'ordre trois de $\frac{g}{f}-1$ soient uniformément dominées par une fonction d'intégrale finie. Mais, en utilisant l'expression (5.4), il est immédiat que ces dérivées s'écrivent comme des combinaisons linéaires de g_i, g'_{θ_i}, g''_{θ_i}, g'''_{θ_i}, $i\in\{1,...,p\}$ et donc il suffit que ces fonctions soient uniformément dominées par une fonction d'intégrale finie.

Vérifions cette dernière affirmation. On commence par calculer g'_{θ_i}, g''_{θ_i}, g'''_{θ_i}, $i \in \{1, ..., p\}$. Comme $\theta_i = (a_i, b_i, \sigma_i)$, les dérivées partielles d'ordre un de g_i, $i = 1, ..., p$ sont :

$$\frac{\partial g_i}{\partial \theta_i}(y_1, y_2) = \frac{1}{\sqrt{2\pi} f(y_2 \mid y_1)} \left(\sqrt{2\pi} \frac{\partial f_i}{\partial a_i}(y_2 \mid y_1), \sqrt{2\pi} \frac{\partial f_i}{\partial b_i}(y_2 \mid y_1), \sqrt{2\pi} \frac{\partial f_i}{\partial \sigma_i}(y_2 \mid y_1) \right)$$

où

$$\sqrt{2\pi} \frac{\partial f_i}{\partial a_i}(y_2 \mid y_1) = \frac{y_1}{\sigma_i^3}(y_2 - F_i(y_1)) e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial f_i}{\partial b_i}(y_2 \mid y_1) = \frac{1}{\sigma_i^3}(y_2 - F_i(y_1)) e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial f_i}{\partial \sigma_i}(y_1 \mid y_2) = \left[\frac{(y_2 - F_i(y_1))^2}{\sigma_i^4} - \frac{1}{\sigma_i^2} \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

La matrice hessienne de g_i, $i = 1, ..., p$, s'écrit comme suit :

$$\frac{\partial^2 g_i}{\partial \theta_i^2} = \frac{1}{\sqrt{2\pi} f(y_2 \mid y_1)} \begin{pmatrix} \sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i^2}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial b_i}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial \sigma_i}(y_2 \mid y_1) \\ \sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial b_i}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial b_i^2}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial b_i \partial \sigma_i}(y_2 \mid y_1) \\ \sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial \sigma_i}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial b_i \partial \sigma_i}(y_2 \mid y_1) & \sqrt{2\pi} \frac{\partial^2 f_i}{\partial \sigma_i^2}(y_2 \mid y_1) \end{pmatrix}$$

avec

$$\sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i^2}(y_2 \mid y_1) = \left[-\frac{y_1^2}{\sigma_i^3} + \frac{y_1^2}{\sigma_i^5}(y_2 - F_i(y_1))^2 \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial b_i}(y_2 \mid y_1) = \left[-\frac{y_1}{\sigma_i^3} + \frac{y_1}{\sigma_i^5}(y_2 - F_i(y_1))^2 \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial^2 f_i}{\partial a_i \partial \sigma_i}(y_2 \mid y_1) = \left[-\frac{3y_1}{\sigma_i^4}(y_2 - F_i(y_1)) + \frac{y_1}{\sigma_i^6}(y_2 - F_i(y_1))^3 \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial^2 f_i}{\partial b_i^2}(y_2 \mid y_1) = \left[-\frac{1}{\sigma_i^3} + \frac{1}{\sigma_i^5}(y_2 - F_i(y_1))^2 \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi} \frac{\partial^2 f_i}{\partial b_i \partial \sigma_i}(y_2 \mid y_1) = \left[-\frac{3}{\sigma_i^4}(y_2 - F_i(y_1)) + \frac{1}{\sigma_i^6}(y_2 - F_i(y_1))^3 \right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_i}{\partial\sigma_i^2}(y_2 \mid y_1) = \left[\frac{1}{\sigma_i^7}(y_2 - F_i(y_1))^4 - \frac{5}{\sigma_i^5}(y_2 - F_i(y_1))^2 + \frac{2}{\sigma_i^3}\right] e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$$

Pour les dérivées d'ordre trois, on remarque qu'elles s'écriront comme des combinaisons linéaires de puissances de $\frac{1}{\sigma_i}$, y_1 et $y_2 - F_i(y_1)$, multipliées par des fonctions de type $\frac{1}{\sqrt{2\pi}f(y_2|y_1)} e^{-\frac{1}{2\sigma_i^2}(y_2 - F_i(y_1))^2}$. Cette remarque, ainsi que les conditions énoncées en début de cette section, c'est-à-dire que pour tout $i \in \{1, ..., p\}$, il existe $k \in \{1, ..., p_0\}$ tel que $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$ et $\left|a_i - a_k^0\right| < \sqrt{\delta\left(2\left(\sigma_k^0\right)^2 - \sigma_i^2\right)}$ avec $\delta > 0$ et $\mathbb{E}\left(e^{\delta Y_t^2}\right) < \infty$, assurent le fait que les fonctions g_i et leurs dérivées partielles jusqu'à l'ordre trois sont dominées uniformément par une fonction intégrable.

Il nous reste à montrer la dernière partie de la proposition et comme l'implication inverse est évidente, il faut prouver uniquement que

$$\left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right)^T g'_{\left(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}\right)} + \frac{1}{2}\left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right)^T g''_{\left(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right) = 0 \Rightarrow \phi_{\mathbf{t}} = \phi_{\mathbf{t}}^0$$

Pour cela, on va énoncer et démontrer le lemme suivant :

Lemme 5.1 *: La famille de fonctions*

$$\left\{g_{\theta_i^0}, \frac{\partial g_{\theta_i^0}}{\partial a_i}, \frac{\partial g_{\theta_i^0}}{\partial b_i}, \frac{1}{\sigma_i^0}\frac{\partial g_{\theta_i^0}}{\partial \sigma_i} + \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i \partial \sigma_i}, i = 1, ..., p_0\right\}$$

est linéairement indépendante.

Preuve du Lemme 5.1 :

Pour prouver l'indépendance linéaire, on a d'abord besoin de l'identité suivante qui est vraie pour $\theta_i^0 = \left(a_i^0, b_i^0, \sigma_i^0\right)$, $i \in \{1, ..., p_0\}$ distincts :

$$\sum_{i=1}^{p_0} P_i(y_1, y_2)\, e^{-\frac{1}{2\left(\sigma_i^0\right)^2}\left(y_2 - F_i^0(y_1)\right)^2} = 0,\ (\forall)\, y_1, y_2 \Leftrightarrow P_i(y_1, y_2) = 0,\ i = 1, ..., p_0$$

où $P_i(y_1, y_2)$ sont des polynômes en y_1 et y_2 et $F_i^0(y_1) = a_i^0 y_1 + b_i^0$, pour $i \in \{1, ..., p_0\}$.

Cette identité se démontre facilement en supposant que, après une éventuelle permutation, les vecteurs des paramètres θ_i^0 sont ordonnés de sorte que l'on ait :

- $\sigma_{p_0}^0 \geq ... \geq \sigma_1^0$
- pour tous les indices i tels que $\sigma_{i+1}^0 = \sigma_i^0$, on a $b_{i+1}^0 \geq b_i^0$
- pour tous les indices i tels que $\sigma_{i+1}^0 = \sigma_i^0$ et $b_{i+1}^0 = b_i^0$, on a $a_{i+1}^0 > a_i^0$

L'identité peut se réécrire :

$$\sum_{i=1}^{p_0} P_i\left(y_1,y_2\right) e^{-\frac{1}{2\left(\sigma_i^0\right)^2}\left(y_2-F_i^0(y_1)\right)^2} = 0,\ (\forall)\, y_1,y_2 \Leftrightarrow$$

$$\Leftrightarrow e^{-\frac{1}{2\left(\sigma_{p_0}^0\right)^2}\left(y_2-F_{p_0}^0(y_1)\right)^2} \left[P_{p_0}\left(y_1,y_2\right) + \right.$$

$$\left. + \sum_{i=1}^{p_0-1} P_i\left(y_1,y_2\right) e^{-\frac{1}{2\left(\sigma_i^0\right)^2}\left(y_2-F_i^0(y_1)\right)^2+\frac{1}{2\left(\sigma_{p_0}^0\right)^2}\left(y_2-F_{p_0}^0(y_1)\right)^2}\right] = 0,\ \forall y_1,y_2$$

Comme les paramètres sont distincts et en utilisant le réarrangement au dessus, le deuxième terme dans la somme tend vers zéro quand $y_2 \to \infty$ et $y_1 \to \infty$ et donc on obtient $P_{p_0}\left(y_1,y_2\right) = 0,\ (\forall)\, y_1,y_2$. Après, par récurrence, on aura $P_i\left(y_1,y_2\right) = 0,\ \forall y_1,y_2,\ i \in \{1,...,p_0\}$.

L'identité démontrée, on peut maintenant prouver l'indépendance linéaire. On considère l'égalité suivante :

$$\textstyle\sum_{i=1}^{p_0} \alpha_i g_{\theta_i^0} + \sum_{i=1}^{p_0} \beta_i^T g_i' + \sum_{i=1}^{p_0} \gamma_i^T g_i'' = 0$$

avec $g_i' = \left(\frac{\partial g_{\theta_i^0}}{\partial a_i}, \frac{\partial g_{\theta_i^0}}{\partial b_i}, \frac{\partial g_{\theta_i^0}}{\partial \sigma_i}\right)^T$, $g_i'' = \left(\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i \partial \sigma_i}\right)^T$, $\beta_i^T = (\beta_{i,1}, \beta_{i,2}, \beta_{i,3})$ et $\gamma_i^T = (\gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3}, \gamma_{i,4}, \gamma_{i,5}, \gamma_{i,6})$.

En utilisant les expressions analytiques des dérivées partielles, on peut réécrire l'égalité

$$\sum_{i=1}^{p_0} \alpha_i g_{\theta_i^0} + \sum_{i=1}^{p_0} \beta_i^T g_i' + \sum_{i=1}^{p_0} \gamma_i^T g_i'' = 0 \Leftrightarrow \sum_{i=1}^{p_0} P_i\left(y_1,y_2\right) e^{-\frac{1}{2\left(\sigma_i^0\right)^2}\left(y_2-F_i^0(y_1)\right)^2} = \sum_{i=1}^{p_0} \alpha_i$$

où

$$P_i\left(y_1,y_2\right) = \frac{1}{f\left(y_2 \mid y_1\right)} \frac{1}{\sqrt{2\pi}\sigma_i^0} \left[\alpha_i + \beta_{i,1} \frac{y_1}{\left(\sigma_i^0\right)^2}\left(y_2 - F_i^0\left(y_1\right)\right) + \beta_{i,2} \frac{1}{\left(\sigma_i^0\right)^2}\left(y_2 - F_i^0\left(y_1\right)\right) + \right.$$

$$+\beta_{i,3}\left(\frac{1}{\left(\sigma_i^0\right)^3}\left(y_2 - F_i^0\left(y_1\right)\right)^2 - \frac{1}{\sigma_i^0}\right) + \gamma_{i,1}\left(-\frac{y_1^2}{\left(\sigma_i^0\right)^2} + \frac{y_1^2}{\left(\sigma_i^0\right)^4}\left(y_2 - F_i^0\left(y_1\right)\right)^2\right) +$$

$$\left. +\gamma_{i,2}\left(-\frac{1}{\left(\sigma_i^0\right)^2} + \frac{1}{\left(\sigma_i^0\right)^4}\left(y_2 - F_i^0\left(y_1\right)\right)^2\right) + \right.$$

$$+\gamma_{i,3}\left(\frac{1}{\left(\sigma_i^0\right)^6}\left(y_2-F_i^0\left(y_1\right)\right)^4-\frac{5}{\left(\sigma_i^0\right)^4}\left(y_2-F_i^0\left(y_1\right)\right)^2+\frac{2}{\left(\sigma_i^0\right)^2}\right)+$$

$$+\gamma_{i,4}\left(-\frac{y_1}{\left(\sigma_i^0\right)^2}+\frac{y_1}{\left(\sigma_i^0\right)^4}\left(y_2-F_i^0\left(y_1\right)\right)^2\right)+$$

$$+\gamma_{i,5}\left(-\frac{3y_1}{\left(\sigma_i^0\right)^3}\left(y_2-F_i^0\left(y_1\right)\right)+\frac{y_1}{\left(\sigma_i^0\right)^5}\left(y_2-F_i^0\left(y_1\right)\right)^3\right)+$$

$$+\gamma_{i,6}\left(-\frac{3}{\left(\sigma_i^0\right)^3}\left(y_2-F_i^0\left(y_1\right)\right)+\frac{1}{\left(\sigma_i^0\right)^5}\left(y_2-F_i^0\left(y_1\right)\right)^3\right)\Bigg]$$

En considérant maintenant $y_2 \to \infty$, le terme de gauche de l'égalité s'annule et donc $\sum_{i=1}^{p_0}\alpha_i=0$. D'après l'identité qu'on vient de montrer, on a alors que $P_i\left(y_1,y_2\right)=0$, pour tout $i=1,...,p_0$. Ensuite, il ne reste qu'à identifier les coéfficients. Pour les termes y_2^4, $y_1y_2^3$, y_2^3, $y_1^2y_2^2$ et $y_1y_2^2$ on obtient immédiatement que $\gamma_{i,3}=\gamma_{i,5}=\gamma_{i,6}=\gamma_{i,1}=\gamma_{i,4}=0$ et les termes qui restent sont

$$\alpha_i+\beta_{i,1}\frac{y_1}{\left(\sigma_i^0\right)^2}\left(y_2-F_i^0\left(y_1\right)\right)+\beta_{i,2}\frac{1}{\left(\sigma_i^0\right)^2}\left(y_2-F_i^0\left(y_1\right)\right)+\beta_{i,3}\left(\frac{1}{\left(\sigma_i^0\right)^3}\left(y_2-F_i^0\left(y_1\right)\right)^2-\frac{1}{\sigma_i^0}\right)+$$

$$+\gamma_{i,2}\left(-\frac{1}{\left(\sigma_i^0\right)^2}+\frac{1}{\left(\sigma_i^0\right)^4}\left(y_2-F_i^0\left(y_1\right)\right)^2\right)=0,\ (\forall)\, y_1,y_2$$

Ici encore on procède par identification des coefficients, d'abord de y_2^2, puis de y_1y_2 et de y_2 et on aura finalement que $\gamma_{i,2}=-\frac{\beta_{i,3}}{\sigma_i}$ et $\beta_{i,1}=\beta_{i,2}=\alpha_i=0$ et l'indépendance linéaire est démontrée.

(fin de la preuve du Lemme 5.1)■

Revenons maintenant à la preuve de la Proposition 5.6. Il nous reste à montrer l'implication suivante

$$\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)^T g'_{\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)}+\frac{1}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)^T g''_{\left(\phi_{\mathbf{t}}^0,\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^0\right)=0\Rightarrow\phi_{\mathbf{t}}=\phi_{\mathbf{t}}^0$$

Pour $\psi_{\mathbf{t}}$ fixé, on considère $\phi_{\mathbf{t}}$ qui vérifie l'égalité

$$\left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right)^T g'_{\left(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}\right)} + \frac{1}{2}\left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right)^T g''_{\left(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}\right)} \left(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\right) = 0.$$

En remplaçant les deux termes par les expressions (5.5) et (5.6), l'égalité est équivalente à :

$$\sum_{i=1}^{p_0} \pi_i^0 \left(\sum_{j=t_{i-1}+1}^{t_i} q_j \theta_j - \theta_i^0 \right)^T g_i' + \sum_{i=1}^{p_0} s_i g_{\theta_i^0} +$$

$$+ \sum_{i=1}^{p_0} \left[2 s_i \left(\sum_{j=t_{i-1}+1}^{t_i} q_j \theta_j - \theta_i^0 \right)^T g_i' + \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(\theta_j - \theta_i^0\right)^T g_i'' \left(\theta_j - \theta_i^0\right) \right] = 0$$

ou encore, en regroupant les termes,

$$\sum_{i=1}^{p_0} s_i g_{\theta_i^0} + \sum_{i=1}^{p_0} \left(\pi_i^0 + 2 s_i\right) \left(\sum_{j=t_{i-1}+1}^{t_i} q_j \theta_j - \theta_i^0 \right)^T g_i' +$$

$$+ \sum_{i=1}^{p_0} \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(\theta_j - \theta_i^0\right)^T g_i'' \left(\theta_j - \theta_i^0\right) = 0$$

D'après l'indépendance linéaire du lemme 5.1, l'égalité a lieu si et seulement si

– les coefficients de $g_{\theta_i^0}$ sont nuls, ce qui implique $s_i = 0$, $i = 1, ..., p_0$

– les coefficients de $\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}$ et $\frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}$ sont nuls et alors, pour tout $i = 1, ..., p_0$

$$\pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(a_j - a_i^0\right)^2 = \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(\sigma_j - \sigma_i^0\right)^2 = 0$$

et comme π_i^0 et $q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l}$ ont été supposés strictement positifs, on aura nécessairement $a_j = a_i^0$ et $\sigma_j = \sigma_i^0$ pour tout $i = 1, ..., p_0$, $j = t_{i-1} + 1, ..., t_i$.

En remplaçant les coefficients s_i, a_j and σ_j par leurs expressions, l'égalité devient

$$\sum_{i=1}^{p_0} \pi_i^0 \left(\sum_{j=t_{i-1}+1}^{t_i} q_j b_j - b_i^0 \right) \frac{\partial g_{\theta_i^0}}{\partial b_i} + \sum_{i=1}^{p_0} \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(b_j - b_i^0\right)^2 \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2} = 0$$

et toujours en raison de l'indépendance linéaire, $b_j = b_i^0$ pour tout $i = 1, ..., p_0$, $j = t_{i-1} + 1, ..., t_i$, ce qui donne finalement $\phi_{\mathbf{t}} = \phi_{\mathbf{t}}^0$.

■

En utilisant le développement de Taylor qu'on vient d'écrire, on peut maintenant prouver que $\mathcal{S}_0$ peut être plongé dans une classe de fonctions qui vérifie la propriété de Donsker. Ce résultat est résumé dans la proposition suivante :

Proposition 5.7 *: Le nombre d'ε-crochets $\mathcal{N}_{[]}(\varepsilon, \mathcal{S}_0, \|\cdot\|_2)$ nécessaires pour couvrir $\mathcal{S}_0$ est d'un ordre plus petit que $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{9p_0}$.*

Preuve :

L'idée de la démonstration est de borner $\mathcal{N}_{[]}(\varepsilon, \mathcal{S}_0, \|\cdot\|_2)$ par le nombre d'ε-crochets nécessaires pour couvrir une classe plus large de fonctions. Pour chaque $g \in \mathcal{F}_0$, on considère la reparamétrisation $\Phi_{\mathbf{t}} = (\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ qui permet d'écrire un développement de Taylor à l'ordre deux de $\frac{g}{f} - 1$ autour de la vraie valeur du paramètre $\phi^0_{\mathbf{t}}$. On verra dans la suite que $\mathcal{N}_{[]}(\varepsilon, \mathcal{S}_0, \|\cdot\|_2)$ est majoré par le nombre d'ε-crochets recouvrant l'ensemble des développements limités, pour tout $\Phi_{\mathbf{t}} = (\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ dans un ensemble compact.

