FRANCK ARNAUD Version du 1er mai 2012 à 15h38

# Markov Switching Models

Les modèles à changements de régime markoviens ont connu un fort développement depuis leur redécouverte par James Hamilton à la fin des années 1980. À cette époque, les macro-économètres disposaient de peu d'outils de modélisation des séries temporelles hors des modèles ARIMA. Hamilton (1989), reprenant et améliorant des travaux de Quandt, propose un modèle non-linéaire mais stationnaire du PNB américain, développe la théorie et l'estimation par maximum de vraisemblance. Il expose aussi l'impact de ce nouveau modèle sur la croissance de long-terme, et l'intérêt pour la datation du cycle économique. Depuis ces travaux, ces modèles ont été employés à de multiples occasions, dans les domaines macroéconomiques et financiers, et de nombreuses extensions ont été envisagées ; la synthèse avec les modèles à facteurs communs, par exemple, est fructueuse.

Actuellement, ces modèles sont utilisés par tous les organismes de conjoncture comme un outil indispensable à l'analyse du cycle économique. C'est une technique complémentaires à d'autres, telle que la décomposition entre tendance et cycle. Les modèles à changement de régimes markoviens ne fournissent que peu d'aide pour décrire le futur à court terme.

Ce document ne contient probablement pas de nouveautés, ce n'est pas son objectif : il a une portée pédagogique. Il s'agit en effet d'une synthèse, nécessairement incomplète, sur cette classe de modèles. Dans une première partie, un modèle très simple est présenté : la souplesse de la modélisation est déjà visible. Les propriétés de ce modèle et plusieurs méthodes d'estimation (ainsi que des considérations sur les choix de modèles) sont détaillées, puis appliquées à des exemples concrets, dans les domaines macro-économiques et financiers. Cette longue introduction constitue une sorte de factorisation des éléments communs à cette modélisation. Les parties suivantes reviennent sur des enrichissements de ce modèle : la seconde partie généralise au cas de la régression multivariée, et en intégrant une dynamique auto-régressive, la suivante aux données qualitatives. La quatrième partie est consacrée à l'intersection entre les modèles à facteurs et les modèles à changements de régimes markoviens. L'annexe est particulièrement développée : y sont relégués (de façon quelque peu désordonnée) de nombreux points d'importance variée. Néanmoins, on y trouve une section présentant (plus ou moins) rapidement un grand nombre de publications, structurée autour des sous-classes de modélisation.

**Ce document est en cours de rédaction. Les passage en rouge ont vocation à disparaître.** N'hésitez pas à me faire part de toute remarque : `franck.arnaud@ensae.org`.

# Table des matières

# Chapitre 1

# Introduction

Ce chapitre introductif présente un modèle de processus à changement de régime markovien très simple, destiné à fixer les idées. Un examen précis du modèle précède l'examen des avantages de cette modélisation, notamment pour le modélisateur. Les propriétés détaillées sont ensuite explicitées. Les procédures d'estimation sont ensuite détaillées ; de façon liée, on présente des outils de choix de modèles. Des applications de ce modèle simple illustrent les développements précédents ; elles portent sur la datation des cycles économiques, l'identification de régimes d'inflation ainsi qu'à l'analyse des rendements financiers.

## 1.1 Présentation

Les processus à changements de régime markovien sont très utilisés en macroéconométrie. Avant de préciser l'intérêt de cette modélisation, il apparaît nécessaire de définir rigoureusement le modèle statistique sous-jacent.

### 1.1.1 Présentation statistique

Introduits par Hamilton (1988, 1989), les processus à changements de régime markovien enrichissent l'analyse ARIMA classique en autorisant la série dépendante à suivre différents sous-modèles, les transitions entre ces différents modèles étant gérés à l'intérieur du cadre général. Voici un exemple assez général de processus : soit $S_t \in [\![1; M]\!]$ le processus d'état [1], inobservé ; la loi du processus $(Y_t)_t$ observé est déterminé par :

$$Y_t = \mu_{S_t} + \sum_{j=1}^{p} \Gamma_{j,S_t} Y_{t-j} + \Sigma_{S_t}^{1/2} \cdot \varepsilon_t$$

où $\varepsilon_t$ est un bruit blanc, communément (car commodément) supposé gaussien [2].

On suppose le processus d'état $S_t$ markovien, dans la mesure où les états intéressants sont fréquemment liés et persistents. Soit donc $S_t$ une chaîne de Markov homogène, irréductible, apériodique, et indépendante du processus $(\varepsilon_t)_t$ :

$$\forall i, j, \qquad \mathbb{P}(S_t = j | S_{t-1} = i, S_{\to t-2}, (\varepsilon_t)_t) = \mathbb{P}(S_t = j | S_{t-1} = i) = p_{ij}$$

où la matrice $M \times M$ de transition $P = (p_{ij})_{ij}$ est stochastique :

$$\forall i, j \quad p_{ij} \in [0; 1] \qquad \text{et} \qquad \forall i, \quad \sum_j p_{ij} = 1 \tag{1.1}$$

On note $\mathbb{1}_M$ le vecteur-colonne de $\mathbb{R}^M$ formé de 1, de sorte que la deuxième contrainte s'écrit : $P\mathbb{1}_M = \mathbb{1}_M$. Outre la matrice de transition, la distribution initiale de $S_1$ est un paramètre du processus d'état. La loi stationnaire [3] associée à une chaîne de Markov irréductible est un choix de loi initiale, mais ce n'est pas le seul.

1. Notation standard : $\forall a, b \in \mathbb{R}, [\![a; b]\!] = [a; b] \cap \mathbb{N}$.
2. Quoique d'autres modélisations soient envisagées : lois de Student, par exemple Hamilton (2005)
3. La loi stationnaire, identifiée au vecteur colonne $\pi^\infty$ des $\pi_i^\infty = \mathbb{P}(S_\infty = i)$, vérifie par définition $\pi' P = \pi'$. Toutes choses égales par ailleurs (ce qui n'a pas grand sens en soi puisque $P\mathbb{1}_M = \mathbb{1}_M$), $\pi_i$ est une fonction croissante de $p_{ii}$ et une fonction décroissante des $p_{ij}$, pour $j \neq i$.

La loi du processus dépend de la loi $\varepsilon_t$ ainsi que des paramètres de la chaine. On peut la calculer par récurrence :

$$\begin{aligned}\mathbb{P}(S_{1\to t}, Y_{1\to t}) &= \mathbb{P}(Y_t|Y_{1\to t-1}, S_{1\to t}) \cdot \mathbb{P}(S_t|S_{1\to t-1}, Y_{1\to t-1}) \cdot \mathbb{P}(S_{1\to t-1}, Y_{1\to t-1}) \\ &= \mathbb{P}(Y_t|Y_{t-p\to t-1}, S_t) \cdot \mathbb{P}(S_t|S_{t-1}) \cdot \mathbb{P}(S_{1\to t-1}, Y_{1\to t-1})\end{aligned}$$

Notons dès à présent que la stationnarité de $Y$ dépend de $S$ : si la loi marginales de $S$ n'est pas $\pi^\infty$ alors $Y$ n'est pas stationnaire.

Dans la suite de ce chapitre, on se concentre sur un cas simple, car la présence d'exogènes (dont l'impact sur le processus dépend ou non du régime) ou de termes *retardés*[4] (lag) n'affecte pas sensiblement les conclusions. Le modèle plus général fait l'objet du chapitre suivant. Nous nous intéressons donc ici au modèle :

$$Y_t = \mu_{S_t} + \sigma_{S_t}\varepsilon_t \tag{1.2}$$

Autrement dit, $Y_t|S_t$ suit une loi normale de moyenne $\mu_{S_t}$ et d'écart-type $\sigma_{S_t}$ :

$$Y_t|S_t \rightsquigarrow \mathcal{N}(\mu_{S_t}, \sigma^2_{S_t}) \tag{1.3}$$

En dépit de la simplicité de sa formulation, ce modèle présente des propriétés stochastiques riches. À titre d'exemple, suivant les valeurs des paramètres, l'applatissement (kurtosis) peut être élevé ou faible : cette caractéristique a motivé l'application de cette modélisation aux données financières.

Le modèle $(1.2)$ admet plusieurs sous-modèles :

- moyennes identiques : $\mu_m \equiv \mu$, auquel cas seuls les écarts-types diffèrent. A l'instar des modèles ARCH et dérivés (GARCH), ils introduisent de l'hétéroscédasticité ; mais sous une forme plus simple que les ARCH, puisque l'ensemble des variances est non pas continu mais fini (et de taille faible).
- variances identiques : $\sigma^2_m \equiv \sigma^2$, auquel cas les régimes diffèrent uniquement par leur moyenne. Hamilton et Chauvet (2005) appliquent ces modèles aux données macroéconomiques, et en déduisent un indicateur de retournement du cycle économique.
- moyenne et variance identique : $\mu_m \equiv \mu$ et $\sigma^2_m \equiv \sigma^2$, auquel cas le processus est un bruit blanc.

Il sera intéressant, lors de l'étape de choix de modèle, de pouvoir tester ces sous-modèles contre le modèle général.

Une trajectoire d'un tel processsus est représenté sur la figure 1.1. Le processus présente trois régimes, dont les moyennes sont $\mu = (-1, 0, 1)$, les variances $\sigma^2 = (1, 2, 3) \times 10^{-1}$, et les probabilités de transition

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

Sur la figure 2 (prog AA), on a représenté les estimations de densité conditionnelles au régime. Les densités conditionnelles ajustent de près les densités gaussiennes théoriques, de moyenne et variance connues. Les moyennes et variances empiriques conditionnelles figurent aussi dans le tableau suivant.

| | Moyenne | Variance |
|---|---|---|
| Etat 1 | $-0.98$ | $0.11$ |
| Etat 2 | $0.00$ | $0.21$ |
| Etat 3 | $1.03$ | $0.27$ |

### 1.1.2 Avantages de cette modélisation

Les processus à changement de régime markovien présentent des intérêts statistiques ainsi que pour le modélisateur.

4. Il en va autrement des termes *avancés* (lead) : leur présence change en revanche tout.

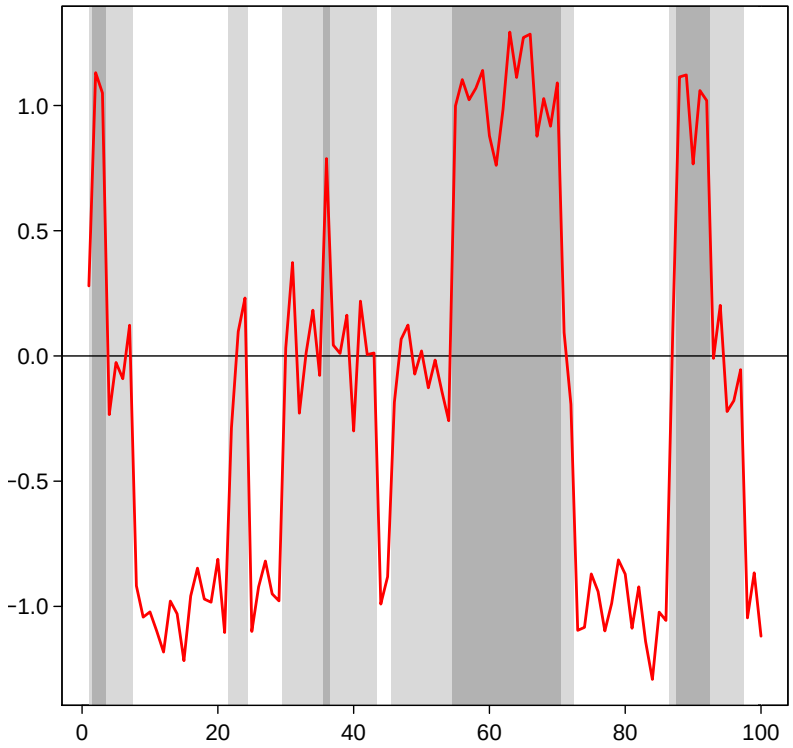


FIGURE 1.1 – Simulation d'un processus simple

#### 1.1.2.1 Intérêt statistique

**1.1.2.1.1 Spécificités de ces modèles** Les modèles Markov Switching enrichissent l'analyse des séries temporelles, encore peu développée au-delà des processus ARMA à la fin des années 1980 :

- les processus MS généralisent la modélisation ARMA en autorisant des changements de régimes dans les paramètres. La section 7 de Hamilton (1989) propose quelques éléments de réflexion.
- ils relèvent d'une approche différente des processus à changements de régime existant à l'époque tels que les modèles dits Threshold Auto Regressive (TAR) (cf Tong (198 ?), ou Fan et Yao (2005) pour une présentation rapide), qui reposent sur l'hypothèse que le régime dépend du passé du processus. À l'inverse, Hamilton [5] considère que

> changes in regime are the result of processes largely unrelated to past realizations of the series and are not themselves directly observable

La modélisation Markov Switching répond explicitement à cette objection.

**1.1.2.1.2 Nécessité de ces modèles** Supposons que le Data Generating Process (DGP) comprenne effectivement des changements de régime, dans quelle mesure l'estimation par des méthodes usuelles (type ARMA/OLS) conduisent-elles à une mauvaise inférence ? Plusieurs travaux ont révélé que l'interprétation pouvait être tout à fait faussée (cf Ceccheti, Lam et Mark (1990) par exemple).

#### 1.1.2.2 Intérêt pour le modélisateur

Ces modèles ont été appliqués principalement dans deux domaines :

- en macro-économie, ils sont utilisés pour l'analyse du cycle des affaires : ils autorisent en particulier une détection rapide des retournements de conjoncture,
- en finance : ils sont utilisés comme une modélisation alternative des variances distinctes, par exemple
- ces modèles ont aussi été appliqués aux prix de l'énergie, et à d'autres champs d'études.

5. cf Hamilton (1994) page 234.

**1.1.2.2.1 Développements macroéconomiques** Hamilton (2005) démontre statistiquement l'intérêt de ces processus pour l'analyse économique : leur adéquation aux données informe l'économiste théoricien sur les mécanismes de transmission des chocs dans l'économie. Par exemple, Hamilton montre que les hausses et les baisses du taux de chômage ont probablement des explications assez différentes ; la comparaison de modélisation de variables réelles et monétaires informe sur le sens de la causalité : la sphère réelle serait influencée par la sphère monétaire.

**1.1.2.2.1.1 'Transitory vs permanent'** Hamilton (1989) (section 8 page 378) montre que ces modèles enrichissent le débat sur 'transitory versus permanent' [6]. Considérons $y_t = \log \mathrm{PIB_t}$ ; à long terme, le taux de croissance est constant : $\mathbb{E}\Delta y_t = \mathbb{E}\Delta \log \mathrm{PIB_t} = \mu'\pi$ n'est pas affecté par un changement d'état. À l'inverse, un changement d'état modifie le *niveau* du PIB :

$$\lim_{h\to\infty} [\mathbb{E}(y_{t+h}|S_t = 2) - \mathbb{E}(y_{t+h}|S_t = 1)] = \frac{\lambda}{\lambda - 1}(\mu_2 - \mu_1)$$

où $\lambda = p_1 + p_2 - 1$ est la seconde valeur propre de $P$. Soit $\pi^0$ la loi initiale du processus d'état, $S_0$, et $\mathbb{E}_{\pi^0}$ la loi du processus associé, alors [7] :

$$\begin{aligned} \mathbb{E}_{\pi^0} \log \mathrm{PIB_t} &= \sum_{\tau=1}^{t}\sum_{s=1}^{M} \mu_s \mathbb{P}(S_\tau = s) = \sum_{s=1}^{M} \mu_s \sum_{\tau=1}^{t} \left[\pi_s^\infty + \lambda^\tau(\pi_s^\infty - \pi_s^0)\right] \\ &= t\sum_{s=1}^{M} \mu_s \pi_s^\infty + \sum_{s=1}^{M} \mu_s(\pi_s^\infty - \pi_s^0)\lambda\frac{1-\lambda^{t+1}}{1-\lambda} \end{aligned} \tag{1.4}$$

en vertu de $(B.2)$ (cf section B.1 page 49). Il ne reste qu'à conclure :

$$\begin{aligned} \mathbb{E}(y_{t+h}|S_t = 2) - \mathbb{E}(y_{t+h}|S_t = 1) &= \mathbb{E}_{\pi^0=(0,1)}(\log \mathrm{PIB_{t+h}}) - \mathbb{E}_{\pi^0=(1,0)}(\log \mathrm{PIB_{t+h}}) \\ &= \lambda\frac{1-\lambda^{t+h+1}}{1-\lambda}(\mu_2 - \mu_1) \end{aligned}$$

À titre d'exemple, Hamilton (1989) estime que la transition d'une récession à une phase d'expansion augmente le niveau de long terme de 3%. Détails en section A.6.1 page 44.