Selon la proposition 5.6, à $\psi_{\mathbf{t}}$ fixé, on a

$$\frac{g_{(\phi_{\mathbf{t}},\psi_{\mathbf{t}})}}{f} - 1 = \left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right)^T g'_{\left(\phi^0_{\mathbf{t}},\psi_{\mathbf{t}}\right)} + \frac{1}{2}\left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right)^T g''_{\left(\phi^0_{\mathbf{t}},\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right) + o\left(D\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)\right)$$

où $\left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right)^T g'_{\left(\phi^0_{\mathbf{t}},\psi_{\mathbf{t}}\right)}$ et $\frac{1}{2}\left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right)^T g''_{\left(\phi^0_{\mathbf{t}},\psi_{\mathbf{t}}\right)}\left(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}\right)$ sont des combinaisons linéaires de $g_{\theta^0_i}$, g'_i, g''_i, $i = 1, ..., p_0$ et donc on aura l'écriture suivante :

$$\frac{g_{(\phi_{\mathbf{t}},\psi_{\mathbf{t}})}}{f} - 1 = \sum_{i=1}^{p_0} \alpha_i g_{\theta^0_i} + \sum_{i=1}^{p_0} \beta_{i,1} \frac{\partial g_{\theta^0_i}}{\partial a_i} + \sum_{i=1}^{p_0} \beta_{i,2} \frac{\partial g_{\theta^0_i}}{\partial b_i} + \sum_{i=1}^{p_0} \beta_{i,3}\left(\frac{1}{\sigma^0_i}\frac{\partial g_{\theta^0_i}}{\partial \sigma_i} + \frac{\partial^2 g_{\theta^0_i}}{\partial b_i^2}\right) +$$

$$+ \sum_{i=1}^{p_0} \gamma_{i,1} \frac{\partial^2 g_{\theta^0_i}}{\partial a_i^2} + \sum_{i=1}^{p_0} \gamma_{i,2} \frac{\partial^2 g_{\theta^0_i}}{\partial \sigma_i^2} + \sum_{i=1}^{p_0} \gamma_{i,3} \frac{\partial^2 g_{\theta^0_i}}{\partial a_i \partial \sigma_i} + \sum_{i=1}^{p_0} \gamma_{i,4} \frac{\partial^2 g_{\theta^0_i}}{\partial a_i \partial b_i} + \sum_{i=1}^{p_0} \gamma_{i,5} \frac{\partial^2 g_{\theta^0_i}}{\partial b_i \partial \sigma_i} + o\left(D\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)\right)$$

Comme, selon le lemme 5.1, la famille de fonctions

$$\left\{ g_{\theta^0_i}, \frac{\partial g_{\theta^0_i}}{\partial a_i}, \frac{\partial g_{\theta^0_i}}{\partial b_i}, \frac{1}{\sigma^0_i}\frac{\partial g_{\theta^0_i}}{\partial \sigma_i} + \frac{\partial^2 g_{\theta^0_i}}{\partial b_i^2}, \frac{\partial^2 g_{\theta^0_i}}{\partial a_i^2}, \frac{\partial^2 g_{\theta^0_i}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta^0_i}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta^0_i}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta^0_i}}{\partial b_i \partial \sigma_i}, i = 1, ..., p_0 \right\}$$

est libre dans L_2, il existe une constante $m > 0$ telle que pour tout vecteur $(\alpha_i, \beta_{i,j}, \gamma_{i,k}, i = 1, ..., p_0, j = 1, ..., 3, k = 1, ..., 5)$ de norme 1, on a

$$\left\| \sum_{i=1}^{p_0} \alpha_i g_{\theta^0_i} + \sum_{i=1}^{p_0} \beta_i^T g_{1,i} + \sum_{i=1}^{p_0} \gamma_i^T g_{2,i} \right\|_{L^2(\mu)} \geq m$$

où $\beta_i^T = (\beta_{i,1}, \beta_{i,2}, \beta_{i,3})$, $\gamma_i^T = (\gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3}, \gamma_{i,4}, \gamma_{i,5})$, $g_{1,i}^T = \left(\frac{\partial g_{\theta_i^0}}{\partial a_i}, \frac{\partial g_{\theta_i^0}}{\partial b_i}, \frac{1}{\sigma_i^0} \frac{\partial g_{\theta_i^0}}{\partial \sigma_i} + \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2} \right)$ et $g_{2,i}^T = \left(\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i \partial \sigma_i} \right)$.

En même temps, on a aussi

$$\left\| \frac{\frac{g(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}{f} - 1}{\left\| \frac{g(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}{f} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} = 1$$

et donc on déduit que la norme euclidienne des coefficients de la combinaison linéaire dans le développement à l'ordre deux de $\frac{\frac{g(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}{f} - 1}{\left\| \frac{g(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}{f} - 1 \right\|_{L^2(\mu)}}$ est majorée par $\frac{1}{m}$.

Remarquons aussi que si la norme des coefficients dans le développement de $\frac{g(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}{f} - 1$ n'est pas unitaire, on peut normaliser en divisant les termes du dénominateur et du numérateur par la norme elle-même, l'identité (5.7) de la Proposition 5.6 assurant le fait qu'elle est non-nulle.

Par conséquant, on peut plonger $\mathcal{S}_0$ dans

$$\mathcal{H} = \left\{ \sum_{i=1}^{p_0} \left(\alpha_i g_{\theta_i^0} + \beta_i^T g_{i,1} + \gamma_i^T g_{i,2} \right) + o(1), \left\| \left(\alpha_i, \beta_i^T, \gamma_i^T, i = \overline{1, p_0} \right) \right\| \leq \frac{1}{m} \right\}$$

Il est alors évident que le nombre de crochets recouvrant $\mathcal{H}$ est $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{9p_0}$ et donc $\mathcal{N}_{[]}(\varepsilon, \mathcal{S}_0, \|\cdot\|_2)$ est d'un ordre plus petit que $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{9p_0}$.

■

La dernière proposition montre donc que l'hypothèse **(A4)** sur l'entropie de la classe des fonctions score généralisées du théorème 5.1 est vérifiée. On peut donc construire un estimateur consistent pour le nombre de régimes d'un modèle autorégressif stationnaire, à fonctions de régression linéaires et à bruit gaussien.

5.5 Est-il possible d'étendre le résultat aux changements de régime markoviens ?

5.5.1 Le modèle

Dans cette section, on étudie le point suivant : les résultats de consistence obtenus pour l'estimateur du nombre de régimes d'un modèle à changements de régime indépendants peuvent-ils s'étendre au cadre plus général des modèles à changements markoviens ? Comme

dans le cas indépendant, on considère que le nombre de retards est connu et égal à un. Soit le processus $(X_t, Y_t)_{t\in\mathbb{Z}}$ qui vérifie le vrai modèle :

$$Y_t = F^0_{X_t}(Y_{t-1}) + \varepsilon_t(X_t) \qquad (5.5)$$

où

- (X_t) est une chaîne de Markov homogène, apériodique et irréductible, à valeurs dans un espace d'états fini $S = \{1, ..., p_0\}$, $p_0 \in \mathbb{N}^\star$ et caractérisée par sa matrice de transition $A = \left(\pi^0_{ij}\right)_{i,j=1,...,p_0}$ et sa mesure invariante $\pi^0 = \left(\pi^0_i\right)_{i=1,...,p_0}$
- pour chaque $i \in \{1, ..., p_0\}$, $F^0_i(y)$ est une fonction nonlinéaire autorégressive qui dépend du paramètre $\theta^0_i \in E$, avec $E \subset \mathbb{R}^d$ ensemble compact
- pour chaque $i \in \{1, ..., p_0\}$, $(\varepsilon_t(i))_{t\in\mathbb{Z}}$ est un bruit i.i.d., les suites $(\varepsilon_t(i))_{1\le i\le p_0}$ étant indépendantes.

Suivant la section 4.1, on introduit les hypothèses suivantes qui garantissent l'existence d'une unique solution strictement stationnaire et ergodique :

Hypothèse (HS)

(HS1) Les fonctions de régression sont continues et sous-linéaires, c'est-à-dire que pour tout ,$i = 1, ..., p_0$ il existe $\left(a^0_i, b^0_i\right) \in \mathbb{R}^2_+$ tels que $\left|F^0_i(y)\right| \le a^0_i |y| + b^0_i$, $y \in \mathbb{R}$

(HS2) Pour tout $i \in \{1, ..., p_0\}$, le bruit $(\varepsilon_t(i))$ admet une densité strictement positive par rapport à la mesure de Lebesgue et il existe $s \ge 1$ tel que $\mathbb{E}|\varepsilon_1(i)|^s < \infty$

(HS3) Le rayon spectral de la matrice Q_s vérifie la relation $\rho(Q_s) < 1$, où

$$Q_s = \begin{pmatrix} \left(a^0_1\right)^s \pi^0_{11} & \cdots & \left(a^0_{p_0}\right)^s \pi^0_{1p_0} \\ \vdots & \ddots & \vdots \\ \left(a^0_1\right)^s \pi^0_{p_0 1} & \cdots & \left(a^0_{p_0}\right)^s \pi^0_{p_0 p_0} \end{pmatrix}$$

L'hypothèse (H3) est immédiatement vérifiée si, par exemple, la matrice Q_s est sous-stochastique et pour cela il suffit d'avoir $a^0_i < 1$ pour tout $i = 1, ..., p_0$.

5.5.2 Existence d'une fonction de coût ?

Soit maintenant $\{y_1, ..., y_n\}$ les valeurs observées d'un n-échantillon de la série Y_t qui vérifie le vrai modèle (5.5). Le but étant d'étendre au cadre markovien la construction de l'estimateur de vraisemblance pénalisée défini à la section 5.2 plusieurs inconvénients apparaissent : d'un côté la non-identifiabilité du modèle, mais on a vu précédemment que ce problème peut être contourné par une reparametrisation bien choisie et d'un autre côté (X_t), qui est un processus markovien non-observé. La dépendance du processus caché ne permet pas une expression explicite de la densité de Y_t conditionnellement au passé et marginale en (X_t) :

$$f(y_t \mid y_{t-1}, ..., y_1) = \sum_{i=1}^{p_0} \mathbb{P}(X_t = i \mid y_{t-1}, ..., y_1)\, f^0_i\left(y_t - F^0_i(y_{t-1})\right)$$

car $\mathbb{P}(X_t = i \mid y_{t-1}, ..., y_0)$ n'a pas une expression analytique et se calcule de manière récursive.

Pourtant, (X_t) étant un processus stationnaire, on peut introduire une nouvelle fonction de coût construite à partir de sa distribution invariante.

Comme dans la section 5.2, on fixe une borne supérieure pour le nombre de régimes $P > 0$ et on considère tous les mélanges de densités jusqu'à P composantes :

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p$$

$$\mathcal{G}_p = \left\{ g \mid \ g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)),\ \pi_i \geq 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

où, pour chaque $i = 1, ..., p$,

- F_i est une fonction nonlinéaire autorégressive qui dépend du paramètre $\theta_i \in E$
- f_i est une densité strictement positive par rapport à la mesure de Lebesgue, qui dépend du paramètre $\theta_i \in E$.

Pour chaque $g \in \mathcal{G}_P$ on définit le nombre de régimes comme

$$p(g) = min\{p \in \{1, ..., P\},\ g \in \mathcal{G}_p\}$$

et on introduit la nouvelle fonction de coût

$$\widetilde{l}_n(g) = \sum_{t=2}^{n} log\, g(y_{t-1}, y_t) = \sum_{t=2}^{n} log \left(\sum_{i=1}^{p} \pi_i f_i(y_t - F_i(y_{t-1})) \right)$$

Cette fonction de coût peut être vue comme une approximation de la vraisemblance marginale. On rappelle que pour un ensemble d'observations non nécessairement indépendantes $\{y_1, ..., y_n\}$ et une densité g, la log-vraisemblance marginale est définie par

$$l_n(g) = \sum_{t=1}^{n} log\, g(y_t)$$

Dans notre cas, pour chaque y_t, on conditionne par rapport à y_{t-1} et on approxime $\mathbb{P}(X_t = i \mid y_{t-1})$ par la probabilité invariante. On s'attendrait donc à ce que $\widetilde{l}_n(g)$ soit maximisé par la "vraie densité"

$$f(y_1, y_2) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0 \left(y_2 - F_i^0(y_1) \right)$$

Vérifions maintenant si $\widetilde{l}_n(g)$ remplit les conditions pour être une fonction de contraste et notamment si elle atteint son maximum en f. Pour cela, soit (X, Y_1, Y_2) une variable générique ayant pour distribution la mesure stationnaire de la chaîne de Markov étendue (X_k, Y_{k-1}, Y_k). On aura alors :

$$\mathbb{E}[ln(g) - ln(f)] = \sum_{i=1}^{p_0} \mathbb{P}(X = i)\, \mathbb{E}\left[ln\frac{g}{f} \mid X = i\right] =$$

$$= \sum_{i=1}^{p_0} \pi_i^0 \int_{y_1, y_2 \in \mathbb{R}} ln\left(\frac{\sum_{j=1}^{p} \pi_j f_j(y_2 - F_j(y_1))}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0\left(y_2 - F_j^0(y_1)\right)}\right) f_i^0\left(y_2 - F_i^0(y_1)\right) \mu_i(y_1)\, dy_1 dy_2$$

où $\mu_i(y_1)$ est la mesure stationnaire de Y_1 conditionnellement à $X = i$ et finalement, par le théorème de Fubini, on obtient :

$$\mathbb{E}[ln(g) - ln(f)] =$$

$$= \int_{y_1, y_2 \in \mathbb{R}} ln\left(\frac{\sum_{j=1}^{p} \pi_j f_j(y_2 - F_j(y_1))}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0\left(y_2 - F_j^0(y_1)\right)}\right) \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0(y_1)\right) \mu_i(y_1)\, dy_1 dy_2$$

L'inégalité de Jensen permet de voir immédiatement que le dernier terme est négatif dans chacune des situations suivantes :

- $\mu_i(y_1) = \mu(y_1)$ pour chaque $i \in \{1, ..., p_0\}$, ce qui correspond au cas des changements de régime indépendants traité aux sections précédentes
- $F_j(y_1)$ et $F_i^0(y_1)$ sont des fonctions constantes pour chaque $j \in \{1, ..., p\}$ et $i \in \{1, ..., p_0\}$, mais ceci revient au cas des chaînes de Markov cachées déjà étudié dans Gassiat (2002).

Cependant, dans le cas général, il n'y a pas de raison pour que le signe de la dernière intégrale soit négatif. Des résultats sur des simulations qu'on verra plus tard dans la section consacrée aux résultats numériques montrent que l'estimateur de vraisemblance pénalisée pour le nombre de régimes $\widehat{p}$ diverge quand le vrai modèle est, par exemple, un autorégressif à deux états et à changements de régime markoviens.

Ceci signifie qu'une généralisation de la "vraisemblance marginale" n'a pas les "bonnes" propriétés pour être une fonction de contraste et que le problème de l'estimation du nombre de régimes pour un modèle autorégressif n'est résolu que partiellement dans ce document. On est arrivé à construire un estimateur consistent du nombre de régimes dans le cadre des changements indépendants, mais pour les changements markoviens la question reste toujours ouverte.

Utiliser la vraisemblance exacte pourrait être une piste possible de recherche, quelques résultats dans ce sens ayant déjà été démontrés. Gassiat and Keribin (2000) ont montré que la statistique du test du rapport de vraisemblance diverge dans le cas particulier des

chaînes de Markov cachées, mais la consistance d'un estimateur de vraisemblance pénalisée a été prouvée sous de bonnes hypothèses. D'un autre côté, Leroux (1992b), Ryden (1995) et Francq, Roussignol et Zakoian (2001) ont déjà démontré qu'un critère de vraisemblance pénalisée ne sous-estimait pas le vrai nombre de régimes dans le cas des mélanges, des chaînes de Markov cachées et, respectivement, des modèles GARCH dont les coefficients dépendent des états d'une chaîne de Markov non-observée. Etendre ces résultats aux modèles autorégressifs à changements markoviens est assez immédiat, la partie difficile qui reste à résoudre est de montrer que le même critère ne surestime pas le vrai nombre de régimes.

5.6 Résultats numériques

Une fois le résultat théorique vérifié dans le cadre des modèles autorégressifs à changements de régimes indépendants, on s'intéresse à quelques exemples numériques pour l'illustrer. Trois questions seront d'intérêt dans cette section : la vitesse de convergence de l'estimateur du nombre de régimes puisqu'on ne l'a pas calculée théoriquement, la stabilité des algorithmes proposés et l'influence du terme de pénalité. A cause du temps de calcul assez élevé, seulement la pénalité correspondant au critère BIC a été considérée, mais d'autres possibilités pourraient être étudiées dans le futur.

Cette partie est organisée comme suit : dans un premier temps, on décrit un algorithme de type EM pouvant être utilisé dans le cas des changements indépendants. Des algorithmes de descente de gradient de type Newton sont présentés ensuite pour les fonctions de coût introduites à la section 5.5. Finalement, des résultats numériques sur des données simulées illustrent la stabilité et la convergence de l'estimateur dans le cas des changements de régime indépendants, ainsi que la divergence dans le cas des changements markoviens.

5.6.1 Algorithme EM

Les algorithmes de type EM (Expectation/Maximization) sont particulièrement adaptés aux calculs des estimateurs de maximum de vraisemblance pour des données incomplètes dont les modèles autorégressifs à changements de régime font partie. Les arguments qui maximisent la vraisemblance sont calculés de manière itérative, une description détaillée de l'algorithme étant disponible dans Dempster, Laird et Rubin (1977) ou Redner et Walker (1984). On présente ici, de manière très succinte, le principe et les pas de l'algorithme EM pour le cas particulier des modèles autorégressifs à changements de régime indépendants.

Selon le modèle (5.1), on dispose d'un échantillon observé $y_1^n = \{y_1, ..., y_n\}$ de la série $(Y_t)_{t\in\mathbb{Z}}$, chaque observation y_t dépendant de y_{t-1} et d'un processus caché i.i.d. (X_t) à valeurs dans l'espace d'états discret $\{1, ..., p_0\}$. Pour une borne supérieure de p_0 fixée, $P > 0$, l'estimateur du nombre de régimes $\widehat{p}$ se calcule donc comme l'argument $p \in \{1, ..., P\}$ qui maximise

$$T_n(p) = sup_{g\in\mathcal{G}_p} l_n(g) - a_n(p)$$

où

$$l_n(g) = \sum_{t=2}^{n} log\, g(y_t \mid y_{t-1})$$

et

$$\mathcal{G}_p = \left\{ g \mid \; g(y_2 \mid y_1) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)),\ \pi_i \geq 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

A p fixé, la partie difficile consiste à maximiser la fonction $l_n(g)$ pour $g \in \mathcal{G}_p$. Ici, on se place dans le cadre décrit à la section 5.4.1, c'est-à-dire que les fonctions de régression sont linéaires et le bruit est gaussien centré : pour chaque $i \in \{1, ..., p\}$, $F_i(y) = a_i y + b_i$ et $f_i \sim \mathcal{N}(0, \sigma_i^2)$, avec $\theta_i = (a_i, b_i, \sigma_i) \in E$, $E \subset \mathbb{R}^2 \times \mathbb{R}_+^\star$ compact. Soit $\Phi = \left(\theta = (\theta_i)_{1 \leq i \leq p}, \pi = (\pi_i)_{1 \leq i \leq p}\right)$ le paramètre global correspondant à g.

Dans ce cadre, la vraisemblance conditionnelle à y_1 des données incomplètes ou la vraisemblance marginale des observations y_1^n conditionnellement à $y_1$est

$$L(y_1^n \mid \Phi) = \prod_{t=2}^{n} g(y_t \mid y_{t-1}) = \prod_{t=2}^{n} \left(\sum_{i=1}^{p} \pi_i f_i(y_t - F_i(y_{t-1})) \right)$$

Supposant qu'on connait aussi les valeurs du processus caché $x_1^n = \{x_1, ..., x_n\}$, la vraisemblance marginale des données complètes conditionnellement aux observations initiales y_1 et x_1 s'écrit :

$$L(y_1^n, x_1^n \mid \Phi) = \prod_{t=2}^{n} \pi_{x_t} f_{x_t}(y_t - F_{x_t}(y_{t-1}))$$

et donc, en utilisant la formule de Bayes, la vraisemblance des variables non-observées conditionnellement aux variables observées est

$$L(x_1^n \mid y_1^n, \Phi) = \prod_{t=2}^{n} \frac{\pi_{x_t} f_{x_t}(y_t - F_{x_t}(y_{t-1}))}{\sum_{i=1}^{p} \pi_i f_i(y_t - F_i(y_{t-1}))}$$

Avec ces notations, on peut maintenant écrire l'algorithme EM qui se déroule de manière itérative selon les étapes suivantes :

- **Pas 1 (Initialisation)** : Poser $k = 0$ et choisir $\Phi^{(0)}$
- **Pas 2 (Expectation)** : Poser $\Phi^\star = \Phi^{(k)}$ et calculer l'espérance de la log-vraisemblance des données complètes conditionnellement à la distribution a posteriori des observations :

$$Q(\Phi \mid \Phi^\star) = \mathbb{E}(log\, L(x_1^n, y_1^n \mid \Phi) \mid y_1^n, \Phi^\star) =$$

$$= \mathbb{E}\left(\sum_{t=2}^{n} log\left(\pi_{x_t} f_{x_t}\left(y_t - F_{x_t}\left(y_{t-1}\right)\right)\right) \mid y_1^n, \Phi^\star\right) =$$

$$= \sum_{i=1}^{p}\sum_{t=2}^{n} log\left(\pi_i f_i\left(y_t - F_i\left(y_{t-1}\right)\right)\right) \frac{\pi_i^\star f_i^\star\left(y_t - F_i^\star\left(y_{t-1}\right)\right)}{\sum_{j=1}^{p} \pi_j^\star f_j^\star\left(y_t - F_j^\star\left(y_{t-1}\right)\right)}$$

- **Pas 3 (Maximization)** : Trouver $\widehat{\theta} = argmax\, Q\left(\Phi \mid \Phi^\star\right)$. On obtient les formules explicites, pour $i \in \{1, ..., p\}$:

$$\widehat{\pi}_i = \frac{1}{n-1}\sum_{t=2}^{n} h_i^\star\left(y_t, y_{t-1}\right)$$

$$\widehat{a}_i = \frac{\sum_{t=2}^{n}\sum_{t'=2}^{n} y_t h_i^\star\left(y_t, y_{t-1}\right) h_i^\star\left(y_{t'}, y_{t'-1}\right)\left(y_{t-1} - y_{t'-1}\right)}{\sum_{t=2}^{n}\sum_{t'=2}^{n} y_{t-1} h_i^\star\left(y_t, y_{t-1}\right) h_i^\star\left(y_{t'}, y_{t'-1}\right)\left(y_{t-1} - y_{t'-1}\right)}$$

$$\widehat{b}_i = \frac{\sum_{t=2}^{n}\sum_{t'=2}^{n} y_t y_{t'-1} h_i^\star\left(y_t, y_{t-1}\right) h_i^\star\left(y_{t'}, y_{t'-1}\right)\left(y_{t-1} - y_{t'-1}\right)}{\sum_{t=2}^{n}\sum_{t'=2}^{n} y_{t-1} h_i^\star\left(y_t, y_{t-1}\right) h_i^\star\left(y_{t'}, y_{t'-1}\right)\left(y_{t'-1} - y_{t-1}\right)}$$

$$\widehat{\sigma}_i = \frac{\sum_{t=2}^{n}\left(y_t - \widehat{a}_i y_{t-1} - \widehat{b}_i\right)^2 h_i^\star\left(y_t, y_{t-1}\right)}{\sum_{t=2}^{n} h_i^\star\left(y_t, y_{t-1}\right)}$$

avec la notation

$$h_i^\star\left(y_t, y_{t-1}\right) = \frac{\pi_i^\star f_i^\star\left(y_t - F_i^\star\left(y_{t-1}\right)\right)}{\sum_{j=1}^{p} \pi_j^\star f_j^\star\left(y_t - F_j^\star\left(y_{t-1}\right)\right)}$$

- **Pas 4** : Poser $\Phi^{(k)} = \widehat{\Phi}$ et recommencer au Pas 2 jusqu'à ce qu'un critère d'arrêt soit satisfait.

En utilisant l'inégalité de Jensen, on peut immédiatement montrer que l'algorithme augmente la valeur de la log-vraisemblance à chaque itération et converge vers un maximum local. Afin d'éviter de tomber dans un maximum de mauvaise qualité, la procedure est initialisée plusieurs fois avec des valeurs différentes : dans notre cas, dix initialisations ont fourni des résultats satisfaisants. Le critère d'arrêt est appliqué soit quand il n'y a plus d'amélioration dans la valeur de la vraisemblance, soit quand un nombre maximal d'itérations a été atteint. Il faut, tout de même, que le nombre d'itérations soit assez élevé , la vitesse de convergence de l'algorithme étant assez faible.

5.6.2 Algorithmes de type Newton

Utiliser un algorithme EM pour optimiser la fonction de coût introduite à la section 5.5.2 n'est pas possible, car cette fonction ne représente une vraisemblance que dans le cas particulier des changements de régime indépendants. On a donc choisi d'implémenter aussi des algorithmes plus classiques d'optimisation de type descente de gradient.

Comme dans la section 5.6.1, on dispose d'un échantillon observé $y_1^n = \{y_1, ..., y_n\}$ de la série $(Y_t)_{t\in\mathbb{Z}}$, chaque observation y_t dépendant cette fois-ci de y_{t-1} et d'une chaîne de Markov cachée (X_t) à valeurs dans l'espace d'états discret $\{1, ..., p_0\}$. Pour une borne supérieure de p_0 fixée, $P > 0$, l'estimateur du nombre de régimes $\widehat{p}$ se calcule donc comme l'argument $p \in \{1, ..., P\}$ qui maximise

$$T_n(p) = sup_{g\in\mathcal{G}_p} \widetilde{l}_n(g) - a_n(p)$$

où

$$\widetilde{l}_n(g) = \sum_{t=2}^{n} log\, g(y_{t-1}, y_t)$$

et

$$\mathcal{G}_p = \left\{ g \mid g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)),\ \pi_i \geq 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

De plus la classe de fonctions $\mathcal{G}_p$ vérifie les hypothèses suivantes : pour chaque $i \in \{1, ..., p\}$, $F_i(y) = a_i y + b_i$ et $f_i \sim \mathcal{N}(0, \sigma_i^2)$, avec $\theta_i = (a_i, b_i, \sigma_i) \in \mathbb{R}^2 \times \mathbb{R}_+^\star$.

Pour chaque $p \in \{1, ..., P\}$, il faut donc trouver les paramètres qui maximisent $\widetilde{l}_n(g)$. Les algorithmes de descente de gradient étant des méthodes d'optimisation sans contraintes, on a choisi de reparamétrer π_i et σ_i par :

- pour chaque $i \in \{1, ..., p\}$, on considère $\alpha_i = ln \frac{\pi_i}{\pi_p}$. On a donc $\alpha_p = 0$ et les π_i s'écrivent en fonction de α_i comme suit :

$$\pi_i = \frac{exp(\alpha_i)}{1 + exp(\alpha_1) + ... + exp(\alpha_p)}$$

- pour chaque $i \in \{1, ..., p\}$, on fait la notation $\beta_i = ln\, \sigma_i$

Soit donc $\Phi = \left(a = (a_i)_{1\leq i\leq p}, b = (b_i)_{1\leq i\leq p}, \beta = (\beta_i)_{1\leq i\leq p}, \alpha = (\alpha_i)_{1\leq i\leq p}\right)$ le nouveau paramètre global correspondant au modèle donné par la fonction $g \in \mathcal{G}_p$.

La fonction de coût devient alors $\widetilde{l}_n(g) := \widetilde{l}_n(\Phi)$ où

$$\widetilde{l}_n(\Phi) = \sum_{t=2}^{n} log \left(\sum_{i=1}^{p} \frac{exp(\alpha_i)}{1 + exp(\alpha_1) + ... + exp(\alpha_p)} \frac{1}{\sqrt{2\pi} exp(\beta_i)} exp\left(-\frac{(y_t - F_i(y_{t-1}))^2}{2exp(2\beta_i)} \right) \right)$$

L'idée des ces algorithmes est de minimiser la fonction $-\widetilde{l}_n(\Phi)$ de manière itérative en modifiant à chaque itération la direction de descente. On aura donc une suite de paramètres $(\Phi_k)_{k\in\mathbb{N}^\star}$ telle que

$$\Phi_{k+1} = \Phi_k - \gamma_k H_k \nabla \widetilde{l}_n(\Phi_k)$$

où H_k est une matrice définie positive qui approxime l'inverse de la matrice hessienne calculée au point Φ_k et γ_k est un pas qui assure que la fonction à minimiser décroit à chaque itération. Plusieurs méthodes sont disponibles, selon le choix de H_k et γ_k. Pour la fonction $-\widetilde{l}_n(\Phi)$, on a codé l'algorithme du gradient congujué de Polack et Ribière et l'algorithme quasi-newtonien de Broyden, Fletcher, Goldfarb et Shanno (BFGS). Ce sont des algorithmes classiques, dont on peut trouver le détail dans Press (1992). Comme pour la méthode EM, plusieurs initialisations sont nécessaires pour éviter de tomber dans des minima locaux de mauvaise qualité.

Remarquons aussi que $\widetilde{l}_n(\Phi)$ représente la vraie vraisemblance dans le cas particulier des changements de régimes indépendants et donc les deux algorithmes présentés donneront les mêmes résultats dans ce cas. L'avantage en temps de calcul reste cependant du côté des algorithmes de descente de gradient qui ont une convergence quadratique par rapport à la méthode EM qui ne converge que linéairement.

5.6.3 Résultats empiriques - stabilité et convergence

Après avoir décrit les algorithmes d'optimisation de la fonction de coût pour les modèles à changements de régime, illustrons maintenant les résultats théoriques présentés dans ce chapitre par quelques exemples numériques. Le but des simulations était d'étudier la vitesse de convergence de l'estimateur, ainsi que l'influence du terme de pénalité. Finallement, comme le temps de calcul est assez important (environ une heure sur un processeur 512 Mhz pour tester si un échantillon de 2000 observations provient d'un modèle à un, deux ou trois régimes), on a considéré uniquement le terme de pénalité du critère BIC (Schwarz (1978)) :

$$a_n(p) = \frac{1}{2} k \, log(n),$$

où k est le nombre de paramètres d'un modèle à p régimes et n est la taille de l'échantillon.

Tout d'abord, on a étudié la convergence de l'estimateur de vraisemblance pénalisée pour les modèles autorégressifs à changements de régime indépendants. On a considéré que le vrai modèle est à deux régimes et sa densité conditionnelle est

$$f(y_2 \mid y_1) = \pi_1^0 f_1^0\left(y_2 - F_1^0(y_1)\right) + \left(1-\pi_1^0\right) f_2^0\left(y_2 - F_2^0(y_1)\right)$$

avec $F_i^0(y_1) = a_i^0 y_1 + b_i^0$ et $f_i^0 \sim \mathcal{N}\left(0, \left(\sigma_i^0\right)^2\right)$, pour $i \in \{1,2\}$.

Pour chaque modèle, on simule 20 échantillons de tailles $n = 200$, 500, 1000, 1500 et 2000 et on fixe la borne supérieure pour le nombre de régimes à $P = 3$. La vraisemblance

est maximisée via un algorithme EM et, pour éviter les maxima de mauvaise qualité, dix initialisations différentes des paramètres ont été considerées. L'algorithme s'arrête soit quand il n'y a plus d'amélioration dans la valeur de la vraisemblance, soit quand le nombre d'itérations, fixé ici à 200 pour des raisons de temps de calcul raisonnable, est atteint.

Pour tous les modèles, les écarts-type sont égaux $\sigma_1^0 = \sigma_2^0 = 0.5$. Les résultats sont résumés dans les Tableaux 5.1 et 5.2. Dans le Tableau 5.1, on considère des termes constants assez éloignés, $b_1^0 = 1$, $b_2^0 = -1$, et on fait varier les coefficients des retards, $a_1^0, a_2^0 \in \{0.1, 0.5, 0.9\}$, et les poids des deux régimes, $\pi_1^0 \in \{0.5, 0.7, 0.9\}$. On remarque que la convergence est atteinte rapidement pour ces exemples, 500 observations étant suffisantes en général.

En ce qui concerne les exemples du Tableau 5.2, on a gardé les mêmes valeurs pour σ_1^0, σ_2^0, a_1^0, a_2^0 et π_1^0, mais on a rapproché les constantes, $b_1^0 = 0.5$ et $b_2^0 = -0.5$. Dans ce cas, discriminer entre les deux régimes est une tâche beaucoup plus difficile et la convergence n'est obtenue en général que sur des échantillons supérieurs à 2000 observations.

Dans la dernière partie concernant les exemples numériques, on s'intéresse à l'illustration de la divergence de la fonction de coût $\widetilde{l}_n(g)$ introduite à la section 5.5 quand le vrai modèle est un autorégressif d'ordre un avec changements de régimes markoviens. Pour cela, on considère de vrais modèles à deux régimes avec trois matrices de transitions possibles : $M_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $M_2 = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$ et $M_3 = \begin{pmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{pmatrix}$. La première matrice de transition correspond en fait à un modèle avec des changements de régime indépendants, ce qui nous ramène au cas traité précédemment, tandis que la deuxième engendre des régimes assez persistants.

En ce qui concerne les fonctions de régression, on a deux possibilités pour chacune des trois matrices de transition : soit des fonctions linéaires, $F_1^0(y) = 0.8y - 1$ et $F_2^0(y) = 0.3y + 1$, soit des fonctions constantes, $F_1^0(y) = -1$ et $F_2^0(y) = 1$. Pour tous les modèles considérés, les deux densités conditionnelles sont des gaussiennes centrées avec les écarts-type $\sigma_1^0 = \sigma_2^0 = 0.5$.

Pour optimiser la fonction de coût on a utilisé l'algorithme du gradient conjugué de Polack-Ribière. L'execution est arrêtée soit quand le gradient devient nul, soit quand le nombre maximum d'itérations, fixé à 250, est atteint. De même que pour EM, les paramètres sont initialisés avec des valeurs différentes dix fois pour éviter des points extrèmes de mauvaise qualité. Les résultats des simulations sont résumés dans le Tableau 5.3.

Les coefficients dans les fonctions de régression étant assez éloignés, la convergence ou, respectivement, la divergence sont atteintes rapidement. Dans le cas des changements de régime indépendants (matrice de transition M_1) ou des fonctions de régression constantes (cas des chaînes de Markov cachées), l'estimateur trouve le bon nombre de régimes. Cependant, dans le cas général des modèle autorégressifs à changements markoviens, le nombre de régimes est très rapidement surestimé.

		M_1			M_2			M_3		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$F_1^0(y)=0.8y-1$	200	0	20	0	**0**	**15**	**5**	**0**	**17**	**3**
$F_2^0(y)=0.3y+1$	500	0	20	0	**0**	**17**	**3**	**0**	**8**	**12**
	1000	0	20	0	**0**	**6**	**14**	**0**	**4**	**16**
	1500	0	20	0	**0**	**1**	**19**	**0**	**5**	**15**
	2000	0	20	0	**0**	**1**	**19**	**0**	**5**	**15**
$F_1^0(y)=-1$	200	0	20	0	0	20	0	0	20	0
$F_2^0(y)=1$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0

TAB. 5.3 – Résultats sur des modèles à changements de régime markoviens. Pour chaque choix des paramètres, on indique le nombre de fois où l'algorithme sélectionne un, deux ou trois régimes (sur 20 simulations)

5.7 Conclusion

Nous avons montré la consistence faible de l'estimateur de maximum de vraisemblance pénalisée du nombre de régimes pour le modèle (5.1), sous des conditions qui tiennent principalement en compte la complexité de la classe de fonctions score généralisées. Les hypothèses du résultat principal ont été vérifiées ensuite dans le cas particulier des modèles à changements de régime indépendants, avec des fonctions de régression linéaires et bruit gaussien. Des exemples sur des simulations ont illustré la vitesse de convergence de l'estimateur et ont montré que choisir comme pénalité le terme du critère BIC donne d'assez bons résultats.

Finalement, on a montré que généraliser la vraisemblance marginale aux modèles à changements de régime markoviens ne fournit pas de fonction de contraste. Le problème reste donc ouvert dans ce cas et l'utilisation de la vraisemblance exacte ou d'un autre contraste bien adapté pourrait permettre de donner une réponse.

A titre d'exemple, on a regardé si de manière empirique le critère BIC sélectionne le bon nombre de régimes. Les données utilisées pour les simulations sont celles des deux cas de divergence du Tableau 5.3. La log -vraisemblance a été maximisée directement par optimisation différentielle via un algorithme proposé par Rynkiewicz (2000). Les nouveaux résultats sont présentés dans le Tableau 5.4 et on remarque que cette fois on arrive à sélectionner le bon nombre de régimes. En conséquence, le critère BIC et l'utilisation de la vraisemblance exacte semblent bien adaptés empiriquement pour les modèles autorégressifs à changements de régime markoviens, il reste à faire l'important travail théorique correspondant.

		M_2			M_3		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$F_1^0(y)=0.8y-1$	200	0	16	4	0	15	5
$F_2^0(y)=0.3y+1$	500	0	16	4	0	19	1
	1000	0	17	3	0	19	1
	1500	0	18	2	0	19	1
	2000	0	19	1	0	20	0

TAB. 5.4 – Résultats sur des modèles à changements de régime markoviens avec la vraisemblance exacte et le critère BIC. Pour chaque choix des paramètres, on indique le nombre de fois où l'algorithme sélectionne un, deux ou trois régimes (sur 20 simulations)

	π_1^0	0.5			0.7			0.9		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$a_1^0=0.1$	200	0	20	0	0	20	0	0	18	2
$a_2^0=0.1$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.1$	200	0	20	0	0	19	1	1	19	0
$a_2^0=0.5$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.1$	200	0	20	0	0	20	0	4	16	0
$a_2^0=0.9$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.5$	200	0	19	1	0	18	2	0	20	0
$a_2^0=0.5$	500	0	20	0	0	20	0	0	18	2
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.5$	200	0	19	1	0	20	0	11	9	0
$a_2^0=0.9$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.9$	200	0	20	0	0	20	0	0	16	4
$a_2^0=0.9$	500	0	20	0	0	20	0	0	20	0
	1000	0	19	1	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0

TAB. 5.1 – Résultats pour $b_1^0=1$, $b_2^0=-1$, $\sigma_1^0=\sigma_2^0=0.5$. Pour chaque choix de a_1^0, a_2^0 et π_1^0, on indique le nombre de fois où l'algorithme sélectionne un, deux ou trois régimes (sur 20 simulations)

	π_1^0	0.5			0.7			0.9		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$a_1^0=0.1$	200	20	0	0	20	0	0	20	0	0
$a_2^0=0.1$	500	18	2	0	18	2	0	20	0	0
	1000	14	6	0	9	11	0	11	9	0
	1500	6	14	0	4	16	0	5	15	0
	2000	5	15	0	0	20	0	1	19	0
$a_1^0=0.1$	200	12	8	0	13	7	0	20	0	0
$a_2^0=0.5$	500	11	19	0	6	14	0	18	2	0
	1000	0	20	0	1	19	0	14	6	0
	1500	0	20	0	0	20	0	8	12	0
	2000	0	20	0	0	20	0	7	13	0
$a_1^0=0.1$	200	0	20	0	4	16	0	17	3	0
$a_2^0=0.9$	500	0	20	0	0	20	0	9	11	0
	1000	0	20	0	0	20	0	9	11	0
	1500	0	20	0	0	20	0	4	16	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.5$	200	18	2	0	20	0	0	19	1	0
$a_2^0=0.5$	500	20	0	0	19	1	0	19	1	0
	1000	14	6	0	13	7	0	10	10	0
	1500	9	11	0	5	15	0	5	15	0
	2000	3	17	0	0	20	0	3	17	0
$a_1^0=0.5$	200	9	11	0	11	9	0	20	0	0
$a_2^0=0.9$	500	0	20	0	7	13	0	19	1	0
	1000	0	20	0	0	20	0	19	1	0
	1500	0	20	0	0	20	0	18	2	0
	2000	0	20	0	0	20	0	14	6	0
$a_1^0=0.9$	200	20	0	0	19	1	0	19	1	0
$a_2^0=0.9$	500	20	0	0	18	2	0	17	3	0
	1000	14	6	0	7	13	0	11	9	0
	1500	7	13	0	5	15	0	3	17	0
	2000	6	14	0	0	20	0	0	20	0

TAB. 5.2 – Résultats pour $b_1^0=0.5$, $b_2^0=-0.5$, $\sigma_1^0=\sigma_2^0=0.5$. Pour chaque choix de a_1^0, a_2^0 et π_1^0, on indique le nombre de fois où l'algorithme sélectionne un, deux ou trois régimes (sur 20 simulations)

Chapitre 6

Une méthode empirique pour calculer le nombre d'états dans un modèle à changements de régime

Dans ce chapitre, le problème du choix du nombre d'états dans un modèle autorégressif à changements de régime est abordé par une méthode différente de celle présentée au chapitre précédent et qui permet une reformulation de la question. Il s'agit d'une méthode empirique qui provient de l'analyse des données et plus particulièrement de la classification non-supervisée. On voit dans la suite comment on peut transformer le problème du choix du nombre de régimes en un problème de classification en utilisant des cartes de Kohonen et une nouvelle dispersion dans la classification hiérarchique. Des exemples sur des données réelles et simulées illustrent ensuite la méthode qui a l'avantage de pouvoir être appliquée à tout type de modèle à changements de régime (modèles à seuil ou avec des chaînes de Markov cachées), mais qui reste néanmoins une méthode descriptive et donc fortement dépendante du modèle et de l'échantillon.