**1.1.2.2.1.2 Datation des cycles économiques** La croissance économique connaît des variations autour d'une tendance [8]. Il est intéressant de décomposer la croissance effective en une composante de long terme (la croissance potentielle) et ses variations cycliques. Les macroéconomètres ont développé de nombreux outils pour réaliser cette décomposition. La modélisation à changement de régime markovien en fait partie. Elle consiste à identifier les phases de récession et d'expansion de l'économie.

Cette approche est complémentaire à celle retenue par le *Business Cycle Datation Committee* du *National Bureau of Economic Research* américain. Ce comité publie les dates de référence de retournement du cycle. La méthode de datation privilégie l'examen qualitatif et quantitatif de 5 grands indicateurs économiques, quoique l'ensemble des indicateurs disponibles soient bien entendus pris en compte. Comme les datations sont publiées avec un certain retard (un an de délai au minimum), des techniques économétriques ont été développées pour détecter le plus tôt possible et à la limite en temps réel les cycles économiques.

Plusieurs méthodes existent [9]. Chauvet et Hamilton (2005) légitime la méthode reposant sur les modèles à sauts markoviens.

**1.1.2.2.1.3 Réduction de la volatilité du PIB américain** La volatilité macroéconomique constitue un environnement négatif à la croissance, pour des agents averses au risque. La littérature macroéconométrique a traité ce sujet de façon extensive. La question de la réduction de la volatilité autour de la seconde guerre mondiale a fait couler beaucoup d'encre : Delong et Summers (1986) (par exemple) soutiennent cette position, tandis que Christina Romer avance que leurs conclusions tiennent beaucoup à des problèmes de mesure. Un consensus s'est depuis dégagé sur ce sujet : réduction de la volatilité, mais de moindre ampleur

6. Voir Nelson et Plosser (1982), Campbell et Mankiw (1987).
7. On suppose que $y_0 = 0$ : il suffit de le soustraire dans les deux termes.
8. Depuis les travaux de Nelson et Plosser, cette tendance n'est plus considérée comme déterministe et linéaire, mais aléatoire et intégrée.
9. Bry-Boschan et Harding-Pagan considèrent les extrema locaux comme les points de retournement du cycle, par exemple.

que ne l'indiquent les statistiques officielles. Backus et Kehoe ont dressé un bilan international sur ce point.

Depuis, la même question a été transposée sur la période après-guerre, avec deux avantages : les données sont plus fiables, et disponibles trimestriellement (et non plus annuellement). Les modèles à changements de régime markovien ont été mobilisés pour étudier cette question. Kim et Nelson (1999) par exemple déduisent d'une telle modélisation [10] du PIB américain que la volatilité macroéconomique américaine a connu une réduction après-guerre : la variance après et avant le premier trimestre 1984 aurait réduite d'un facteur 4.

**1.1.2.2.1.4 Long swings in the dollar** La littérature sur les taux de change a bénéficié de la modélisation des processus à changements de régime markovien. Il n'est pas question de se livrer ici à une analyse détaillée de cette littérature, mais nous pouvons toutefois donner deux éléments de contexte :

- Meese et Rogoff (1983) montre que le marche aléatoire constitue le meilleur modèle de prévision possible des taux de change. La meilleure prévision possible du taux de change est donc la dernière valeur connue : cette méthode domine dans de nombreuses institutions.
- entre 1977 et 1988, le dollar avait connu trois phases (cf figure 1.2) :
  - dépréciation entre 1977 et fin 1980
  - puis 4 années d'appréciation, entre 1981 et 1984,
  - et enfin une nouvelle phase de dépréciation

  Or l'alternance de telles phases était incompatible avec la théorie existant alors.

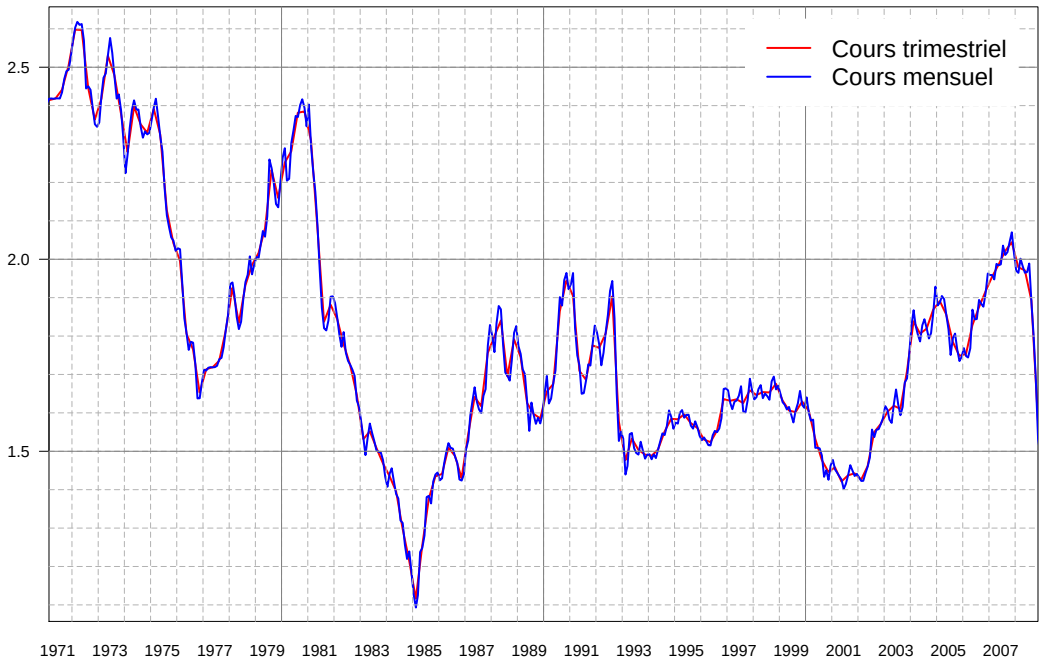


FIGURE 1.2 – Cours de la livre britannique contre le dollar américain (1 GBP=... USD)

Engel et Hamilton (1990) ont appliqué la modélisation markovienne au taux de change (trimestriel) du dollar contre le deutsch mark, avec un modèle très simple $Y_t = \mu_{S_t} + \sigma_{S_t}\varepsilon_t$. Ils interprètent le bon ajustement comme une amélioration de Meese et Rogoff (1983) : ils enrichissent en effet le diagnostic en montrant que le taux de change alterne entre deux régimes [11]. Ils montrent finalement que leur modèle constitue un bon outil de prévision **montrent-ils que c'est une meilleur outil que la marche aléatoire ?**

Depuis, de nombreux travaux empiriques se sont succédé : Engel, Mark et West (2007) constitue probablement un état récent de la littérature. Klaasen (2005) a poursuivi dans la voie ouverte par Engel et Hamilton (1990) et corroboré leurs résultats.

10. Présentée rapidement en annexe, section A.1.2.1.
11. Hamilton (1996) revient sur le sujet, uniquement pour mettre en application ses procédures de test, et montre que le modèle proposé dans Engel et Hamilton (1990) ne présente pas de défaut de spécification majeur, au sens où il ne manque pas d'auto-corrélation, entre autres.