6.1 Reformulation du problème du choix du nombre de régimes en problème de classification

Le problème du choix du "vrai" nombre de régimes peut se réécrire comme un problème de classification en transformant l'échantillon observé par une fenêtre glissante. Soit $y_1^n = \{y_1, ..., y_n\}$ une observation d'un n-échantillon de la série $(Y_t)_{t\in\mathbb{Z}}$. On suppose que (Y_t) suit le vrai modèle :

$$Y_t = F_{X_t}^0 \left(Y_{t-1}, ..., Y_{t-l}\right) + \varepsilon_t \left(X_t\right) \qquad (6.1)$$

où

- le nombre de retards $l \geq 1$ est connu
- (X_t) est une suite non-observée de variables aléatoires à valeurs dans un espace fini $\{1, ..., p_0\}$

- pour chaque $i = 1, ..., p_0$, $F_i^0(y_t) = a_{i,0}^0 + a_{i,1}^0 y_{t-1} + ... + a_{i,l}^0 y_{t-l}$ est une fonction autorégressive linéaire d'ordre l qui dépend du paramètre $\theta_i^0 = \left(a_{i,0}^0, a_{i,1}^0, ..., a_{i,l}^0\right) \in \mathbb{R}^{l+1}$
- pour chaque $i = 1, ..., p_0$, $(\varepsilon_t(i))_{t\in\mathbb{Z}}$ est un bruit i.i.d., les suites $(\varepsilon_t(i))_{1\leq i\leq N}$ étant indépendantes.

Le nombre de retards étant connu, le vecteur des observations de dimension n peut s'écrire comme un tableau de données en lui appliquant une fenêtre glissante de pas l. On obtient alors

$$\mathbb{Y} = \{y_t, y_{t-1}, ..., y_{t-l}\}_{t=l+1,...,n} \in \mathbb{R}^{(n-l)\times(l+1)}$$

Trouver le nombre de régimes revient donc à trouver le nombre d'hyperplans de régression qui s'ajustent le mieux aux données $\mathbb{Y}$.

Pour illustrer le problème, un exemple trivial est présenté dans la Figure 6.1. Il est évident qu'ajuster une seule droite de régression aux données n'est pas le bon choix. En revanche, si on suppose qu'on est arrivé à séparer les données en deux classes et qu'on ajuste une droite de régression dans chaque classe, le nouveau modèle semble mieux adapté aux données. Le critère de comparaison le plus naturel est de regarder les sommes des résidus au carré du premier modèle et le total des sommes des résidus au carré dans les deux classes du deuxième. Une première question serait donc comment classer les observations de la manière la plus pertinente ?

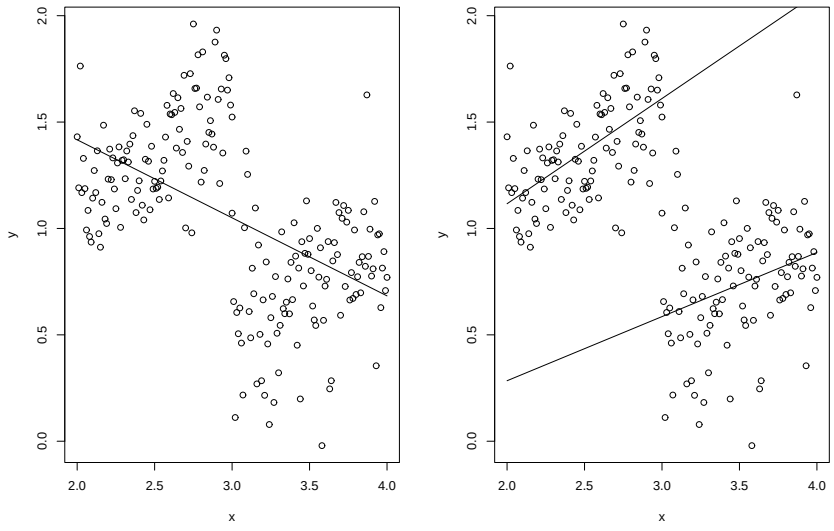

FIG. 6.1 – Ajustement d'un modèle à deux régimes

Remarquons tout d'abord qu'une classification qui conduit à une bonne séparation des observations est fortement liée au modèle à partir duquel les données ont été obtenues. Même si le nombre de méthodes de classification est très grand, des centaines d'algorithmes étant

disponibles, la plupart d'entre elles reposent plus sur l'échantillon observé que sur le modèle qui l'a généré. Cependant, l'idée de classer des données en utilisant des conditions tirées du vrai modèle n'est pas nouvelle dans la littérature. Le concept de classification probabiliste est présent dans quelques méthodes comme les méthodes à partition fixée ("fixed-partition methods"), un exemple étant la classification de type régression ("regression-type clustering"), la classification en composantes principales, la méthode des directions révélatrices ("projection pursuit") ou celle des supports convexes, les modèles de mélange ou les clusters à "haute densité" ("high-density clusters").

La méthode proposée dans ce chapitre est assez proche du "regression-type clustering" qui est basé sur l'idée suivante : identifier les hyperplans qui s'ajustent le mieux aux données suivant un critère basé soit sur la minimisation de la somme des résidus au carré, soit sur la maximisation de la somme des projections orthogonales des observations. Dans le premier cas, il s'agit d'une généralisation de la régression linéaire, dans le deuxième d'une généralisation de l'analyse en composantes principales (voir la Figure 6.2). La méthode "regression-type clustering" a été introduite par Charles (1977) et des développements ont été proposés ultérieurement par Spath (1979), DeSarbo et Cron (1988) ou Lou, Jiang et Keng (1993). En résumé, l'algorithme est une généralisation de la méthode des centres mobiles en remplaçant les vecteurs centroïdes par des hyperplans de régression.

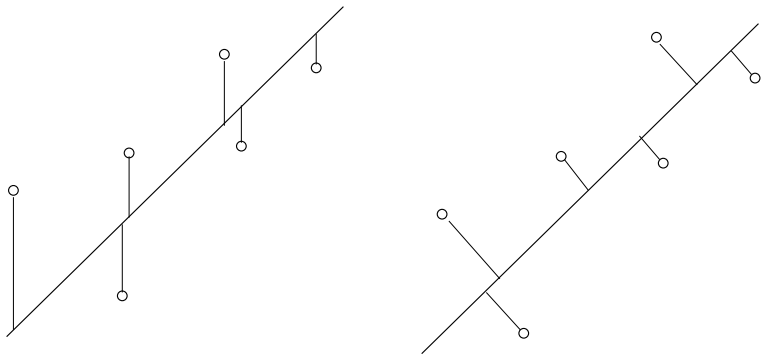

FIG. 6.2 – Régression linéaire (à gauche) et analyse en composantes principales (à droite)

Dans notre cas, on utilise la méthode suivante : on commence par regrouper les observations en clusters les plus homogènes possible, avec un nombre de clusters largement supérieur au "vrai" nombre de régimes. Ensuite, on ajuste le contenu de chaque cluster par un modèle, on calcule la somme des carrés résiduels obtenus dans chaque classe et on définit une nouvelle inertie intra-classes par la somme des carrés résiduels de chaque classe. Le dernier pas consiste à regrouper les deux clusters qui font augmenter le moins possible cette inertie et ensuite on recommence : on ajuste le contenu de chaque classe par un modèle, on calcule la somme des carrés résiduels, etc jusqu'à ce qu'on obtienne une seule classe. Si le bruit n'a pas une variance très importante, on s'attend à avoir un saut important dans les valeurs de l'inertie intra-classes dès qu'on passe du vrai nombre de régimes à un nombre plus petit. Remarquons aussi qu'une "bonne" classification initiale signifie qu'on a obtenu des classes qui contiennent des observations provenant du même régime et, en même temps, avec assez de données pour avoir de bons estimateurs des coefficients de régression.

En ce qui concerne la classification initiale, on a choisi l'algorithme de Kohonen (Kohonen (1984)) en raison de ses propriétés de visualisation, d'homogénéité des clusters et de la possibilité de fixer un grand nombre de clusters dans cette étape d'initialisation. La méthode de regroupement successif des clusters est une variante de la classification hiérarchique où l'inertie usuelle (somme des carrés des écarts aux centres des gravités des classes) est remplacée par la somme des carrés résiduels après ajustement des données sur un modèle à l'intérieur de chaque classe.

6.2 Méthodes hiérarchiques et cartes de Kohonen - rappels sur les méthodes de segmentation destinées à mettre en évidence des classes dans les données

Cette section est consacrée à un bref rappel des deux méthodes de segmentation de données utilisées dans la suite, les cartes de Kohonen et la classification hiérarchique. Les deux algorithmes font partie des méthodes de segmentation non-supervisées qui reviennent à partitionner un ensemble de points sur la base de leur similarité sans avoir d'information "a priori" sur le "vrai" nombre de classes ou sur l'etiquetage de chaque observation. Dans la classification hiérarchique, le nombre de classes n'est pas fixé à l'avance, il est déterminé "a posteriori" selon un critère spécifié, tandis que dans les cartes de Kohonen, le nombre de classes est un paramètre fixé dès le début de l'algorithme qui est choisi supérieur au "vrai" nombre de classes.

6.2.1 Classification hiérarchique, méthode de Ward

La classification hiérarchique correspond à quoi, vu le titre, on s'attend de manière intuitive. Elle produit une représentation hiérarchique des données dans laquelle, par exemple, une classe au niveau k est obtenue en regroupant deux classes du niveau $k-1$. On obtient une structure arborescente appelée aussi dendrogramme avec, sur le niveau le plus bas, un nombre de classes égal au nombre d'observations et contenant chacune un seul point et, sur le niveau le plus haut, une seule classe qui contient toutes les données.

Ici, on s'intéresse aux méthodes agglomératives : on commence au niveau le plus bas, avec N observations et donc N classes, et à chaque pas on regroupe les deux classes les plus proches au sens d'une distance ou d'une dissimilarité qui reste à définir, jusqu'à ce qu'il reste une seule classe.

La première étape consiste à calculer toutes les distances deux à deux de toutes les observations et regrouper les deux plus proches pour former une nouvelle "classe". Il s'agit alors de définir une distance entre ce groupe et les observations restantes, pour remplir à nouveau la table de toutes les distances et continuer les regroupements successifs. Il est donc nécessaire de définir la distance d'une observation à une classe et la distance entre deux classes.

Il existe plusieurs méthodes pour définir des distances entre les classes, mais celle qu'on développe dans la suite est celle de Ward définie ci-dessous. On rapelle que si $X = (x_i)_{i=1,\dots,N}$

est un nuage de points dans $\mathbb{R}^p$ et si on suppose que chaque point à la même masse $\frac{1}{N}$, on définit l'inertie totale du nuage par

$$I = \sum_{i=1}^{N} \|x_i - g\|^2$$

où $g = \frac{1}{N}\sum_{i=1}^{N} x_i$ est le centre de gravité du nuage et $\|\cdot\|$ est la norme euclidienne.

S'il existe une partition du nuage des points notée $C_1, ..., C_s$, la relation de Huygens fournit une décomposition de l'inertie totale :

$$I = \sum_{k=1}^{s} \#C_k \|g_k - g\|^2 + \sum_{k=1}^{s} \sum_{x_i \in C_k} \|x_i - g_k\|^2 \qquad (6.2)$$

où $g_k = \frac{1}{\#C_k}\sum_{x_i \in C_k} x_i$ est le centre de gravité de la classe C_k.

Le premier terme de la relation (6.2) mesure l'inertie inter-classes I_{inter}, tandis que le deuxième est la somme des inerties à l'intérieur de chaque classe, appelé I_{intra}. Quand on regroupe deux classes, I_{intra} augmente et I_{inter} diminue. Au premier pas d'une classification hiérarchique, $I_{intra} = 0$ et $I_{inter} = I$, au dernier on a le contraire $I_{intra} = I$ et $I_{inter} = 0$. On souhaite que les variations de ces deux inerties soient les plus progressives possibles et cela conduit à choisir la distance de Ward, qui a pour propriété de conduire aux regroupements qui font le moins gagner d'inertie intra (et le moins perdre de l'inertie inter).

Pour définir cette distance, considérons le gain d'inertie inter qui résulte du regroupement de deux classes C_i et C_j. Il vaut

$$\#C_i \|g_i - g\|^2 + \#C_j \|g_j - g\|^2 - (\#C_i + \#C_j) \left\|g_{C_i \cup C_j} - g\right\|^2 \qquad (6.3)$$

où $g_{C_i \cup C_j}$ est le centre de gravité de la classe $C_i \cup C_j$ et, d'après le théorème de la médiane généralisé, on obtient

$$\frac{\#C_i \cdot \#C_j}{\#C_i + \#C_j} \|g_i - g_j\|^2$$

A chaque étape, on regroupe donc les classes (à un ou plusieurs éléments) qui réalisent le minimum de cette quantité qu'on appelle distance de Ward des classes C_i et C_j.

Remarquons que cet algorithme est aussi très sensible aux valeurs extrêmes et fournit souvent des classes déséquilibrées.

6.2.2 Cartes de Kohonen - principe de l'algorithme

Les cartes de Kohonen ou les cartes auto-organisées sont des processus itératifs, qui regroupent les observations en un nombre fini de classes en ayant comme caractéristique le

fait de préserver la topologie des données. En effet, une notion de voisinage est définie a priori sur les classes et les observations proches dans l'espace des variables se retrouvent, après classement, dans la même classe ou dans des classes voisines. Il y a plusieurs méthodes pour choisir le voisinage entre les classes et la dimension de la carte qui dépendent de la dimension intrinsèque des données, même si en général les utilisateurs préfèrent des grilles à deux dimensions.

Le principe de l'algorithme, qui a une structure itérative, est le suivant : on associe à chaque classe des vecteurs code initialisés aléatoirement dans l'espace des observations et, à chaque etape, une observation est tirée au hasard parmi les données et présentée en entrée, le vecteur code le plus proche est alors désigné comme gagnant et il est rapproché, ainsi que ses voisins, de l'observation présentée.

Soit donc $X = (x_i)_{i=1,...,N}$ un nuage de points défini dans $\mathbb{R}^p$ et soit les vecteurs codes $C_1, ..., C_s \in \mathbb{R}^p$ correspondant aux s classes de la carte. En notant $C_i(t)$, $i = 1, ..., s$, les valeurs des vecteurs codes à l'étape t et $x(t+1)$ l'observation tirée au hasard, les pas de l'algorithme à l'étape $t+1$ sont les suivants :

– on choisit la classe gagnante

$$i_0(t+1) = argmin_{i \in \{1,...,s\}} \|x(t+1) - C_i(t)\|$$

où $\|\cdot\|$ est le plus souvent la distance euclidienne

– on met à jour le gagnant ainsi que ses voisins

$$C_i(t+1) = C_i(t) + \varepsilon(t+1)\,\sigma(i_0(t+1), i)\,(x(t+1) - C_i(t))$$

où $\varepsilon(t+1)$ est un paramètre d'adaptation, constant ou décroissant et $\sigma(i,j)$ est une fonction de voisinage définie par

$$\sigma(i,j) = \begin{cases} 1, \text{ si les classes i et j sont voisines} \\ 0, \text{ sinon} \end{cases}$$

Remarque : la taille du voisinage doit décroître lentement vers zéro au cours du temps pour que l'algorithme soit convergent.

La méthodes que nous proposons pour les données provenant des modèles à changements de régime consiste à faire une classification initiale par l'algorithme de Kohonen et à pratiquer ensuite une classification hiérarchique en modifiant les définitions de l'inertie intra et de l'inertie totale.

6.3 Généralisation de la méthode de Ward, classification des données provenant des modèles à changements de régime

6.3.1 Classification initiale des données

Le but étant de trouver le meilleur ajustement et de choisir le nombre d'hyperplans de régression qui expliquent le mieux les données, on a besoin d'adapter la méthode de classification hiérarchique à ce cas. Deux points sont à noter : les classes ne sont plus caractérisées

par leur centre de gravité, mais par un hyperplan et, de plus, puisque dans chaque classe on doit estimer des coefficients de régression, on ne peut pas avoir, sur le niveau le plus bas de l'arbre de classification, de classes contenant une seule observation.

Le deuxième problème sera contourné en appliquant une classification initiale qui n'impose pas cette contrainte comme les cartes de Kohonen. On utilise ici la propriété des cartes auto-organisées de créer des classes homogènes et qui regrouperont donc des observations provenant du même régime. Cependant, si les coefficients des fonctions de régression dans les différents régimes sont proches et si le bruit est important, cette propriété peut être mise en défaut.

Une autre question qui apparaît est alors : est-ce qu'il y a assez de points dans chaque classe de la classification initiale pour estimer une régression ? Au cas où il existe des classes qui ne contiennent pas assez d'observations, leur contenu est rattaché, en utilisant la topologie de la carte, à la plus proche classe ayant suffisament de points et les classes qui restent ainsi vides sont enlevées dans la suite.

Une fois la classification initiale obtenue, reste à définir la classification hiérarchique à appliquer. Dans tout le reste de ce chapitre, on fait la convention d'appeler "clusters" les classes obtenues par une carte de Kohonen et "classes" les regroupements éventuels de deux ou plusieurs "clusters".

6.3.2 Classification hiérarchique autour des hyperplans de régression

Soit donc $y_1^n = \{y_1, ..., y_n\}$ un n-échantillon observé de la série $(Y_t)_{t\in\mathbb{Z}}$ qui vérifie le modèle (6.1) et soit

$$\mathbb{Y} = \{y_t, y_{t-1}, ..., y_{t-l}\}_{t=l+1,...,n} \in \mathbb{R}^{(n-l)\times(l+1)}$$

le tableau de données avec $n-l$ lignes et $l+1$ colonnes obtenu en appliquant une fenêtre glissante de taille l.

On suppose aussi que les données $\mathbb{Y}$ ont été classées par une carte de Kohonen bidimensionnelle (M lignes, M colonnes, M^2 clusters) et que la distance considérée est celle générée par la norme euclidienne. Les M^2 clusters obtenus sont regroupés ensuite dans un arbre hiérarchique, en utilisant une distance définie ci-dessous :

Supposons qu'on se place au niveau k de l'arbre avec les classes $C_1, ..., C_k$ et qu'on veuille passer au niveau supérieur $k-1$. Au niveau k, on définit la somme des carrés résiduelle par :

$$SSE_{intra}\,(k) = \sum_{j=1}^{k} SSE_{C_j}$$

où

$$SSE_{C_j} = \sum_{t\in C_j} \left(y_t - \widehat{a}_{C_j,0} + \widehat{a}_{C_j,1} y_{t-1} + ... + \widehat{a}_{C_j,l} y_{t-l}\right)^2$$

est la somme des carrés des résidus pour la régression estimée dans la classe C_j, $\widehat{a}_{C_j,0}$, $\widehat{a}_{C_j,1}$,...,$\widehat{a}_{C_j,l}$ étant les estimateurs des paramètres de régression dans la classe C_j pour $j = 1, ..., k$.

Si, pour passer au niveau $k-1$, les classes C_j et $C_{j\prime}$ ont été regroupées, la somme des carrés résiduelle devient :

$$SSE_{intra}^{j,j\prime}(k) = \sum_{i=1, i\neq j, i\neq j\prime}^{k} SSE_{C_i} + SSE_{C_j \cup C_{j\prime}}$$

Le principe de regroupement des classes est le même que celui de Ward : on regroupe les classes qui minimisent le gain dans la somme des carrés résiduelle. On cherche donc $j, j\prime \in \{1, ..., k\}$ qui minimisent

$$\Delta SSE_{intra}^{j,j\prime}(k, k-1) = SSE_{intra}^{j,j\prime}(k) - SSE_{intra}(k) = SSE_{C_j \cup C_{j\prime}} - SSE_{C_j} - SSE_{C_{j\prime}}$$

Cette quantité est toujours positive et si les deux classes proviennent du même régime, alors elle devrait être très proche de zéro. De plus, Toyoda (1974) a montré que si les bruits des classes C_j et $C_{j\prime}$ suivent des lois gaussiennes centrées de variances σ_j^2 et $\sigma_{j\prime}^2$, alors $\Delta SSE_{intra}^{j,j\prime}(k, k-1)$ suit approximativement une loi $\sigma^2\chi^2(l+1)$, où σ^2 est une moyenne pondérée de σ_j^2 et $\sigma_{j\prime}^2$ bien choisie.

6.3.3 Algorithme

On peut maintenant écrire les pas d'un nouvel algorithme qui réalise, en même temps, une classification des données et une modélisation dans chaque classe. On remarquera au passage que même si l'objet principal de l'étude est l'analyse des modèles autorégressifs, l'algorithme suivant peut être appliqué pour des modèles de régression à changements de régime comprenant aussi d'autres variables explicatives que le passé.

Alg.1

Pas 1 : Choisir les variables explicatives ou/et le nombre des retards l avec un critère d'information de type AIC, BIC

Pas 2 : Construire la matrice des données $\mathbb{Y}$ en utilisant une fenêtre glissante de taille l

Pas 3 : Choisir la dimension et la topologie pour la carte de Kohonen et faire une classification initiale des données en M^2 clusters. Dans cette étude on a utilisé uniquement des grilles carrées, le choix de la taille de la grille étant liée au nombre de données et au nombre de variables de régression. Les clusters devraient contenir assez de points pour pouvoir estimer une régression, mais pas trop, pour ne pas mélanger des observations provenant de régimes différents dans le même cluster. Bien sûr, il n'y a pas de réponse théorique pour ce problème et éviter le mélange des observations devient impossible quand les modèles dans chaque régime sont très proches. Ceci soulève une autre question : quand les modèles sont proches, est-ce que cela a du sens, en pratique, de considérer plusieurs régimes ? Comme règle empirique, les grilles ont été choisies telles que M^2 soit égal au nombre d'observations divisé par cinq fois le nombre de variables explicatives.

Pas 4 : Construire la classification hiérarchique des M^2 clusters.

4.1 : Poser $k = M^2$

4.2 : Estimer les k hyperplans de régression dans chaque classe

4.3 : Calculer (j_0, j_0') qui minimisent $\Delta SSE_{intra}^{j,j'}(k, k-1)$

4.4 : Regrouper les classes j_0 et j_0' et poser $k = k - 1$. Si $k \neq 1$, retourner au pas 4.2

A la dernière étape, il n'y a qu'une seule classe et une seule régression. Pour choisir le nombre de régimes, on représente graphiquement l'évolution de la somme des carrés résiduelle au cours du regroupement des classes, en s'attendant à observer un saut important lorsqu'on passe du vrai nombre de régimes à un nombre plus petit.

De plus, à chaque pas de la classification hiérarchique, des tests d'égalité des coefficients peuvent confirmer si les classes qui sont regroupées proviennent du même régime. Si, par exemple, au passage de k à $k-1$ classes on regroupe C_i et C_j, on considère les hypothèses "H_0 : C_i et C_j proviennent du même régime" contre "H_1 : C_i et C_j appartiennent à des régimes différents" et on pratique un test de Fisher sur l'égalité des coefficients de régression dans les deux classes. Utiliser ce type de tests a pourtant l'inconvénient de surestimer le nombre de régimes. En effet, si la classification initiale contient des clusters qui mélangent des observations de deux ou plusieurs régimes, le test deviendra significatif dès que ces clusters interviendront dans les regroupements de classes.

6.4 Résultats numériques

La méthode proposée étant empirique, sa validation ne peut être apportée que par des exemples numériques. On l'a testée d'abord sur plusieurs modèles simulés à changements de régime et ensuite on l'a appliquée à des données réelles.

6.4.1 Modèles à seuil de type TAR

Les modèles à seuil de type TAR ("Treshold Autoregressive models") introduits par Tong (1983) font partie des premiers modèles non-linéaires à changements de régime. L'idée est de donner des comportements autorégressifs différents à une série, selon un ou plusieurs seuils pour les retards.

Pour les simulations, on a pris l'exemple suivant :

$$Y_t = \begin{cases} 0.5Y_{t-1} - 0.9Y_{t-2} - 3 + \varepsilon_t & , si\ Y_{t-2} \leq 1.5 \\ 0.3Y_{t-1} - 0.2Y_{t-2} + 2 + \varepsilon_t & , si\ Y_{t-2} > 1.5 \end{cases} \tag{6.4}$$

avec (ε_t) un bruit i.i.d. de densité gausienne centrée et réduite.

Trois échantillons contenant 200, 400 et, respectivement, 800 points ont été simulés.

Regardons d'abord l'échantillon de taille 200. Pour la classification initiale, on a choisi une grille de taille 5×5, les variables utilisées étant $\{Y_t, Y_{t-1}, Y_{t-2}\}$. Les résultats sont

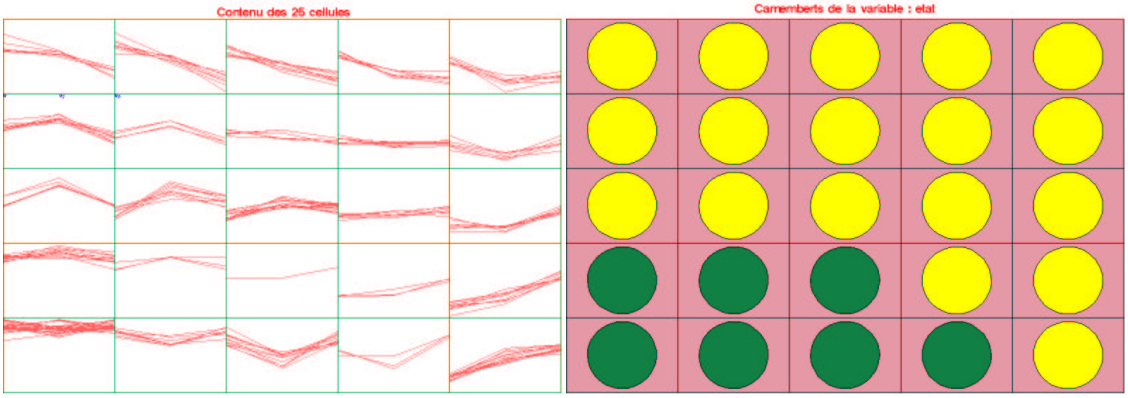

FIG. 6.3 – Classification initiale par carte de Kohonen pour un modèle TAR (200 observations)

	Regime 1			*Regime 2*		
	Constante	Y_{t-1}	Y_{t-2}	*Constante*	Y_{t-1}	Y_{t-2}
Valeurs estimées	-2.82	0.59	-0.89	1.4	0.37	0.32
T-statistique (Student)	-19.15	13.09	15.89	8.24	5.96	4.79
Vraies valeurs	-3	0.5	-0.9	2	0.3	-0.2

TAB. 6.1 – Les coefficients estimés du modèle TAR (200 observations)

présentés dans la Figure 6.3. Les profils des individus dans chaque cluster sont représentés à gauche. On a choisi ce graphique pour confirmer l'utilisation des cartes auto-organisées : les clusters sont bien homogènes, les observations similaires se retrouvent ensemble et, de plus, la préservation de la topologie permet d'avoir des clusters voisins avec des observations qui se ressemblent et qui devraient appartenir au même régime.

Cette observation est renforcée par le graphique de droite sur lequel on a croisé les clusters avec la variable booléenne $1_{\{Y_{t-2} \leq 1.5\}}$. La carte de Kohonen est bien arrivée à séparer les deux régimes dans chaque cluster, mais aussi au niveau topographique.

Si on fait tourner maintenant l'algorithme de classification hiérarchique sur les 25 clusters, la somme des carrés résiduelle évolue comme dans la Figure 6.4 et le choix de deux régimes semble raisonnable.

De plus, l'ajustement de régressions dans les deux classes finales fournit de bons estimateurs des paramètres (voir le Tableau 6.1), le seul paramètre qu'on n'arrive pas à estimer avec cette méthode étant la valeur du seuil. Ce type de classification permet de distinguer l'existence de plusieurs régimes, mais la question du type de modèle - à seuil, à changements de régime markoviens ou indépendants - ne peut être résolue que par des hypothèses a priori.

Pour les échantillons de 400 et, respectivement, 800 observations les résultats sont très similaires. On a choisi de ne pas les détailler, mais d'illustrer seulement (Figure 6.5) les clusters obtenus par des cartes de Kohonen de dimensions 7×7 et, respectivement, 9×9.

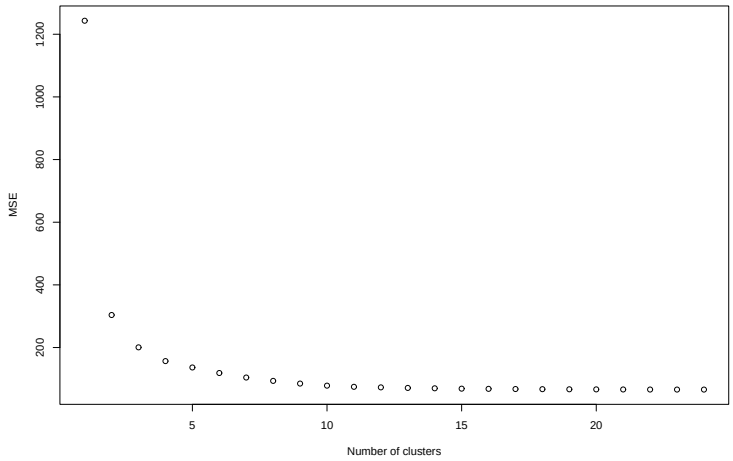

FIG. 6.4 – Somme des carrés résiduelle pour le modèle TAR (200 observations)

Ils sont homogènes et chaque cluster correspond à un regime, suivant le croisement avec la variable $1_{\{Y_{t-2}\leq 1.5\}}$. Dans la Figure 6.6, l'évolution de la somme des carrés résiduelle au cours des classifications hiérarchiques indique l'existence d'au moins deux régimes dans chaque cas.

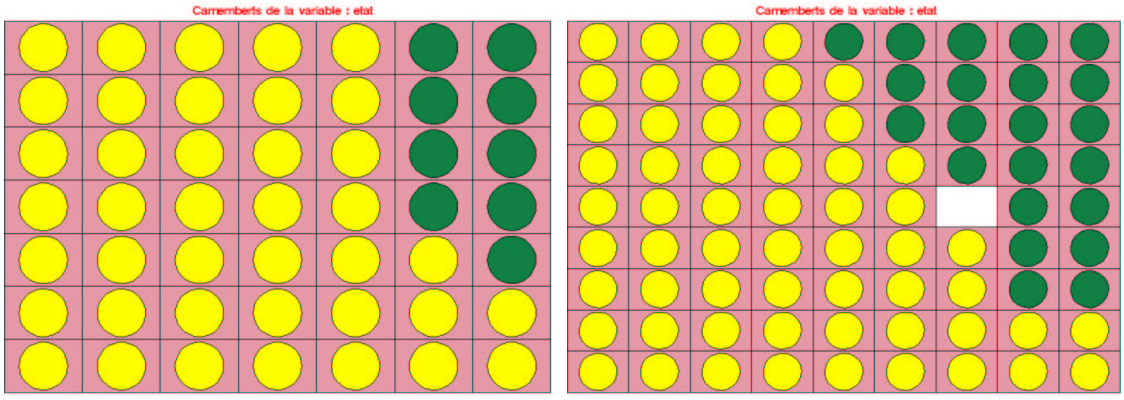

FIG. 6.5 – Classification initiale par carte de Kohonen pour un modèle TAR (400 et 800 observations)

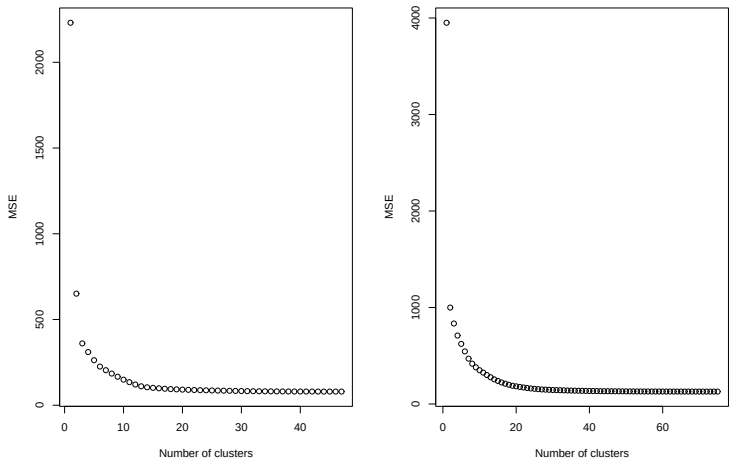

FIG. 6.6 – Somme des carrés résiduelle pour le modèle TAR (400 et 800 observations)

6.4.2 Modèles autorégressifs à changements de régime markoviens

6.4.2.1 Un modèle à deux régimes

Le modèle théorique est le suivant :

$$Y_t = F^0_{X_t}\left(Y_{t-1}, Y_{t-2}\right) + \sigma^0_{X_t}\varepsilon_t \qquad (6.5)$$

où

$$\begin{cases} F^0_1\left(Y_{t-1}, Y_{t-2}\right) = 0.5Y_{t-1} + 0.1Y_{t-2} + 0.2 \\ F^0_2\left(Y_{t-1}, Y_{t-2}\right) = 0.9Y_{t-1} - 0.1Y_{t-2} + 0.3 \end{cases},$$

$\sigma^0_1 = 0.03$, $\sigma^0_2 = 0.02$ et la matrice de transition de la chaîne de Markov (X_t)

$$A^0 = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix}$$

Comme pour les modèles TAR, trois échantillons de 200, 400 et, respectivement, 800 observations ont été simulés.

Les résultats pour l'échantillon de 200 observations sont présentés dans les Figures 6.7, 6.8 et le Tableau 6.2. En croisant les clusters obtenus par les cartes auto-organisées avec une variable catégorielle indiquant le régime, on remarque qu'il existe des cas où les deux régimes se chevauchent dans le même cluster. Ce problème se répercute sur la classification hiérarchique. Même si la somme des carrés résiduelle indique clairement qu'il y a au moins deux régimes, choisir entre deux et trois n'est pas évident.

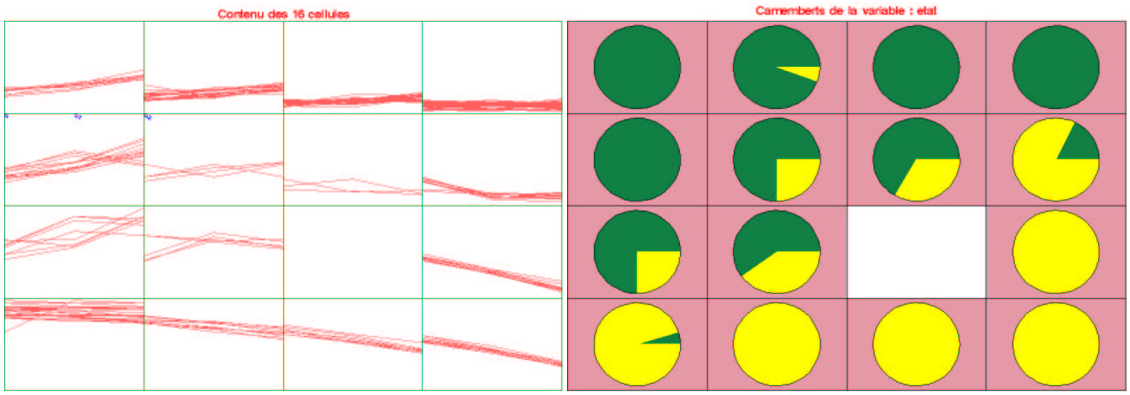

FIG. 6.7 – Classification initiale par carte de Kohonen pour un modèle à changements markoviens et à deux régimes (200 observations)

	Regime 1			*Regime 2*		
	Constante	Y_{t-1}	Y_{t-2}	*Constante*	Y_{t-1}	Y_{t-2}
Valeurs estimées	0.19	0.50	0.14	0.47	0.43	0.23
T-statistique (Student)	9.59	6.55	2.15	4.76	1.42	0.98
Vraies valeurs	0.2	0.5	0.1	0.3	0.9	-0.1

TAB. 6.2 – Les coefficients estimés du modèle à changements markoviens et à deux régimes (200 observations)

Remarquons aussi que l'utilisation de la distance euclidienne dans les cartes de Kohonen apporte une perte de l'information contenue par les données qui est leur caractère temporel. Une distance qui prendrait en compte cette propriété pourrait donc fournir de meilleurs résultats. On discutera cette possibilité dans les sections suivantes.

En ce qui concerne les échantillons de 400 et 800 observations, le nombre de clusters mélangeant des points des deux régimes est moins important et les résultats de la classification hiérarchique sont en faveur d'un modèle à deux régimes (Figures 6.9 et 6.10).

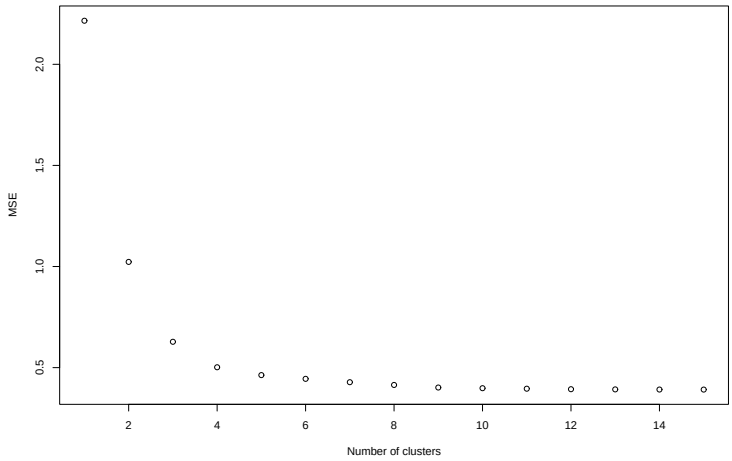

FIG. 6.8 – Somme des carrés résiduelle pour un modèle à changements markoviens et à deux régimes (200 observations)

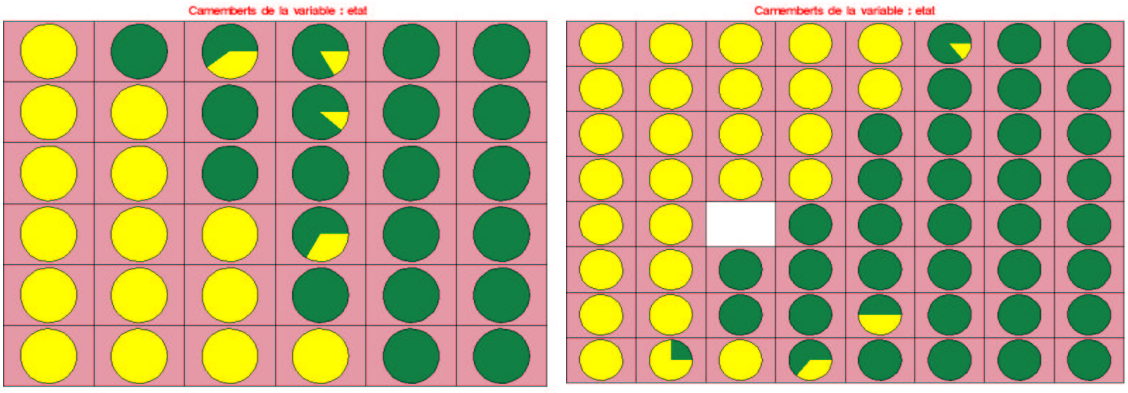

FIG. 6.9 – Classification initiale par carte de Kohonen pour un modèle à changements markoviens et à deux régimes (400 et 800 observations)

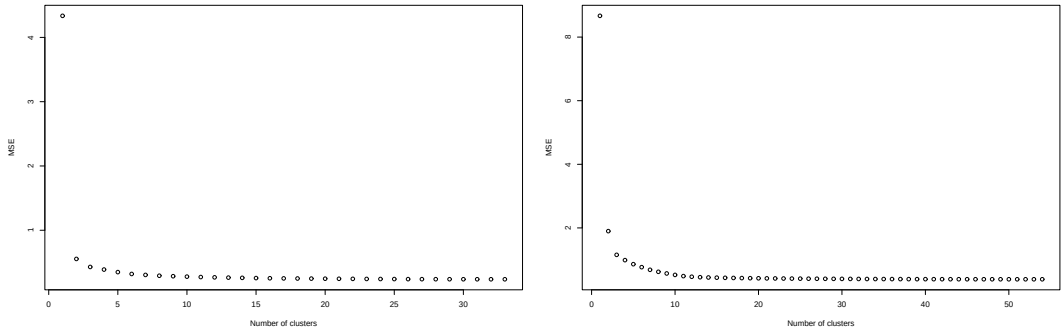

FIG. 6.10 – Somme des carrés résiduelle pour un modèle à changements markoviens et à deux régimes (400 et 800 observations)

6.4.2.2 Un modèle à trois régimes

Regardons maintenant ce qui se passe si au modèle à deux regimes, on rajoute un troisième régime à caractère explosif. On a choisi comme exemple un modèle qui admet un régime explosif, mais reste globalement stationnaire (les fonctions de régression vérifient les hypothèses (HS) de la section 4.1) :

$$Y_t = F^0_{X_t}\left(Y_{t-1}, Y_{t-2}\right) + \sigma^0_{X_t}\varepsilon_t \qquad (6.6)$$

où

$$\begin{cases} F^0_1\left(Y_{t-1}, Y_{t-2}\right) = 0.5Y_{t-1} + 0.1Y_{t-2} + 0.2 \\ F^0_2\left(Y_{t-1}, Y_{t-2}\right) = 0.9Y_{t-1} - 0.1Y_{t-2} + 0.3 \\ F^0_3\left(Y_{t-1}, Y_{t-2}\right) = 1.2Y_{t-1} + 0.5Y_{t-2} + 0.5 \end{cases},$$

$\sigma^0_1 = 0.03$, $\sigma^0_2 = 0.02$, $\sigma^0_3 = 0.03$ et la matrice de transition de la chaîne de Markov (X_t) est

$$A^0 = \begin{pmatrix} 0.8 & 0.1 & 0.6 \\ 0.1 & 0.8 & 0.2 \\ 0.1 & 0.1 & 0.2 \end{pmatrix}$$

Les résultats de la classification présentés dans les Figures 6.11 et 6.12 correspondent à un échantillon de 400 observations. La classification initiale par une carte de Kohonen révèle une bonne séparation des régimes, avec le troisième isolé en bas de la carte à droite et quelques clusters où les deux premiers régimes se mélangent.