## 1.2 Propriétés

Les modèles ARMA ont été créés par Box et Jenkins pour répliquer les autocorrélations et les autocorrélations partielles. Ces modèles s'intéressent donc essentiellement à la dimension temporelle du processus, plus qu'aux distributions instantannées. Les modèles à changements de régime markovien intègrent au contraire les deux dimensions : construits pour refléter des distributions, ils tiennent aussi compte de la dimension temporelle, par le biais du processus inobservable $(S_t)_t$. La découverte principale est qu'une dépendance temporelle peu élaborée, de type markovienne, et une modélisation statique simple, de type gaussienne, produise une classe de modèle aux propriétés riches. On montre en effet dans cette section que ces modèles peuvent être unimodaux ou bi-modaux, présenter des queues de distribution épaisses, de la fausse longue mémoire, etc.

### 1.2.1 Stationnarité

Les processus, admettant la représentation (1.2) sont stationnaires au sens faible, dès lors que $S_1 \overset{(d)}{=} S_\infty \overset{(d)}{=} \pi^\infty$ (où $\overset{(d)}{=}$ indique l'égalité des lois) :

– la moyenne ne dépend pas du temps :

$$\forall t, \qquad \mathbb{E}(Y_t) = \sum_s \mathbb{E}(Y_t|S_t = s) = \sum_s \mu_s \pi_s = \pi'\mu$$

– la covariance entre $Y_s$ et $Y_t$ dépend uniquement de $t-s$. Soit $s < t$, et $\tilde{Y}_t = Y_t - \mathbb{E}Y_t$ de moyenne $\nu = \mu - (\pi'\mu)\mathbb{1}$, alors :

$$\begin{aligned} \operatorname{cov}(Y_s, Y_t) = \mathbb{E}(\tilde{Y}_s\tilde{Y}_t) &= \sum_{i,j} \mathbb{P}(S_s = i, S_t = j)\mathbb{E}(\tilde{Y}_s\tilde{Y}_t|S_s = i, S_t = j) \\ &= \sum_{i,j} \pi_i (P^{t-s})_{ij} \nu_i \nu_j \\ \Longrightarrow \operatorname{cov}(Y_s, Y_t) &= (\pi \odot \nu)' P^{t-s} \nu \end{aligned}$$

où $\odot$ désigne la multiplication terme à terme de deux vecteurs de même dimension : $(a_i) \odot (b_i) = (a_i b_i)$.

En particulier, si $\mu_\cdot \equiv \mu$, alors $\nu = 0$ et le processus n'est pas corrélé : la corrélation provient uniquement du régime dans la moyenne.

La variance vaut $\pi'(\nu^{\odot 2} + \sigma^2)$.

Les autocorrélations sont asymptotiquement nulles [12]. De plus elles décroisssent exponentiellement vite vers 0 :

$$\exists \beta \in [0;1[, c > 0 \qquad / \qquad \forall h, \quad |\gamma(h)| \leq c \cdot \beta^h \tag{1.5}$$

où $\beta$ est la seconde valeur propre de $P$. Cette décroissance est intuitivement d'autant plus rapide que la chaîne mélange : si la probabilité de ne pas bouger $\mathbb{P}(S_1 = S_0)$ est élevée, alors la chaîne converge plus lentement. Notons $\bar{p}$ ce paramètre de la chaîne de Markov :

$$\bar{p} = \mathbb{P}(S_1 = S_0) = \sum_s \pi_s p_{ss}$$

**Cas particulier des AR switching** Ce paragraphe fait exception à la règle, puisque le cas particuliers des modèles avec retards est abordé ; soit en effet :

$$Y_t = \phi_{S_t} Y_{t-1} + \varepsilon_t$$

on montre que la stationnarité nécessite des conditions particulières sur les coefficients $(\phi_m)_m$. Dans le cas non-switching, ces conditions sont du type $|\phi| < 1$. Ici, on montre que **Francq et Zakoian ? autre ?** que la condition sur les coefficients est plus souple. Une condition nécessaire est : **développer**

12. En effet, $P^h \underset{h\to\infty}{\longrightarrow} \mathbb{1}_M \pi'$, de sorte que :

$$\gamma(h) = (\pi \odot \nu)' P^h \nu \underset{h\to\infty}{\longrightarrow} (\pi \odot \nu)' \mathbb{1}\pi'\nu = 0$$

puisque $\pi'\nu = 0$.

### 1.2.2 Multimodalité de la distribution

Le processus $Y_t$ est un mélange de gaussienne : suivant la position des paramètres, la densité de $Y$ est soit unimodale soit multi-modale. Cette caractéristique est intéressante, car le modélisateur rencontre parfois des distributions. Une question intéressante est : peut-on partitionner l'espace des paramètres $\Theta$ en parties $\Theta_i$ où $\theta \in \Theta_i$ signifie que toutes les distributions admettent exactement $i$ modes ? Ray et Lindsay (2005) ont résolu le problème dans le cas général.

Dans le cas d'un mélange de 2 gaussiennes à variances identiques, la figure 1.3 identifie les paramètres générant des distributions bimodales ou unimodales. Deux paramètres déterminent le nombre de modes de la distribution :

- l'écart des moyennes, normalisé par l'écart-type (supposé commun) : $\tilde{\mu} = \frac{\mu_2 - \mu_1}{\sigma}$,
- la probabilité de mélange $\pi$

Intuitivement :

- plus $\tilde{\mu}$ est petit (en valeur absolue), plus la probabilité nécessaire pour faire survenir la bimodalité est élevée,
- plus $\pi$ est proche de 0 (ou de 1), moins un écart important n'est nécessaire pour générer une distribution unimodale.

Bien entendu, changer $\pi$ en $1-\pi$ ou $\tilde{\mu}$ en $-\tilde{\mu}$ ne modifie pas les conclusions. La figure ne fait que quantifier proprement l'intuition précédente.

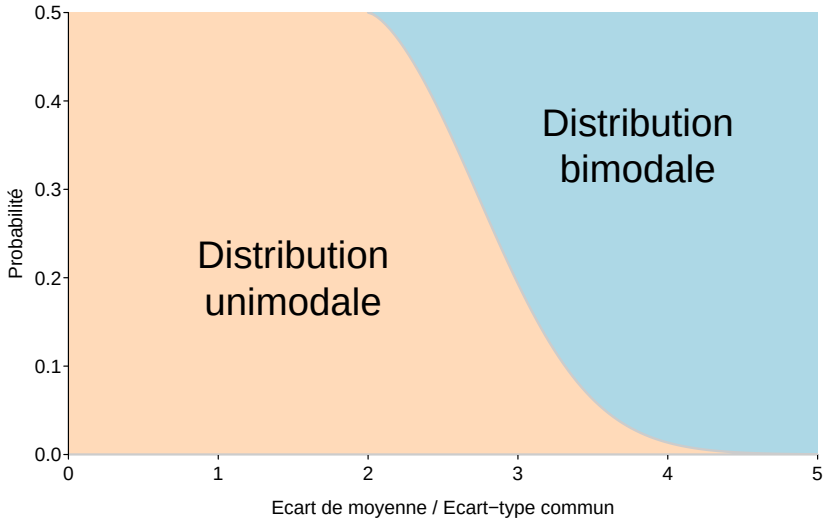


FIGURE 1.3 – Distribution unimodale ou bimodale

### 1.2.3 Moments, asymétrie, applatissement

De façon générale, on obtient aisément le moment centré en raisonnant sur le moment centré pour lequel on a remplacé la moyenne $\mu$ par $\nu = \mu - \pi'\mu\mathbb{1}$. On retrouve alors

$$\begin{array}{rcccl} m_2 & = & \mathbb{V}(Y_t) & = & \pi'(\nu^{\odot 2} + \sigma^2) \\ m_3 & = & \mathbb{E}\left[(Y_t - \mathbb{E}Y_t)^3\right] & = & \pi'\left[\nu^{\odot 3} + 3\nu \odot \sigma^2\right] \\ m_4 & = & \mathbb{E}\left[(Y_t - \mathbb{E}Y_t)^4\right] & = & \pi'\left[\nu^{\odot 4} + 6\nu^{\odot 2} \odot \sigma^2 + 3(\sigma^2)^{\odot 2}\right] \end{array}$$

On en déduit les coefficients d'asymétrie (skewness) $s = \frac{m_3}{m_2^{3/2}}$ et d'applatissement (kurtosis) $\kappa = \frac{m_4}{m_2^2}$.