Remarquons aussi le cluster situé en bas à gauche de la carte. Il ne contient que quatre observations, avec des profils très différents et qui proviennent des trois régimes. De plus, un représentation en trois dimensions des variables $\{Y_t, Y_{t-1}, Y_{t-2}\}$ montre que ces quatre valeurs sont très éloignées du reste des données (voir la Figure 6.13).

En regardant la classification hiérarchique sur les clusters, il semblerait y avoir quatre classes, mais, en regardant leur composition, trois classes correspondent bien aux trois régimes tandis que dans la quatrième on retrouve uniquement le cluster contenant des valeurs extrêmes. L'algorithme de classification hiérarchique proposé est donc sensible aux valeurs aberrantes, comme la méthode de Ward.

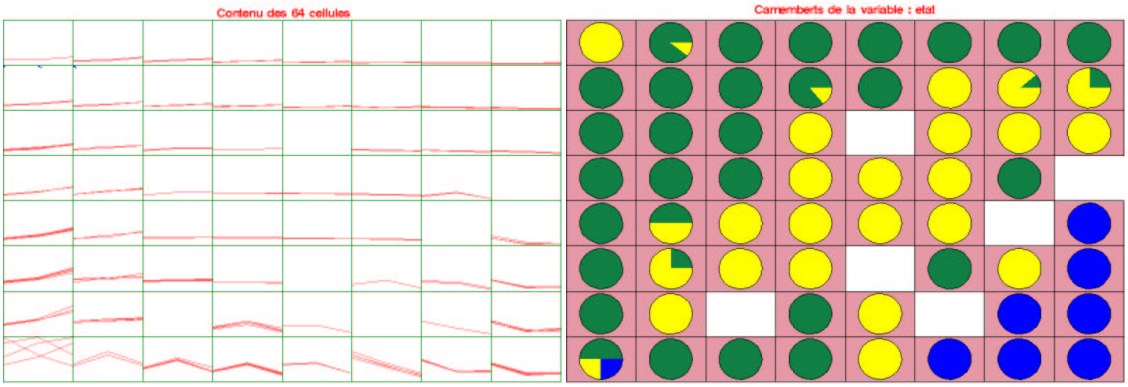

FIG. 6.11 – Classification initiale par carte de Kohonen pour un modèle à changements markoviens et à trois régimes (400 observations)

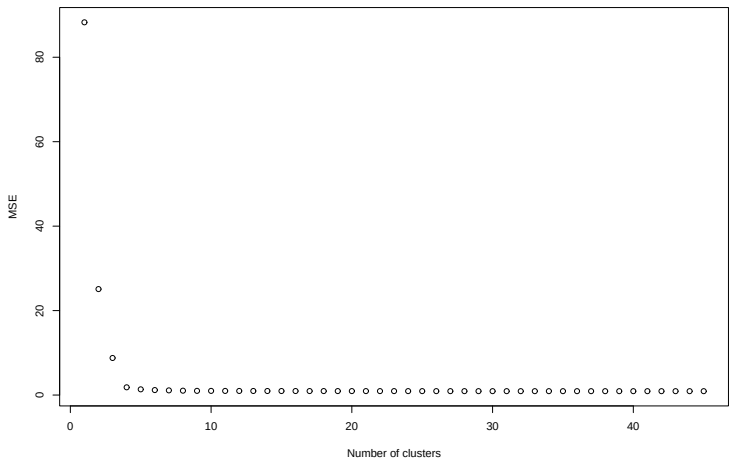

FIG. 6.12 – Somme des carrés résiduelle pour un modèle à changements markoviens et à trois régimes (400 observations)

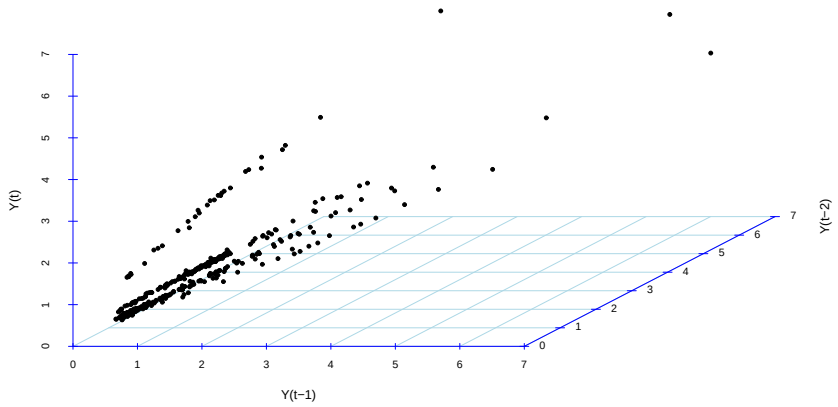

FIG. 6.13 – L'échantillon provenant d'un modèle à changements markoviens et à trois régimes (400 observations)

	Regime 1			*Regime 2*			*Regime 3*		
	Const.	Y_{t-1}	Y_{t-2}	*Const.*	Y_{t-1}	Y_{t-2}	*Const.*	Y_{t-1}	Y_{t-2}
Valeurs estimées	0.23	0.20	0.09	0.28	1.33	-0.53	0.52	1.20	0.48
T-stat (Student)	28.37	53.94	10.71	6.62	22.07	-9.15	31.64	122.86	30.85
Vraies valeurs	0.2	0.5	0.1	0.3	0.9	-0.1	0.5	1.2	0.5

TAB. 6.3 – Les coefficients estimés du modèle à changements markoviens et à trois régimes (400 observations)

6.4.3 Données réelles

Les résultats sur les données simulées étant encourageants, reste à appliquer la méthode à des jeux de données réelles : le premier est le "classique" Old Faithful Geyser Data, le deuxième est la série du PNB (Produit National Brut) aux Etats-Unis, série utilisée par Hamilton (1989) pour introduire les modèles autorégressifs à changements de regime markoviens.

6.4.3.1 Old Faithful Geyser Data

Les données Old Faithful Geyser du parc national Yellowstone sont disponibles dans la plupart des logiciels statistiques de type R ou S-Plus et constituent l'un des bancs d'essai pour la validation des méthodes de classification. Elles contiennent 299 observations à deux composantes qui sont le temps d'attente entre deux éruptions w_t et la durée de

l'éruption suivante d_t. Les données ont été enregistrées entre le 1er et le 15 août 1985 et sont exprimées en minutes. A cause des mesures inexactes pendant la nuit (les appareils ne sont plus surveillés, il y a des erreurs), pour plusieurs relevés la durée a été approximée par des entiers : 53 points où $d_t = 4$ (durée longue), 20 où $d_t = 2$ (durée courte) et un point où $d_t = 3$ (durée moyenne).

Il existe plusieurs études sur ces données. La majorité des auteurs se sont intéressés soit à trouver une classification pertinente des données, soit à souligner une dépendance temporelle entre événements successifs. Un passage en revue de la littérature, ainsi qu'une modélisation par des séries temporelles en supposant "a priori" l'existence de deux classes sont disponibles dans Azzalini et Bowman (1990).

Une approche pour expliquer les données simultanément par une classification et par une dépendance entre les variables a été proposée par Hennig (2000) qui a introduit la méthode "fixed-point regression clustering". Notre algorithme est assez proche de cette dernière par le fait qu'on essaie de déterminer le nombre de classes, tout en ajustant un modèle de régression dans chaque classe.

On a vu que l'algorithme de classification décrit dans ce chapitre utilise la distance euclidienne entre vecteurs dans les cartes auto-organisées et perd ainsi l'information selon laquelle les variables explicatives sont des retards de la série. Il peut être appliqué dans un cadre plus général que les modèles autorégressifs et donc en particulier aux données Old Faithful Geyser. On cherche donc à expliquer la durée des éruptions en fonction des temps d'attente par le modèle :

$$d_t = a^0_{X_t} w_t + b^0_{X_t} + \sigma^0_{X_t} \varepsilon_t \qquad (6.7)$$

où X_t est un processus non-observé à valeurs dans l'espace fini $\{1, ..., p_0\}$, p_0 inconnu, et ε_t est un bruit blanc.

Pour déterminer p_0, on applique la méthode proposée : on commence par une classification initiale des variables $\{d_t, w_t\}$ avec une carte de Kohonen sur une grille de taille 6×6 et ensuite on regroupe suivant une classification hiérarchique les clusters obtenus. La représentation de la somme des carrés résiduelle au cours du regroupement des clusters dans la Figure 6.14, indique l'existence de deux classes.

Choisir deux classes semble donc la décision raisonnable à prendre et ceci est confirmé par les Figures 6.15 et 6.16. Dans la première, on a représenté les classes obtenues "a posteriori" par la méthode de classification hiérarchique. La séparation au niveau topologique est évidente. Dans la Figure 6.16, on a représenté les observations en prenant les temps d'attente en abscisse et la durée en ordonnée. Les cercles sont les observations de la première classe, les triangles de la deuxième. Au niveau de l'interprétation, la première classe contient les points avec des durées d'éruption longues précédées de temps d'attente variables, tandis que la deuxième est concentrée autour des points avec durée $d_t = 2$ et de longs temps d'attente (les coefficients de régression pour les deux classes sont présentés au Tableau 6.4). Cette conclusion est cohérente avec l'hypothèse "a priori" de Azzalini et Bowman (1990) qui apportent des arguments géologiques pour l'existence de deux régimes dans le modèle de la durée d'éruption.

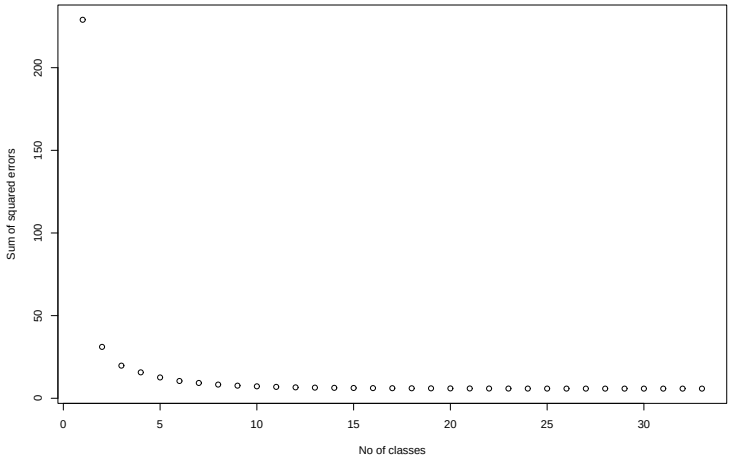

FIG. 6.14 – Somme des carrés résiduelle pour les données Old Faithful Geyser

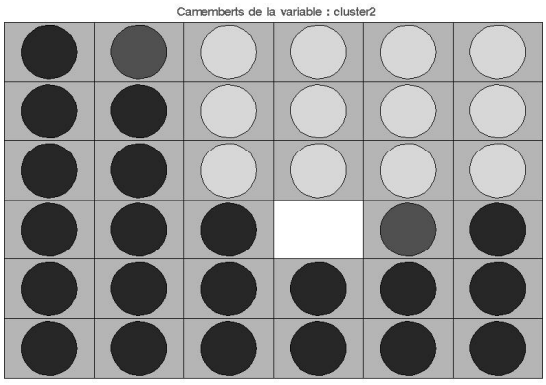

FIG. 6.15 – Classification initiale par carte de Kohonen pour les données Old Faithful Geyser

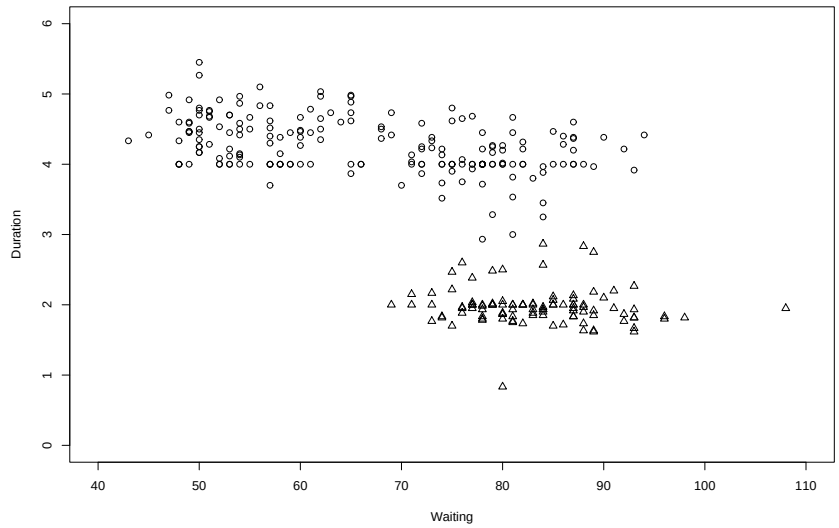

FIG. 6.16 – Représentation à deux classes pour les données Old Faithful Geyser

	Regime 1		*Regime 2*	
	Constante	w_t	*Constante*	w_t
Valeurs estimées	5.095	-0.012	2.315	-0.004
T-statistique (Student)	38.71	-6.42	7.11	-1.06

TAB. 6.4 – Les coefficients estimés du modèle à deux régimes pour les données Old Faithful Geyser

6.4.3.2 La série PNB aux Etats-Unis

La série du PNB d'après guerre aux Etats-Unis a été utilisée par Hamilton (1989) pour introduire et illustrer les modèles autorégressifs à changements de regime markoviens. Ces modèles ont permis une segmentation de la série qui est cohérente avec le cycle économique : les taux de croissance positifs sont associés à des périodes normales de l'économie, tandis que les taux de croissance négatifs correspondent à des périodes de récession. Cependant, le choix de deux régimes n'est pas justifié de manière statistique dans ses travaux.

La série, notée GNP *(Gross National Product)* et disponible dans *Business Conditions Digest* (Février 1986), contient 136 observations enregistrées tous les trois mois entre 1951 et 1984. En raison de l'existence évidente d'une tendance (voir la Figure 6.17), on cherche d'abord à la stationnariser et on considère la série des rendements :

$$y_t = 100 \cdot ln\left(\frac{GNP_t}{GNP_{t-1}}\right)$$

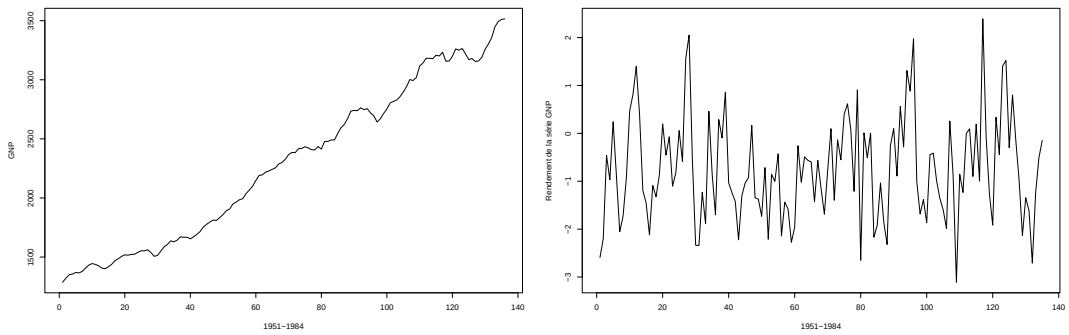

FIG. 6.17 – La série GNP et la série stationnarisée des rendements

Le nombre de retards dans Hamilton (1989) a été choisi arbitrairement égal à 4. Pour des raisons de cohérence et d'interprétabilité des résultats, on a conservé ce nombre dans la suite. La classification initiale est réalisée sur une grille 4×4 pour les vecteurs $(y_t, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. Le résultat de la classification hiérarchique sur les seize clusters est présenté dans les Figures 6.18 et 6.19. Le premier graphique de la Figure 6.19 contient la somme des carrés résiduelle selon le nombre de classes, le deuxième détaille la modification en pourcentage de la somme des carrés résiduelle. Deux sauts importants, même s'ils sont d'une amplitude moindre que dans les exemples précédents, indiquent l'existence de plusieurs régimes.

Pour confirmer l'hypothèse de l'existence de plusieurs régimes, on a utilisé le test de Chow (Chow (1960)) pour l'égalité des coefficients dans deux régressions linéaires.

Si on étudie par exemple le passage de trois à deux classes, notons C_1, C_2, C_3 les trois classes existantes à ce niveau et supposons que C_2 et C_3 sont regroupées. Dans ce cas, on considère les hypothèses "H_0 : les observations des classes C_2 et C_3 proviennent du même régime" et "H_1 : les observations des classes C_2 et C_3 appartiennent à deux régimes différents". La statistique du test est

$$F = \frac{\left(SSE_{C_2 \cup C_3} - SSE_{C_2} - SSE_{C_3}\right)/l}{\left(SSE_{C_2} + SSE_{C_3}\right)/\left(\#C_2 + \#C_3 - 2l\right)}$$

où l est le nombre de retards. Sous des conditions de normalité du bruit et homoscédasticité, $F \sim \mathbb{F}(l, \#C_2 + \#C_3 - 2l)$ sous l'hypothèse nulle. Puisqu'on travaille avec des rendements d'une série financière, on sait qu'en général la distribution est proche d'une log-normale. De plus, Toyoda (1974) a montré que même si on n'a pas l'homoscédasticité, le test reste significatif si au moins une des classes contient suffisamment d'observations. L'emploi de ce test est donc justifié. On a $\#C_2 = 92$, $\#C_3 = 21$ et la statistique du test est $F_{obs} = 9,938$. Le $(1-\alpha)$-quantile de $\mathbb{F}(4, 105)$ vaut 2,458 pour $\alpha = 0,05$ et 3,5031 pour $\alpha = 0,01$ et donc l'hypothèse d'égalité des coefficients est rejetée à 5% et même à 1%.

De la même manière, on montre que la statistique du test est significative au passage de deux à un régime. On déduit donc l'existence de plusieurs régimes, mais on n'arrive pas à trancher entre deux et trois. La décision finale en faveur de deux régimes est prise en s'appuyant sur l'argument économique de l'existence d'une périodicité récessions - périodes

normales (les coefficients de régression pour les deux régimes sont présentés dans le Tableau 6.5).