Dans un modèle où moyenne ne dépend pas du régime, $\nu = 0$, $s = 0$ et $\kappa = 3\sum_s \pi_s(\tilde{\sigma}^2)^2$ où $\tilde{\sigma}^2 = \frac{\sigma^2}{\sum_s \pi_s \sigma_s^2}$. Cette classe de modèle conduit uniquement à des distributions épaisses [13], et toute valeur de

13. En vertu de l'inégalité de Cauchy-Schwart :

$$\sum_s \pi_s \sum_s \pi_s(\tilde{\sigma}_s^2)^2 \geq \left(\sum_s \pi_s \tilde{\sigma}_s^2\right) = 1$$

kurtosis peut être atteinte [14].

Détaillons un autre cas simple :
- le processus d'état ne comporte que deux états ($M = 2$),
- l'état n'influence que la moyenne : $\sigma_\cdot \equiv \sigma$

Ainsi, $\forall m \in \{1, 2\}, \quad Y_t|S_t = m \rightsquigarrow \mathcal{N}(\mu_m, \sigma^2)$. Soit $\Delta\mu = \mu_2 - \mu_1$, le kurtosis s'écrit alors (cf annexe C.1 page 51) :

$$\kappa = 3 + \frac{(\Delta\mu)^4 \pi_1 \pi_2}{\left(\sigma^2 + (\Delta\mu)^2 \pi_1 \pi_2\right)^2}(1 - 6\pi_1\pi_2)$$

de sorte que $\kappa \geq 3 \iff \pi_1\pi_2 \leq \frac{1}{6}$. Comme $\pi_2 = 1 - \pi_1$ :

$$\kappa \geq 3 \quad \iff \quad \pi_1 \text{ ou } \pi_2 \leq \pi^* = \frac{1}{2} - \frac{1}{\sqrt{12}} \approx 0.211$$

Ainsi, des queues de distribution épaisses nécessitent que la chaîne soit asymétrique.

### 1.2.4 Spurious long memory

Les autocorrélations des processus ARMA décroissent exponentiellement vite vers 0 :

$$\exists \beta \in ]0; 1[, c > 0 \quad / \quad \forall h \in \mathbb{N}, \quad |\gamma(h)| \leq c.\beta^h \tag{1.6}$$

À l'inverse, les autocorrélations empiriques des processus intégrés décroissent très lentement vers 0 : c'est un critère visuel immédiat d'identification des séries intégrées. Entre ces deux formes polaires, se trouvent les processus à mémoire longue : on peut les caractériser de plusieurs façons, par exemple par leurs autocorrélations. Un processus est dit à mémoire longue si ses autocorrélations décroissent hyperboliquement vers 0 :

$$\exists \delta \in ]0; \frac{1}{2}[, c > 0 \quad / \quad \gamma(h) \sim_{h\to\infty} c.|h|^{1-12d}$$

La décroissance des autocorrélations est plus lente que dans le cas exponentiel. Les séries à longue mémoire présentent des propriétés particulières : voir Beran (1992) ou Baillie (1996).

Les autocorrélations empiriques de certains processus MS induisent le modélisateur en erreur car elles décroissent lentement vers 0, suggérant la longue mémoire ou, à l'extrême, la non-stationnarité. Or nous avons déjà mis en évidence la décroissance exponentielle pour les processus markoviens (cf eq $(1.5)$). Dans le cas $M = 2$, le calcul est aisé [15] :

$$\gamma(h) = (\pi \odot \nu)' P^h \nu = b\lambda^h$$

où $\lambda = p_1 + p_2 - 1$ est de module inférieur à 1 ; $b$ dépend de $(p_1, p_2, \mu)$ : $b = (\mu_1 - \mu_2)\left[\pi_1\nu_1(p_1 - \pi_1) - \pi_2\nu_2(p_2 - \pi_2)\right]$.

La figure 1.4 présente les autocorrélations empiriques d'un processus MS avec deux états fortement récurrents : pour voir apparaître ce phénomène de *spurious long memory*, il est nécessaire que les états soient fortement persistants. Sur la base de ces autocorrélations, un modélisateur pourrait prendre une mauvaise décision quant à la structure de dépendance temporelle du processus.

### 1.2.5 Temps de séjour

Nous montrons ici que le caractère markovien du processus d'état implique que les durées de séjour sont distribuées géométriquement.

**Notations** Précisément, introduisons les processus $T_n$ et $D_n$ indiquant l'état du nième "spell" et la durée de ce spell. Autrement dit :

$$\underbrace{T_1, \ldots, T_1}_{\text{pendant } D_1 \text{ périodes}} \quad , \quad \underbrace{T_2, \ldots, T_2}_{\text{pendant } D_2 \text{ périodes}} \quad , \ldots$$

Pour définir les processus $T$ et $D$, il est pratique d'introduire deux autres processus :
- l'instant d'entrée dans le spell $n$, noté $E_n$,
- l'instant de sortie du spell $n$, noté $F_n$

Le terme de gauche n'est autre que $\kappa/3$.

14. En effet, avec $\sigma^2 = (1, 0_{M-1})$, $\kappa = \frac{3}{\pi_1}$.

15. Voir annexe B.1 page 49.

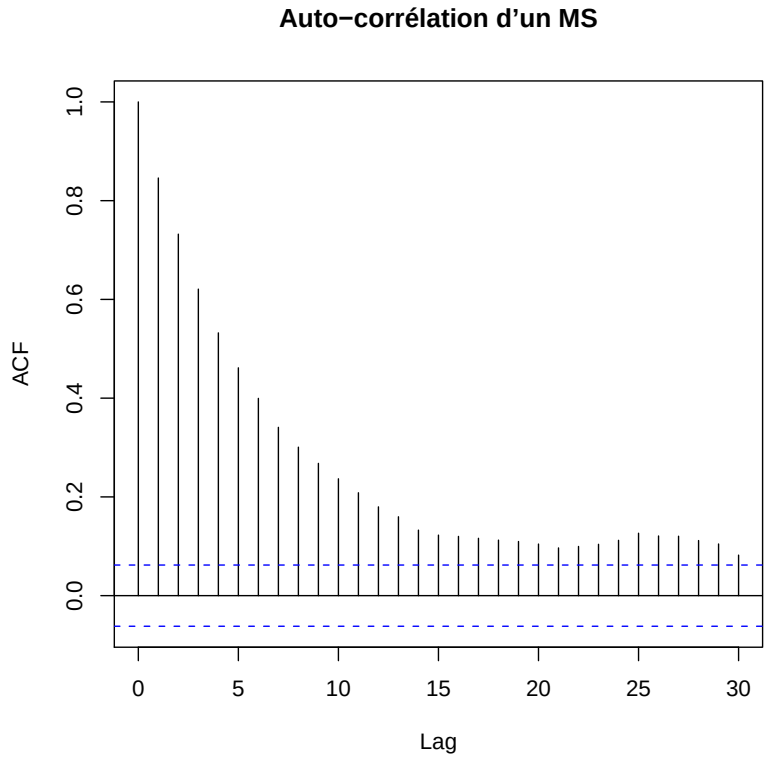


FIGURE 1.4 – Autocorrélations empiriques d'un MS

Le calcul de ces 4 processus (état $T$, durée $D$, entrée $E$ et sortie $F$) procède par récurrence à partir de $S$ :

- Initialisation : $\begin{cases} E_1 &= 1 \\ T_1 &= S_{E_1} = S_1 \end{cases}$
- Hérédité : $\forall n \in \mathbb{N}^*$,
  - $F_n = \inf\{l \geq E_n / S_{l+1} \neq T_n\}$
  - $E_{n+1} = F_n + 1$
  - $D_n = F_n - E_n + 1 = E_{n+1} - E_n$
  - $T_n + 1 = S_{E_{n+1}}$

**Exemple** Soit les 20 premières réalisations suivantes du processus d'état :

| $t$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S_t$ | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Construisons les processus $(T_n)_n$ et $(D_n)_n$ :

- le premier état est le 3 : $T_1 = 3$, on y reste pendant $D_1 = 4$ périodes,
- le deuxième spell est dans l'état $T_2 = 2$ et dure $D_2 = 4$,
- troisième spell : $T_3 = 1$ et $D_3 = 2$,
- enfin : $T_4 = 2$ et $D_4 \geq 10$.