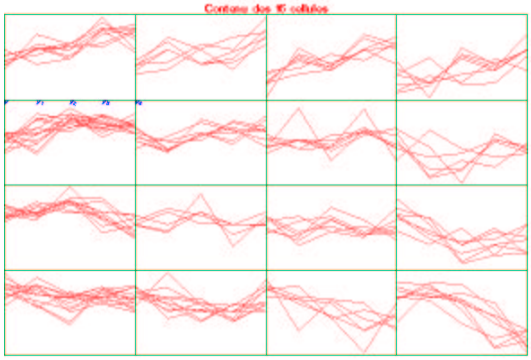

FIG. 6.18 – Classification initiale par carte de Kohonen pour les données GNP

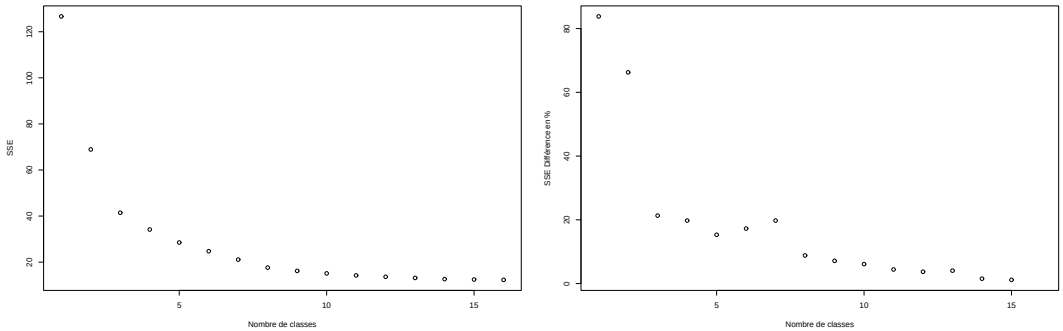

FIG. 6.19 – Somme des carrés résiduelle et variation en % de la somme des carrés résiduelle pour les données GNP

	Regime 1					*Regime 2*				
	Const.	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	*Const.*	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}
Valeurs estimées	-0.16	-0.07	0.21	0.11	-0.01	1.60	-0.03	-0.14	0.13	0.08
T-stat. (Student)	-0.89	-0.65	1.79	0.93	-0.11	10.32	-0.25	-1.70	1.50	1.11

TAB. 6.5 – Les coefficients estimés du modèle à deux régimes pour les données GNP

6.5 Alternative à la distance euclidienne dans la classification initiale

6.5.1 Une distance fonctionnelle pour les séries temporelles

Les sections précédentes ont montré, de manière empirique, que l'utilisation des cartes de Kohonen avec distance euclidienne fournissait de bons résultats pour mettre en évidence l'existence de plusieurs régimes dans un modèle autorégressif. Cependant, la distance euclidienne ne tient pas compte des caractéristiques temporelles des données, des structures des corrélations des variables retardées entre elles et une partie de l'information est perdue.

Lee et Verleysen (2005) ont proposé récemment une généralisation de la norme fonctionnelle L_2 qui revient à la discrétiser pour pouvoir l'utiliser sur des données vectorielles. On étudie dans la suite si le remplacement de la distance euclidienne par la distance proposée par Lee et Verleysen améliore les résultats de la classification hiérarchique.

Pour cela, rappelons d'abord quelques notations et propriétés.

Si $y : \mathbb{R} \to \mathbb{R}$ est une fonction de carré intégrable, on définit la norme L_2 de y par :

$$\|y\|_2 = \left(\int_{\mathbb{R}} y^2 (t)\, dt \right)^{\frac{1}{2}}$$

Si on considère maintenant un échantillon observé $y_1^n = \{y_1, ..., y_n\}$ de la série Y_t et la matrice obtenue par une fenêtre glissante de taille l,

$$\mathbb{Y} = \{y_t, y_{t-1}, ..., y_{t-l}\}_{t=l+1,...,n} \in \mathbb{R}^{(n-l)\times(l+1)},$$

Lee et Verleysen définissent, pour chaque vecteur $y(t) = (y_t, y_{t-1}, ..., y_{t-l})^T$, la norme $L_{TS,2}$ obtenue par la discrétisation de la norme L_2 :

$$\|y(t)\|_{TS,2} = \left(\sum_{i=0}^{l} (A_{i-1} + A_{i+1})^2 \right)^{\frac{1}{2}}$$

où

$$A_{i-1} = \begin{cases} \frac{\tau}{2} |y_i| & , \ y_i y_{i-1} \geq 0 \\ \frac{\tau}{2} \frac{|y_i|^2}{|y_i| + |y_{i-1}|} & , \ y_i y_{i-1} < 0 \end{cases} \text{ et } A_{i+1} = \begin{cases} \frac{\tau}{2} |y_i| & , \ y_i y_{i+1} \geq 0 \\ \frac{\tau}{2} \frac{|y_i|^2}{|y_i| + |y_{i+1}|} & , \ y_i y_{i+1} < 0 \end{cases}$$

avec τ strictement positif fixé, $A_{-1} = 0$ et $A_{l+1} = 0$.

Avec cette définition, $\|y(t)\|_{TS,2}$ représente la norme L_2 de la fonction linéaire par morceaux obtenue en reliant les composantes du vecteur $y(t)$ (voir la Figure 6.20).

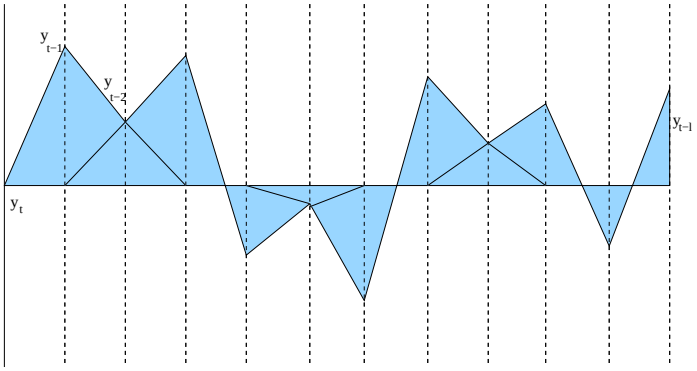

FIG. 6.20 – Calcul de la norme $\|y(t)\|_{TS,2}$

La norme $\|\cdot\|_{TS,2}$ engendre alors la distance suivante :

$$d_{TS,2}\left(y(t), y'(t)\right) = \left\|y(t) - y'(t)\right\|_{TS,2}$$

Dans toute la suite on considère $\tau = 1$ et on a la relation :

$$d_{TS,2}\left(y(t), y'(t)\right) \leq \left(\sum_{i=0}^{l}\left(y_{t-i} - y'_{t-i}\right)^2\right)^{\frac{1}{2}}$$

avec égalité si toutes les coordonnées sont soit positives, soit négatives.

On peut alors modifier le Pas 3 de l'algorithme *Alg.1* (section 6.3.3) en remplaçant la distance euclidienne par la nouvelle distance $d_{TS,2}$. Dans un premier temps, on a implémenté la méthode de Kohonen à zéro voisin. Implémenter l'algorithme avec des cartes auto-organisées et comparer ses performances par rapport aux méthodes "regression type clustering" sont sur la liste "à faire" des prochains mois.

Etudions maintenant si la distance $d_{TS,2}$ est mieux adaptée aux données sur les exemples suivants.

6.5.2 Utilisation de la distance fonctionnelle pour des modèles autorégressifs à changements de régime markoviens

6.5.2.1 Un premier exemple

Le premier exemple reprend la série à deux régimes de la section 6.4.2.1. On rappelle que la série est générée par l'équation suivante :

$$Y_t = F^0_{X_t}\left(Y_{t-1}, Y_{t-2}\right) + \sigma^0_{X_t}\varepsilon_t \qquad (6.8)$$

avec

$$\begin{cases} F^0_1\left(Y_{t-1}, Y_{t-2}\right) = 0.5Y_{t-1} + 0.1Y_{t-2} + 0.2 \\ F^0_2\left(Y_{t-1}, Y_{t-2}\right) = 0.9Y_{t-1} - 0.1Y_{t-2} + 0.3 \end{cases},$$

$\sigma^0_1 = 0.03$, $\sigma^0_2 = 0.02$ et $A^0 = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix}$ comme matrice de transition de la chaîne de Markov X_t.

Sur un échantillon de 200 observations, l'algorithme utilisant la distance euclidienne (voir la section 6.4.2.1) montrait l'existence de plusieurs régimes avec des ruptures importantes dans la somme des résidus au carré aux passages de trois à deux régimes et de deux à un régime. La classification initiale était assez mauvaise, la moitié des clusters mélangeant des observations des deux régimes. Pourtant, les deux régimes avaient été choisis de manière à être assez éloignés (Figure 6.21) et une bonne métrique aurait dû mieux séparer les données dans les clusters.

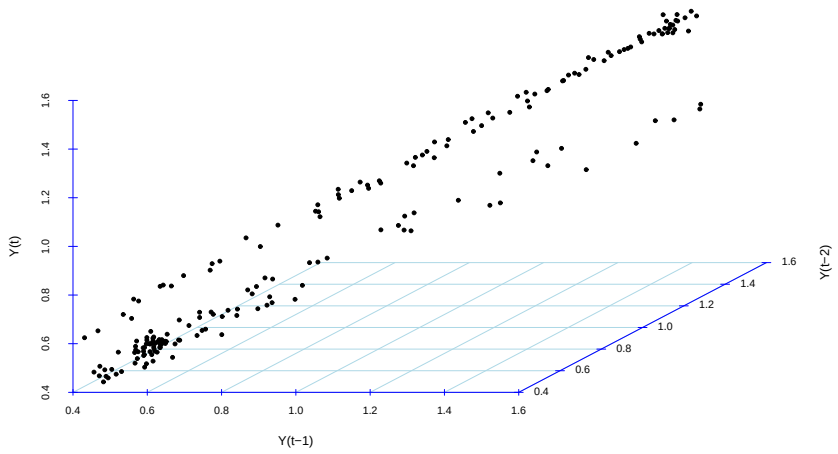

FIG. 6.21 – La représentation *3D* du nuage des points correspondant au modèle à changements de régimes (6.8)

Si maintenant on fait tourner l'algorithme de classification en utilisant la distance $d_{TS,2}$ et une classification initiale ayant toujours 16 clusters, on obtient les résultats illustrés dans la Figure 6.22. Le premier graphique représente les observations (y_t, y_{t-1}, y_{t-2}) contenues dans les clusters. Le deuxième graphique présente le croisement des clusters avec la variable indiquant le régime. Cette fois-ci, le nombre d'observations "mal classées" est moins important, ce qui montre que la nouvelle distance est mieux adaptée à ce type de données. De plus, dans la classification hiérarchique des clusters (Figure 6.23), la rupture au niveau du passage de deux à un régime est beaucoup plus importante et permet de décider qu'il existe deux régimes.

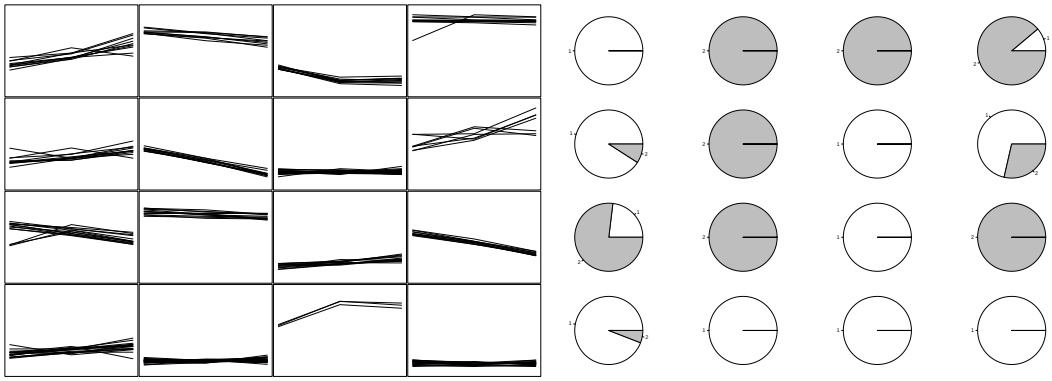

FIG. 6.22 – Classification initiale par la méthode Kohonen à 0-voisin pour le modèle à changements de regimes (6.8)

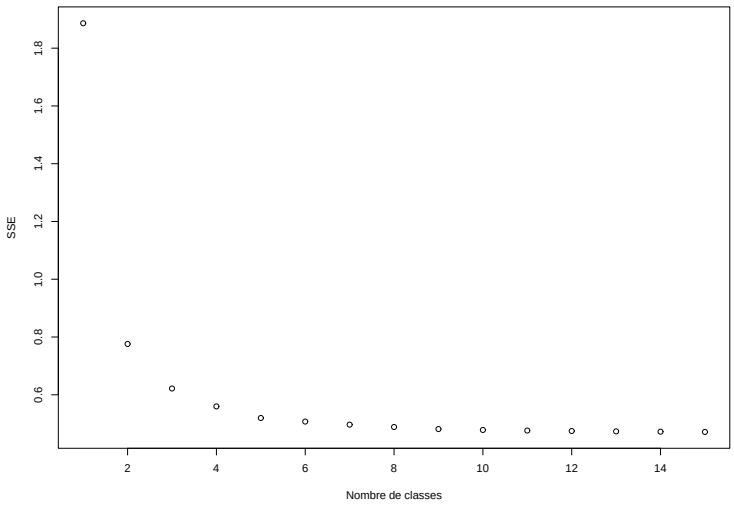

FIG. 6.23 – Somme des carrés résiduels pour le modèle à changements de régime (6.8)

	Regime 1			*Regime 2*		
	Constante	Y_{t-1}	Y_{t-2}	*Constante*	Y_{t-1}	Y_{t-2}
Valeurs estimées	0.17	0.41	0.24	0.35	0.73	0.01
T-statistique (Student)	10.37	6.29	5.05	8.79	5.17	0.11
Vraies valeurs	0.2	0.5	0.1	0.3	0.9	-0.1

TAB. 6.6 – Les coefficients estimés du modèle à changements markoviens et à deux régimes (200 observations)

6.5.2.2 Un second exemple

Pour le second exemple, on a gardé les mêmes équations pour générer la série, le même nombre de régimes, les mêmes fonctions de régression et la même matrice de transition, mais on a augmenté la variance du bruit. Le nouveau modèle s'écrit alors

$$Y_t = F^0_{X_t}\left(Y_{t-1}, Y_{t-2}\right) + \sigma^0_{X_t}\varepsilon_t \qquad (6.9)$$

où

$$\begin{cases} F^0_1\left(Y_{t-1}, Y_{t-2}\right) = 0.5Y_{t-1} + 0.1Y_{t-2} + 0.2 \\ F^0_2\left(Y_{t-1}, Y_{t-2}\right) = 0.9Y_{t-1} - 0.1Y_{t-2} + 0.3 \end{cases},$$

$\sigma^0_1 = 0.3$, $\sigma^0_2 = 0.2$ et $A^0 = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix}$ comme matrice de transition de la chaîne de Markov X_t.

On a simulé une série de longueur 200 et dans la Figure 6.24 on a représenté le nuage de points (y_t, y_{t-1}, y_{t-2}) en trois dimensions. On remarque que dans ce cas les deux régimes se superposent.

L'utilisation de l'algorithme *Alg.1* (section 6.3.3) avec la distance euclidienne n'arrivant pas à donner un résultat satisfaisant, on a testé l'algorithme avec la distance $d_{TS,2}$. Les 25 clusters produits par la classification initiale avec la méthode Kohonen à 0-voisin sont assez homogènes, même si la plupart contiennent des observations appartenant à des regimes différents (Figure 6.25).

La classification hiérarchique des 25 clusters (Figure 6.26) révèle un saut important au passage de deux à une seule classe et on décide alors de sélectionner deux régimes.

	Regime 1			*Regime 2*		
	Constante	Y_{t-1}	Y_{t-2}	*Constante*	Y_{t-1}	Y_{t-2}
Valeurs estimées	0.32	-0.06	0.42	0.16	1.11	-0.28
T-statistique (Student)	8.54	-0.69	5.47	3.00	14.30	-3.27
Vraies valeurs	0.2	0.5	0.1	0.3	0.9	-0.1

TAB. 6.7 – Les coefficients estimés du modèle à changements markoviens et à deux régimes (200 observations)

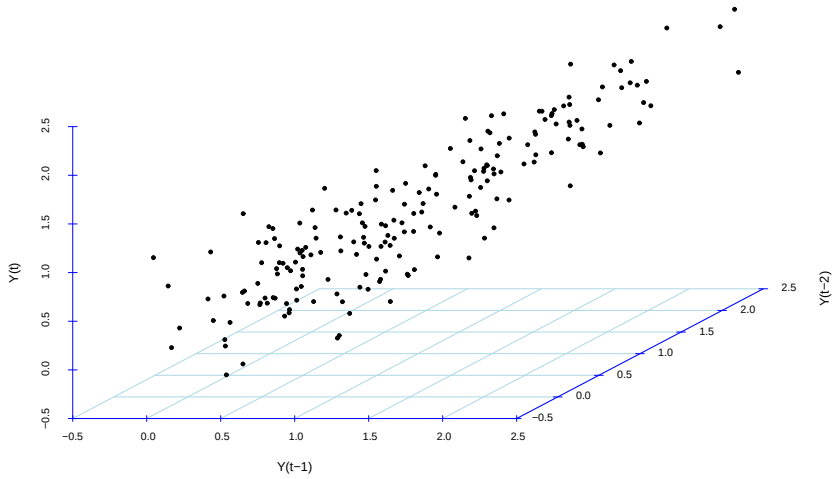

FIG. 6.24 – La représentation *3D* du nuage des points correspondant au modèle à changements de régimes (6.9)

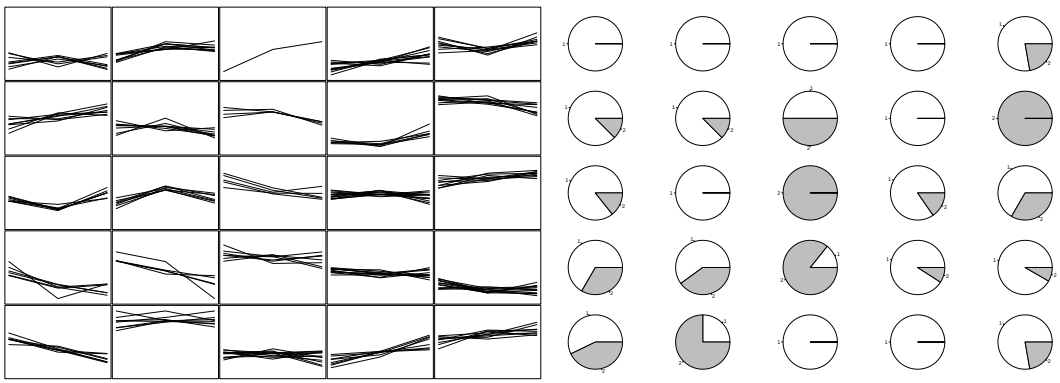

FIG. 6.25 – Classification initiale par la méthode Kohonen à 0-voisin pour le modèle à changements de régime (6.9)

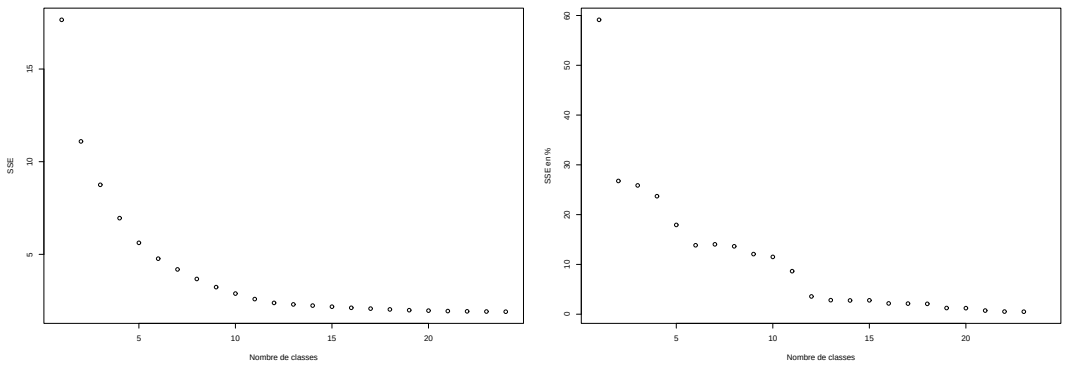

FIG. 6.26 – Somme des résidus au carré et augmentation en % de la somme de résidus au carré pour le modèle à changements de regime (6.7)

6.6 Conclusion

On a proposé une méthode empirique basée sur des méthodes de classification non-supervisées qui permet d'illustrer la présence d'éventuels régimes dans un modèle autorégressif. L'étude a porté sur des modèles à fonctions de régression linéaires dans chaque régime, mais une modification pour d'autres fonctions non-linéaires comme les perceptrons multi-couches pourrait être envisagée. Cependant, dans ce cas il faudrait travailler sur des ensembles de données suffisamment larges pour permettre une bonne estimation des paramètres.

L'algorithme proposé a été testé sur plusieurs exemples simulés ou réels. Sur les exemples simulés on a montré que les résultats sont corrects et qu'on choisit le bon nombre de régimes lorsque les paramètres des modèles dans chaque régime sont assez éloignés et que le résultat risque de surestimer le vrai nombre de régimes lorsqu'on rapproche les hyperplans de régression caractérisant chaque régime. De plus, il semble que la classification initiale est très sensible aux valeurs extrêmes. Les résultats sur les données réelles (Old Faithful Geyser Data et le Produit National Brut aux Etats-Unis) sont encourageants par leur concordance avec la littérature existante.

Ensuite, on a abordé le problème de perte d'information sur les données à cause de la non-prise en compte de leur caractère temporel dans la classification initiale. On a utilisé la modification de la norme L_2 proposée par Lee et Verleysen (2005) de façon à ce qu'elle devienne une distance vectorielle et on a montré une amélioration générale des résultats.

En absence d'une réponse théorique complète concernant la sélection du nombre de régimes dans un modèle auto-régressif, cette méthode peut être utilisée avec, pourtant, de la précaution. En effet, la classification initiale peut mélanger des observations provenant de régimes différents si les coefficients de régression sont très proches ou si la variance du bruit est importante. Dans ce cas, il serait indiqué d'accompagner le choix empirique du nombre de régimes d'un test statistique sur l'égalité des coefficients de régression, en sachant qu'on risque de surestimer le modèle si les clusters initiaux mélangent des observations provenant de régimes différents.

Chapitre 7

Conclusion et perspectives

Le but de ce document a été de mettre en évidence l'utilité des modèles autorégressifs à changements de régime dans l'analyse des séries temporelles, en particulier des séries financières, et d'illustrer les problèmes théoriques qu'ils soulèvent.

Dans un premier temps, on a mené une étude qui essaie de quantifier les turbulences sur les marchés en modélisant un indice de crise appelé IMS (*Index of Market Shocks*). Cette étude a montré que les modèles à changements de régime fournissent une bonne segmentation de la suite de données et permettent une évaluation de la durée des crises. Cependant, il n'y a pas de justification théorique pour le choix du nombre de régimes, ce qui conduit à une certaine subjectivité de la modélisation. On a donc choisi de s'intéresser à ce problème ensuite.

Le principal inconvénient de ce type de modèle est la non-identifiabilité des paramètres. En effet, la théorie classique sur les tests du rapport de vraisemblance ne s'applique pas et Gassiat et Keribin (2000) ont montré la divergence de ce rapport dans le cas des chaînes de Markov cachées. En revanche, une approche par maximum de vraisemblance pénalisée semble envisageable et la consistance faible d'un estimateur du nombre de régimes par critère de vraisemblance pénalisée a été montrée par Gassiat (2002) pour les chaînes de Markov cachées.

Dans ce document, on a abordé deux classes de modèles autorégressifs : à changements de régime indépendants et à changements de régime markoviens. Dans les deux cas, le nombre de retards a été considéré connu et fixé à l'avance.

Dans le cadre des modèles autorégressifs à changements de régime indépendants, on a montré la consistance faible d'un estimateur de vraisemblance pénalisée pour le nombre de régimes. La preuve est faite pour des processus stationnaires et absolument réguliers, sous des hypothèses de régularité et de complexité de la classe des fonctions scores généralisés. Ces hypothèses ont été vérifiées ensuite pour des modèles autorégressifs à bruit gaussien et les résultats de convergence ont été illustrés par des simulations. Le but des simulations a été principalement d'étudier la vitesse de convergence de l'estimateur selon les différentes valeurs des paramètres dans chaque régime. L'influence du terme de pénalité n'a pas été abordée dans ce document et c'est un des points que nous souhaitons étudier maintenant. Toutefois, il faut remarquer que l'utilisation du terme de pénalité du critère BIC a fourni de bons résultats.

En ce qui concerne la généralisation aux modèles autorégressifs à changements de régime markoviens, on a montré que l'utilisation d'une vraisemblance marginale conduit à une fonction de contraste et à un estimateur consistant uniquement dans le cas des chaînes de Markov cachées pour lesquelles les fonctions de régression sont constantes. Le résultat de divergence dans le cas général a été illustré par des simulations. Une approche par un critère de vraisemblance pénalisée en utilisant la vraie vraisemblance pourrait alors être envisagée. Démontrer qu'un critère de ce type ne sous-estime pas le vrai nombre de régimes est assez immédiat. Il suffit d'étendre les résultats de Leroux (1992), Rydén (1995) et Francq, Roussignol et Zakoian (2001) qui portent sur les mélanges de lois, les chaînes de Markov cachées et, respectivement, les modèles GARCH à coefficients dépendant d'une chaîne de Markov cachée. La partie difficile qui resterait à résoudre est de montrer que le même critère ne surestime pas le vrai modèle.

L'approche par maximum de vraisemblance pénalisée ne répondant que partiellement à la question de la sélection du nombre de régimes, on a proposé au dernier chapitre une méthode empirique qui combine les cartes de Kohonen et une classification hiérarchique pour mettre en évidence l'existence de régimes dans un modèle autorégressif. Dans ce cas, le type de changements de régime (modèles à seuil, à changements de régime indépendants ou markoviens) ne compte plus, la seule hypothèse est que les données observées sont caractérisées par plusieurs régressions linéaires.

La méthode proposée a été testée sur plusieurs exemples de données simulées ou réelles, les résultats obtenus étant assez encourageants. De plus, une modification de la distance euclidienne dans l'algorithme de Kohonen pour qu'elle prenne en compte le caractère de dépendance temporelle entre les données a été introduite. Avec cette nouvelle distance, les résultats ont été généralement améliorés. Cependant, cette méthode est à utiliser avec précaution. Des coefficients trop proches dans les différents régimes ou un bruit avec une variance trop importante peuvent conduire à une classification initiale mélangeant les régimes dans les clusters et le vrai nombre de régimes est souvent surestimé.

Une autre question qui est alors apparue concerne les données provenant de modèles non-stationnaires. L'approche par vraisemblance pénalisée étant démontré uniquement dans le cas stationnaire, on aurait souhaité trouver au moins une réponse partielle au problème du choix du nombre de régimes en utilisant la méthode empirique du Chapitre 6. La réponse n'est pourtant pas immédiate. Si, par exemple, dans le modèle (6.6) on considère que le troisième régime qui est explosif apparaît plus souvent (il suffit pour cela d'augmenter la probabilité de rester dans cet état dans la matrice de transition A^0), alors les observations correspondant à ce régime sont réparties sur la plus grande partie de la carte de Kohonen et les deux régimes stationnaires se retrouvent écrasés dans un coin de la carte. Que faire dans ce cas ?

Une généralisation importante de la problématique étudiée dans ce document serait de considérer que le nombre de retards est aussi inconnu et de repenser la question de sélection de modèle comme une double sélection : du nombre de régimes et du nombre de retards. De nouveaux problèmes apparaissent dans ce cas, le plus difficile étant, peut-être, de trouver une bonne reparamétrisation de cette double non-identifiabilité. Une réponse partielle a été apportée par Chambaz et Matias (2006) qui étudient la sélection jointe de la mémoire d'une chaîne de Markov à chaîne de Markov cachée et du nombre de régimes de la chaîne

de Markov cachée dans le cas ou le processus observé est à espace d'états fini, mais la généralisation à un espace d'états continu n'est pas encore établie.

Bibliographie

[1] Andersen T., Bollerslev T., Diebold F., Ebens H. (2001) The distribution of stock returns volatilities, *Journal of Financial Economics*, **61(1)**, 43-76

[2] Anderson H., Nam K., Vahid F. (1999) Asymmetric nonlinear smooth transition GARCH models, *Nonlinear Time Series Analysis of Economic and Financial Data*, Rothman (Ed.) Kluwer

[3] Azzalini A., Bowman A.W. (1990) A look at some data on the Old Faithful Geyser, *Applied Statistics*, **39**, 357-365

[4] Bakry D., Milhaud X., Vandekerkhove P. (1997) Statistique de chaînes de Markov cachées à espace d'états fini. Le cas non-stationnaire, *note C. R. Acad. Sci. Paris*, **325-I**, 1-29

[5] Baum L., Petrie T. (1966) Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Statist.,* **37**, 1554-1563

[6] Bickel P., Ritov Y. (1996) Inference in hidden Markov models I : local asymptotic normality in the stationary case, *Bernoulli*, **2**, 199-228

[7] Bickel P., Ritov Y., Ryden T. (1998) Asymptotic normality of the maximum likelihood estimator for general hidden Markov models, *Ann. Statist.*, **26**, 1614-1635

[8] Bollerslev T. (1986) Generalized autoregressive conditional heteroscedasticity, *J. Econ.*, **31**, 307-327

[9] Bourlard H.A., Morgan N. (1994) *Connectionist speech recognition : a hybrid approach*, Kluwer academic publ.

[10] Boyer-Xambeu M.T., Deleplace G., Gaubert P., Gillard L., Olteanu M. (2006) The periodization of the international bimetalism : 1821-1873, *preprint*

[11] Bradley R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions, *Probability Surveys*, **2**, 107-144

[12] Chambaz A., Matias C. (2006) Number of hidden states and memory : a joint order estimation problem for Markov chains with Markov regime, *Preprint*

[13] Chan K., Ledolter J. (1995) Monte-Carlo EM estimation in time series models involving counts, *J. Amer. Statist. Assoc.*, **90**, 242-252

[14] Charles C. (1977) *Régression typologique*, Rapport de recherche no. 257, Le Chesnay, INRIA

[15] Chow G.C. (1960) Tests of equality between sets of coefficients in two linear regressions, *Econometrica*, **28**, 591-605

[16] Coe P. (2002) Financial crisis and the Great Depression : a regime switching approach, *Journal of Money, Credit and Banking*, **34(1)**, 76-93

[17] Cullogh W.M., Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics*, **5**, 115-133

[18] Dacunha-Castelle D., Gassiat E. (1997a) The estimation of the order of a mixture model, *Bernoulli*, **3**, 279-299

[19] Dacunha-Castelle D., Gassiat E. (1997b) Testing in locally conic models, *ESAIM Prob. and Stat.*, **1**, 285-317

[20] Dacunha-Castelle D., Gassiat E. (1999) Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes, *The Annals of Statistics*, **27(4)** , 1178-1209

[21] DeJong P., Shephard N. (1995) The simulation smoother for time series models, *Biometrika*, **82(2)**, 339-350

[22] Dempster A.P., Laird N.M., Rubin D.B (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statist. Soc. (B)*, **39(1)**, 1-38

[23] DeSarbo W.S., Cron W.L. (1988) A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, **5**, 249-282

[24] Douc R., Matias C. (2001) Asymptotics of the maximum likelihood estimator for general hidden Markov models, *Bernoulli*, **7(3)**, 381-420

[25] Douc R., Moulines E., Ryden T. (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime, *Ann. Statist*, **32(5)**, 2254-2304

[26] Doukhan P. (1995) *Mixing : Properties and Examples*, New York, Springer-Verlag

[27] Doukhan P., Massart P. Rio E. (1995) Invariance principles for absolutely regular empirical processes, *Ann. Inst. Henri Poincaré*, **31(2)**, 393-427

[28] Durbin R., Eddy S.R., Krogh A., Mitchison G. (1998) *Biological Sequence Analyses : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press

[29] Engle R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation, *Econometrica*, **50**, 987-1008

[30] Francq Ch., Roussignol M. (1998) Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator, *Statistics*, **32(2)**, 151-173

[31] Francq C., Roussignol M., Zakoian J-M. (2001) Conditional heteroskedasticity driven by hidden Markov chains, *Journal of Time Series Analysis*, **22**, 197-220

[32] Gassiat E., Keribin C. (2000) The likelihood ratio test for the number of components in a mixture with Markov regime, *ESAIM P&S*, **4**, 25-52

[33] Gassiat E. (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst. Henri Poincaré*, **38**, 897-906

[34] Hamilton J.D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, **57**, 357-384

[35] HAMILTON J.D. (1990) Analysis of time series subject to changes in regime, *J. Econometrics*, **45**, 39-70

[36] HAMILTON J.D. (1994) *Time Series Analysis*, Princeton University Press

[37] HAMILTON J.D., SUSMEL R. (1994) Autoregressive conditional heteroskedasticity and changes in regime, *Journal of Econometrics*, **64**, 307-333

[38] HAMILTON J.D., RAJ B (2002) *Advances in Markov-Switching Models*, Springer-Verlag

[39] HENNA J. (1985) On estimating the number of constituents of a finite mixture of continuous distributions, *Ann. Inst. Statist. Math.*, **37**, 235-240

[40] HENNIG C. (2000) Regression fixed point clusters : motivation, consistency and simulations, *Preprint 2000-02*, Fachbereich Mathematik, Universitat Hamburg

[41] HOLST U., LINDGREN G., HOLST J., THUVESHOLMEN M. (1994) Recursive estimation in switching autoregressions with a Markov regime, *J. Time. Ser. Anal.*, **15**, 489-506

[42] HORNIK K., STINCHCOMBE M., WHITE H. (1989) Multilayer feedforward networks are universal approximators, *Neural Networks*, **4**, 359-366

[43] IZENMAN A.J., SOMMER C. (1988) Philatelic mixtures and multivariate densities, *Journal of the American Stat. Assoc.*, **83**, 941-953

[44] JENSEN J.L., PETERSEN N.V. (1999) Asymptotic normality of the maximum likelihood estimator in state space models, *Ann. Statist.*, **27(2)**, 514-535

[45] KERIBIN C. (2000) Consistent estimation of the order of mixture models, *Sankhya : The Indian Journal of Statistics*, **62**, 49-66

[46] KIM S., SHEPHARD N., CHIB S. (1998) Stochastic volatility : Likelihood inference and comparison with ARCH models, *Review of Economic Studies*, **65**, 361-393

[47] KOHONEN T. (1984) *Self-organization and Associative Memory*, Springer-Verlag

[48] KRISHNAMURTHY V., RYDEN T. (1998) Consistent estimation of linear and non-linear autoregressive models with Markov regime, *J. Time. Ser. Anal.*, **19(3)**, 291-307

[49] LECUN Y. (1985) Une procédure d'apprentissage pour réseau à seuil assymétrique, *Cognitiva*, **85**, 599-604

[50] LE GLAND F., MEVEL L. (2000a) Basic properties of the projective product with application to products of column-allowable nonnegative matrices, *Math. Control Signals Systems*, **13(1)**, 41-62

[51] LE GLAND F., MEVEL L. (2000b) Exponential forgetting and geometric ergodicity in hidden Markov models, *Math. Control Signals Systems*, **13(1)**, 63-93

[52] LEE J.A., VERLEYSEN M. (2005) Generalization of the Lp norm for time series and its application to Self Organizing Maps, *Proceedings of WSOM'05*, Paris (France), 733-740

[53] LEROUX B.G. (1992a) Maximum-likelihood estimation for hidden Markov models, *Stochastic Process. Appl.*, **40**, 127-143

[54] LEROUX B.G. (1992b) Consistent estimation of a mixing distribution, *The Annals of Statistics,* **20**, 1350-1360

[55] LETREMY P. (2000) Notice d'installation et d'utilisation de programmes basés sur l'algorithme de Kohonen et dédiés à l'analyse de données, *Prepub. SAMOS 131*

[56] LINDSAY B.G. (1983) Moment matrices : application in mixtures, *The Annals of Statistics*, **17**, 722-740

[57] LIU X., SHAO Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, **31(3)**, 807-832

[58] LOU S., JIANG J., KENG K. (1993) Clustering objects generated by linear regression models, *Journal of the American Statistical Association*, **88**, 1356-1362

[59] MAC DONALD L., ZUCCHINI W. (1997) *Hidden Markov and other models for discrete-valued time series*, London New-York Chapman&Hall

[60] MAILLET B., MICHEL TH. (2002) Quelle a été l'ampleur de la crise financière de Septembre 2001 ? Une mise en perspective, *Revue d'Economie Financière*, **67**, 269-276

[61] MAILLET B., MICHEL TH. (2003) An index of market shocks based on multiscale analysis, *Quantitative Finance*, **3(2)**, 88-97

[62] MAILLET B., OLTEANU M., RYNKIEWICZ J. (2004) Caractérisation des crises financières à l'aide de modèles hybrides HMC-MLP, *Revue d'Economie Politique*, **114(4)**, 489-506

[63] MILLNERT M. (1986) Identification of ARX models with Markovian parameters, *Int. J. Control*, **45**, 2045-2058

[64] OLTEANU M. (2006) A descriptive method to evaluate the number of regimes in a switching autoregressive model, *Neural Networks*, **19**, 963-972

[65] OLTEANU M., RYNKIEWICZ J. (2006) Estimating the number of regimes in a switching autoregressive model, *Prepub. SAMOS 245*

[66] PRESS W.H. ET ALII. (1992) *Numerical Recipes in C : The art of scientific computing*, Cambridge University Press

[67] RABEMANANJARA R., ZAKIOAN J. (1993) Threshold ARCH models and asymmetries in volatility, *Journal of Applied Econometrics*, **8(1)**, 31-49

[68] RABINER R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77(2)**, 257-286

[69] REDNER R.A., WALKER H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26(2)**, 195-239

[70] ROEDER K. (1994) A graphical technique for determining the number of components in a mixture of normals, *Journal of the American Stat. Assoc.*, **89**, 487-495

[71] ROSENBLATT F. (1962) *Principles of neurodynamics*, Spartan, New York

[72] RUMELHART D.E., MCCLELLAND J.L. (1986) *Parallel distributed processing : exploration in the microstructure of cognition*, vol. 1, 318-362, MIT Press/ Bradford Books

[73] RYDÉN T. (1995) Estimating the order of hidden Markov models, *Statistics*, **26**, 345-354

[74] RYNKIEWICZ J. (2000) *Modèles hybrides intégrant des réseaux de neurones artificiels à des modèles de chaînes de Markov cachées : application à la prédiction de séries temporelles*, Thèse de doctorat

[75] SCHWARZ G. (1978) Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464

[76] SORNETTE D. (2000) Self-organized "slimming" of power law tails by increasing market returns, *working paper*

[77] SPATH H. (1979) Clusterwise linear regression, *Computing*, **22**, 367-373

[78] TEICHER H. (1963) Identifiability of finite mixtures, *Ann. Math. Statist.*, **34(2)**, 1265-1269

[79] TONG H. (1983) *Threshold models in non-linear time series analysis*, Lecture Notes in Statistics, 21, Springer, Heidelberg

[80] TOYODA T. (1974) Use of the Chow test under heteroscedasticity, *Econometrica*, **42**, 601-608

[81] VAN DER VAART A. (2000) *Asymptotic Statistics*, Cambridge University Press, New York

[82] VOLKONSKII V.A., ROZANOV Y.A. (1959) Some limit theorems for random functions, Part I, *Theory Probab. Appl.*, **4**, 178-197

[83] WEIGEND A.S., GERSHENFELD N.A. (1993) *Time series prediction*, Addison-Wesley

[84] WEIGEND A.S., MANGEAS M., SRIVASTA A. (1995) Nonlinear gated expert for time series : discovering regimes and avoiding overfitting, *International Journal of Neural Systems*, **6(4)**, 373-399

[85] WONG C.S., LI W.K. (2000) On a mixture autoregressive model, *J. Royal Statist. Soc. Series B*, **62(1)**, 95-115

[86] YAO J.-F., ATTALI J.-G. (2000) On stability of nonlinear AR processes with Markov switching, *Adv. in Applied Probability*, **32(2)**, 394-407

[87] YULE U. (1927) On a method of investigating periodicities in disturbed series with special reference to Wolfers's sunspot numbers, *Philos. Trans. Royal Soc. London, Series A*, **226**, 267-298

[88] ZUMBACH G., DACOROGNA M., OLSEN J., OLSEN R. (2000a) Measuring shocks in financial markets, *International Journal of Theoretical and Applied Finance*, **3(3)**, 347-355

[89] ZUMBACH G., DACOROGNA M., OLSEN J., OLSEN R. (2000b) Shock of the news, *Risk*, March 2000, 110-114