#### 1.2.5.1 Loi des processus $T$ et $D$

Dans le cas markovien [16], on explicite aisément la loi du processus (cf annexe C.2 page 52) :

$$\mathbb{P}(T_1, \ldots, T_n, D_1, \ldots, D_n) = q(T_n).\prod_{i=2}^{n} p(T_{i-1}, T_i) \prod_{i=1}^{n} p(T_i)^{D_i - 1}.\pi(T_1) \tag{1.7}$$

[notations : $p(i,j)$ pour $p_{ij}$ et $q(i) = 1 - p_{ii}$]

16. Pas nécessairement stationnaire : la distribution de $S_1$ n'est pas imposée.

**1.2.5.1.1 Processus des durées** Les lois de dimension finies du processus des durées découlent de (1.7) :

$$\mathbb{P}(D_1,\dots,D_n|T_1,\dots,T_n)=\prod_{i=1}^{n}\left\{(1-p(T_i))\cdot p(T_i)^{D_i-1}\right\} \tag{1.8}$$

On déduit de (1.8) que :

- $\forall i\in[\![1;n]\!],\quad D_i|T_1,\dots,T_n\overset{(d)}{=}D_i|T_i$ :
$$\forall i\in[\![1;n]\!],\quad \mathbb{P}(D_i|T_1,\dots,T_n)=\mathbb{P}(D_i|T_i)$$
- si $i\neq j$, alors $D_i|T_i$ et $D_j|T_j$ sont indépendants :
$$i\neq j\quad\Longrightarrow\quad \mathbb{P}(D_i,D_j|T_i,T_j)=\mathbb{P}(D_i|T_i)\cdot\mathbb{P}(D_j|T_j)$$
- les temps de séjour sont distribués géométriquement : $\forall i\in[\![1;M]\!],\quad D_i|T_i\rightsquigarrow\mathcal{G}\left(1-p(T_i)\right)$ ; en outre, $D_i|T_i=s$ ne dépend pas de $i$.

Comme $\mathbb{E}\mathcal{G}(p)=p^{-1}$, $\mathbb{E}(D_n|T_n=m)=\frac{1}{1-p_{mm}}$ : le temps moyen de séjour croît avec $p_{mm}$. Plus un état est "fort", plus la chaîne y reste.

**1.2.5.1.2 Loi du processus d'état** Le processus $T$ est une nouvelle chaîne de Markov, dont la matrice de transition $\tilde{P}$ se déduit de la matrice de transition initiale $P$ :

$$\forall(i,j)\in[\![1;M]\!]^2,\quad \tilde{p}_{ij}=\begin{cases}0 & \text{si}\quad i=j\\ \dfrac{p_{ij}}{1-p_{ii}} & \text{si}\quad i\neq j\end{cases}$$

## 1.2.6 Inférence sur le processus inobservé

**Où insérer cette section ? Dans les propriétés ou dans l'estimation ; dans la partie estimation, ça fait un peu loin, il me semble que les probabilités lissées et filtrées ont un intérêt en dehors de l'estimation.**

**Débuter par une analogie avec le filtre de Kalman ?** Le modélisateur cherche à inférer l'état du processus sous-jacent $(S_t)_t$ sur la base de réalisations du processus observé $(Y_t)_t$. On note $S_{\tau|t}\in\mathbb{R}^M$ la distribution de probabilités $\left(\mathbb{P}(S_\tau=s|Y_{1\to t})\right)_{s\in[\![1;M]\!]}$ : $S_{\tau|t}=\mathbb{P}(S_\tau|Y_{1\to t})$. Bien entendu, $\mathbb{1}_M'S_{\tau|t}=\sum_{s=1}^{M}\mathbb{P}(S_\tau=s|Y_{1\to t})=1$.

L'étude de certaines de ces probabilités est privilégiée :
- les probabilités *filtrées*, $S_{t|t}=\mathbb{P}(S_t|Y_{1\to t})$, fournissent un diagnostic en temps réel[17],
- les probabilités *lissées*, $S_{t|T}=\mathbb{P}(S_t|Y_{1\to T})$, à l'inverse s'intéressent au passé : ayant constaté l'intégralité de l'échantillon $Y_1,\dots,Y_T$, on estime l'état passé $S_t$.

Étant données les lois des processus, ces probabilités sont aisées à calculer par récurrence.

### 1.2.6.1 Probabilités filtrées

Il s'agit ici de calculer, pour tout $t\in[\![1;T]\!]$ et tout $m\in[\![1;M]\!]$, la probabilité $\mathbb{P}(S_t=m|Y_{1\to t})$.

**1.2.6.1.1 Algorithme récursif de calcul des probabilités filtrées** L'algorithme repose sur l'exploitation extensive de la formule des probabilités conditionnelles : $\mathbb{P}(A,B)=\mathbb{P}(A|B)\cdot\mathbb{P}(B)$. En effet :

$$\begin{aligned}\mathbb{P}(S_t|Y_{1\to t}) &= \frac{\mathbb{P}(S_t,Y_{1\to t})}{\mathbb{P}(Y_{1\to t})}\\ &= \frac{\mathbb{P}(Y_t|S_t,Y_{1\to t-1})\cdot\mathbb{P}(S_t,Y_{1\to t-1})}{\mathbb{P}(Y_{1\to t})}\\ &= \frac{\mathbb{P}(Y_t|S_t)}{\mathbb{P}(Y_{1\to t})}\cdot\mathbb{P}(S_t,Y_{1\to t-1})\\ &= \mathbb{P}(Y_t|S_t)\frac{\mathbb{P}(Y_{1\to t-1})}{\mathbb{P}(Y_{1\to t})}\mathbb{P}(S_t|Y_{1\to t-1})\end{aligned}$$

17. Ce diagnostic en temps réel repose sur les données utilisées : si celles-ci révisent, alors le diagnostic est conditionnel à l'état des données. Les données de comptabilité nationale, par exemple, connaissent des révisions, d'amplitude variable suivant les pays et les variables examinées. En tout état de cause, il convient d'être prudent sur l'analyse des points les plus récents.

Soient alors $V_t = \mathbb{P}(Y_{1\to t})$ la vraisemblance de l'échantillon $(Y_1, \dots, Y_t)$ et $G_t \in \mathbb{R}^M$ le vecteur des densités conditionnelles $(\mathbb{P}(Y_t|S_t = s))_{s\in[\![1;M]\!]}$, la formule précédente s'écrit :

$$S_{t|t} = \frac{V_{t-1}}{V_t} G_t \odot S_{t|t-1}$$

Le dernier terme $S_{t|t-1}$ se calcule ainsi :

$$\begin{aligned}
\mathbb{P}(S_t|Y_{1\to t-1}) &= \sum_s \mathbb{P}(S_t|S_{t-1} = s, Y_{1\to t-1}).\mathbb{P}(S_{t-1} = s|Y_{1\to t-1}) \\
&= \sum_s \mathbb{P}(S_t|S_{t-1} = s).\mathbb{P}(S_{t-1} = s|Y_{1\to t-1}) \\
\text{soit } S_{t|t-1} &= P' S_{t-1|t-1}
\end{aligned}$$

Comme le vecteur $S_{t|t}$ somme à 1 (*ie* $\mathbb{1}'_M S_{t|t} = 1$), l'ignorance du quotient $\frac{V_{t-1}}{V_t}$ n'est pas problématique : il suffit de normaliser pour supprimer ce terme. La propriété d'hérédité s'écrit donc :

$$\boxed{S_{t|t} = \frac{G_t \odot (P' S_{t-1|t-1})}{\mathbb{1}'(G_t \odot (P' S_{t-1|t-1})}} \tag{1.9}$$

Pour initialiser, on peut utiliser la loi stationnaire $\pi$ :

$$S_{1|1} = \frac{\mathbb{P}(Y_1|S_1)\mathbb{P}(S_1)}{\mathbb{P}(Y_1)} \Longrightarrow S_{1|1} = \frac{G_1 \odot \pi}{\mathbb{1}'(G_1 \odot \pi)} \tag{1.10}$$

Les probabilités filtrées sont par construction des lois de probabilité sur $[0;1]^M$ :

$$\forall t, \quad S_{t|t} \in \Sigma_{M-1} = \left\{ p = (p_1, \dots, p_M) \in [0;1]^M / \mathbb{1}'_M p = \sum_{m=1}^{M} p_i = 1 \right\}$$

**1.2.6.1.2 Loi des probabilités filtrées** Les probabilités filtrées $S_{t|t} = \mathbb{P}(S_t|Y_{1\to t})$ sont des variables aléatoires. En tant que telles, il est légitime de s'interroger sur leur loi. Le processus est-il stationnaire ? Peut-on identifier cette loi, en fonction des paramètres sous-jacents ?

La complexité de $(1.9)$ explique que difficulté de la réponse à ces questions. Néanmoins, des simulations indiquent que :
- les lois seraient identiques. Question suivante : le processus est-il stationnaire ?
- la loi se concentrerait sur les sommets de l'hypercube $[0;1]^M$ : les probabilités filtrées seraient soit faibles soit fortes, mais relativement peu "moyennes".

### 1.2.6.2 Probabilités lissées

Pour estimer les probabilités lissées $S_{t|T} = \mathbb{P}(S_t|Y_{1\to T})$, Kim (1994) propose un algorithme récursif, reposant sur :
- les probabilitées lissées jointes consécutives $\mathbb{P}(S_t, S_{t+1}|Y_{1\to T})$,
- les probabilités filtrées précédemment calculées.

**1.2.6.2.1 Calcul des probabilités lissées** On commence par séparer les lois lissées jointes $(S_t, S_{t+1}|Y_{1\to T})$ en deux parties :

$$\mathbb{P}(S_t, S_{t+1}|Y_{1\to T}) = \mathbb{P}(S_t|S_{t+1}, Y_{1\to T}) \cdot \mathbb{P}(S_{t+1}|Y_{1\to T}) = \mathbb{P}(S_t|S_{t+1}, Y_{1\to T}) \cdot S_{t+1|T} \tag{1.11}$$

Connaissant $S_{t+1|T}$, et ayant explicité le premier terme, la probabilité lissée $S_{t|T}$ se déduira par sommation sur $S_{t+1}$.

Pour expliciter le premier terme de $(1.11)$, on élimine le futur du conditionnement :

$$\begin{aligned}
\mathbb{P}(S_t|S_{t+1}, Y_{1\to T}) &= \mathbb{P}(S_t|S_{t+1}, Y_{1\to t}, Y_{t+1\to T}) \\
&= \frac{\mathbb{P}(S_t, S_{t+1}, Y_{1\to t}, Y_{t+1\to T})}{\mathbb{P}(S_{t+1}, Y_{1\to t}, Y_{t+1\to T})} \\
&= \frac{\mathbb{P}(Y_{t+1\to T}|S_t, S_{t+1}, Y_{1\to t})}{\mathbb{P}(Y_{t+1\to T}|S_{t+1}, Y_{1\to T})} \cdot \frac{\mathbb{P}(S_t, S_{t+1}, Y_{1\to t})}{\mathbb{P}(S_{t+1}, Y_{1\to t})} \\
\Longrightarrow \mathbb{P}(S_t|S_{t+1}, Y_{1\to T}) &= \frac{\mathbb{P}(Y_{t+1\to T}|S_t, S_{t+1}, Y_{1\to t})}{\mathbb{P}(Y_{t+1\to T}|S_{t+1}, Y_{1\to T})} \cdot \mathbb{P}(S_t|S_{t+1}, Y_{1\to t})
\end{aligned}$$

Le premier terme de ce produit vaut 1, car conditionellement à $S_{t+1}$, $S_t$ et $Y_{1\to t}$ n'apportent aucune information sur $Y_{t+1\to T}$, de sorte que numérateur et dénominateurs sont égaux à $\mathbb{P}(Y_{t+1\to T}|S_{t+1})$. Nous avons donc montré que : $\mathbb{P}(S_t|S_{t+1}, Y_{1\to T}) = \mathbb{P}(S_t|S_{t+1}, Y_{1\to t})$. Il faut alors distinguer deux cas de figure :

- l'événement $\{S_{t+1} = m, Y_{1\to t}\}$ est de probabilité nulle, auquel cas $\mathbb{P}(S_t, S_{t+1} = m|Y_{1\to T}) = 0$,
- l'événement $\{S_{t+1} = m, Y_{1\to t}\}$ n'est pas de probabilité nulle.

Dans ce dernier cas, on peut continuer à développer :

$$\mathbb{P}(S_t|S_{t+1}, Y_{1\to t}) = \frac{\mathbb{P}(S_t, S_{t+1}, Y_{1\to t})}{\mathbb{P}(S_{t+1}, Y_{1\to t})} = \frac{\mathbb{P}(S_{t+1}|S_t, Y_{1\to t}) \cdot \mathbb{P}(S_t, Y_{1\to t})}{\mathbb{P}(S_{t+1}, Y_{1\to t})}$$

Or : $\mathbb{P}(S_{t+1}|S_t, Y_{1\to t}) = \mathbb{P}(S_{t+1}|S_t)$, donc :

$$\mathbb{P}(S_t|S_{t+1}, Y_{1\to t}) = \frac{\mathbb{P}(S_{t+1}|S_t).\mathbb{P}(S_t|Y_{1\to t})}{\mathbb{P}(S_{t+1}|Y_{1\to t})} = \frac{P_{S_t S_{t+1}} S_{t|t}}{S_{t+1|t}}$$

où $S_{t+1|t} = P' S_{t|t}$. On en déduit les lois lissées jointes [18] :

$$\boxed{\mathbb{P}(S_t, S_{t+1}|Y_{1\to T}) = \frac{\mathbb{P}(S_{t+1}|S_t).\mathbb{P}(S_t|Y_{1\to t})}{\mathbb{P}(S_{t+1}|Y_{1\to t})}.\mathbb{P}(S_{t+1}|Y_{1\to T}) = \frac{P_{S_t S_{t+1}} \cdot S_{t|t}}{S_{t+1|t}}.S_{t+1|T}} \tag{1.12}$$

(1.12) s'écrit matriciellement :

$$\mathbb{P}(S_t, S_{t+1}|Y_{1\to T}) = P \odot (S_{t|t} \mathbb{1}'_M) \odot (\mathbb{1}_M S'_{t+1|T}) \oslash (\mathbb{1}_M S'_{t+1|t})$$

où $S_{t+1|t} = P' S_{t|t}$. On en déduit l'expression des probabilités lissées d'ordre 1 :

$$\boxed{S_{t|T} = S_{t|t} \odot \left[P(S_{t+1|T} \oslash S_{t+1|t})\right]} \tag{1.13}$$

Ce résultat est compatible avec l'examen du cas $\mathbb{P}(S_{t+1} = m, Y_{1\to t}) = 0$, puisque $S_{t+1|t}(m) = \mathbb{P}(S_{t+1} = m|Y_{1\to t}) = 0 \Longrightarrow S_{t+1|T}(m) = \mathbb{P}(S_{t+1} = m|Y_{1\to T}) = 0$ de sorte qu'il suffit de considérer que le terme $(S_{t+1|T} \oslash S_{t+1|t})(m) = \dfrac{\mathbb{P}(S_{t+1} = m|Y_{1\to T})}{\mathbb{P}(S_{t+1} = m|Y_{1\to t})}$ vaut 0.

**1.2.6.2.2 En pratique** Les probabilités lissées sont encore plus polaires que les probabilitées filtrées (cf par exemple la figure 1.5), c'est-à-dire que :

- si $S_{t|t}$ est faible (proche de 0), alors $S_{t|T} \leq S_{t|t}$
- si $S_{t|t}$ est fort (proche de 1), alors $S_{t|T} \geq S_{t|t}$

## 1.2.7 Fitted values, résidus et prévisions

### 1.2.7.1 Fitted values et résidus

Statistiquement, avec un critère usuel de minimisation de la variance [19], les fitted values et résidus s'écrivent :

$$\begin{cases} \hat{Y}_t &= \hat{\mathbb{E}}\left(\mu_{S_t} + \sigma_{S_t}\varepsilon_t|Y_{1\to T}\right) &= \displaystyle\sum_{m=1}^{M} \hat{\mathbb{P}}(S_t = m|Y_{1\to T})\hat{\mu}_m \\ \hat{\varepsilon}_t &= \hat{\mathbb{E}}\left(\dfrac{Y_t - \mu_{S_t}}{\sigma_{S_t}}|Y_{1\to T}\right) &= \displaystyle\sum_{m=1}^{M} \hat{\mathbb{P}}(S_t = m|Y_{1\to T})\frac{Y_t - \mu_m}{\sigma_m} \end{cases}$$

Néanmoins, il peut être intéressant de travailler non pas avec les probabilités lissées (estimées) $\hat{\mathbb{P}}(S_t|Y_{1\to T})$ mais directement avec une inférence sur les états $\hat{S}_t$ :

$$\hat{S}_t = \underset{m\in[\![1;M]\!]}{\operatorname{argmax}}\, \hat{\mathbb{P}}(S_t = m|Y_{1\to T})$$

Utiliser $\hat{\mathbb{P}}(S_t|Y_{1\to T})$ ou $\hat{S}_t$ n'a qu'un impact relatif, dans la mesure où les probabilités lissées sont fréquemment proches d'un sommet du simplexe.

18. Écrivons cette expression sous forme matricielle, avec les indices, pour lever toute ambiguité :

$$\mathbb{P}(S_t = i, S_{t+1} = j|Y_{1\to T}) = \frac{p_{ij}\mathbb{P}(S_t = i|Y_{1\to t})\mathbb{P}(S_{t+1} = j|Y_{1\to T})}{\mathbb{P}(S_{t+1} = j|Y_{1\to t})}$$

19. Au sens où $\mathbb{E}X = \operatorname{argmin}_{m\in\mathbb{R}} \mathbb{E}(X - m)^2$.

#### 1.2.7.2 Prévisions

Soit $\hat{Y}_t(k) = \mathbb{E}(Y_{t+k}|Y_{1\to t})$ la prévision de $Y_{t+k}$ connaissant $Y_{1\to t}$ :

$$\begin{aligned}\hat{Y}_t(k) &= \mathbb{E}(Y_{t+k}|Y_{1\to t}) \\ &= \sum_s \left[\mu_s + \sigma_s \mathbb{E}(\varepsilon_{t+k} \mathbb{1}_{S_{t+k}=s}|Y_{1\to t})\right]\end{aligned}$$

Or, conditionnellement à $Y_{1\to t}$, les variables aléatoires $\varepsilon_{t+k}$ et $S_{t+k}$ sont indépendantes [20]. Comme de plus $\mathbb{E}(\varepsilon_{t+k}|Y_{1\to t}) = \mathbb{E}(\varepsilon_{t+k}) = 0$, on en déduit que :

$$\hat{Y}_t(k) = \mu' S_{t+k|t} = \mu'(P')^k S_{t|t} = (P^k\mu)' S_{t|t}$$

Apprécier cette prévision nécessite de connaître sa loi. Intuitivement, cette loi présente plusieurs modes. Supposons en effet qu'à la date $t$ le processus soit dans l'état $m$, avec la moyenne $\mu_m$. Un des deux événements suivants s'est réalisé en $t+1$ :

- le processus d'état est resté dans l'état $m$, ce qui se produit avec probabilité $p_{mm}$, il est alors encore proche de sa moyenne $\mu_m$.
- le processus d'état a sauté vers l'état $i \neq m$, ce qui se produit avec probabilité $p_{mi}$, $Y_{t+1}$ est alors proche de la moyenne $\mu_i$

L'intervalle de confiance de prévision est donc composé de plusieurs intervalles : il n'est pas d'un seul tenant. **Je trouve ça assez déstabilisant : voir le graphique de droite du programme 11. cf 11fig5.**

De la même façon, la convergence de la prévision $\hat{Y}_t(k)$ vers $\mu'\pi$ quand l'horizon de prévision s'étend ($k \to \infty$) est très compréhensible : asymptotiquement, $S_{\infty|t} = S_\infty$.

**Erreur $\mathbb{V}\left(Y_{t+k} - \hat{Y}_t(k)\right)$, ou alors spécifier la loi de l'erreur...**

#### 1.2.7.3 Appréciations qualitatives

Les processus Markov Switching présentent de bonnes propriétés *in sample* mais de moins bonnes propriétés *out of sample* au sens où :

- les résidus sont faibles dans l'échantillon,
- mais les prévisions ne sont pas parfaites.

La cause de ce phénomène est identifiée : il s'agit de l'inférence sur la chaîne de Markov cachée. La bonne qualité de cette inférence dans l'échantillon explique l'ajustement correct sur le passé. Mais la moindre rupture dans un horizon proche n'est pas anticipée, et cause de fortes erreurs de prévision. Bessec et Bouabdallah (2005) détaillent les faiblesses de cette approche en prévision.

## 1.3 Estimation

Le paramètre $\theta$ à estimer est constitué :

- des paramètres du modèle AR, notés $\theta_Y$ : les moyennes $\mu = (\mu_m)_{m\in[\![1;M]\!]}$ et variance $\sigma^2 = (\sigma_m^2)_{m\in[\![1;M]\!]}$ forment un ensemble de $2M$ paramètres
- des paramètres de la chaîne de Markov, , notés $\theta_M$ : les probabilités de transition : $(p_{ij})_{(i,j)\in[\![1;M]\!]^2}$. Cet ensemble comporte $M^2$ paramètres liés par $M$ relations linéaires, de sorte qu'il nous faut estimer $M(M-1)$ probabilités. Eventuellement, s'ajoutent aux probabilités de transition la distribution initiale de la chaîne.

### 1.3.1 Identification

Sans contrainte particulière, l'estimation n'est pas dénuée d'ambiguité. En effet, intervertir les états et les paramètres correspondants produit le même modèle probabiliste. Autrement dit, le modèle n'est pas statistiquement identifiable :

$$\exists(\theta_0, \theta_1) \in \Theta^2,\ \theta_0 \neq \theta_1 \qquad / \qquad \mathbb{P}_{\theta_0} = \mathbb{P}_{\theta_1}$$

20. Plus généralement : $\forall s, \forall t, \tau > s$, $\varepsilon_t$ et $S_\tau$ sont indépendants, conditionnellement à $Y_{1\to s}$ :

$$\mathbb{P}(\varepsilon_t, S_\tau|Y_{1\to s}) = \mathbb{P}(\varepsilon_t|Y_{1\to s}).\mathbb{P}(S_\tau|Y_{1\to s})$$

Concrètement : soit $E = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ la matrice de permutation des deux états (on se place dans le cas $M = 2$), alors $(P, \mu, \sigma^2)$ et $(EPE, E\mu, \sigma^2)$ imposent la même structure probabiliste.

Pour résoudre cette identification, il suffit d'imposer des contraintes identifiantes. Plusieurs sont envisageables :
- imposer que l'état 1 ait la moyenne la plus basse : $\mu_1 < \mu_2$,
- imposer que la diagonale de la matrice de transition $P$ soit croissante : $p_{11} < p_{22}$.

Chacune de ces contrainte suffit à lever l'ambiguité sur les états. La seconde présente l'avantage de la généralité : elle ne dépend pas spécifiquement de la modélisation non-markovienne. Toutefois, aucune de ces deux contraintes ne présente d'avantage technique sur l'autre.

Notons que la question de l'identifiabilité n'est que trop rarement dans les papiers théoriques ou dans les implémentations informatiques de cette techniques.

Une fois les paramètres judicieusement contraints, on peut procéder à l'estimation

### 1.3.2 Estimation par maximum de vraisemblance

#### 1.3.2.1 Calcul de la log-vraisemblance

Soit $f(\cdot; \mu, \sigma^2)$ la densité de $\mu + \sigma\varepsilon$ ; si $\varepsilon$ est gaussien alors $f(\cdot; \mu, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{\cdot - \mu}{\sigma}\right)$, où $\phi$ est la densité d'une gausienne centrée réduite. La vraisemblance s'écrit :

$$\begin{aligned} \mathcal{V}_\theta(Y_{1\to T}) &= \sum_{s_1,\dots s_T} \mathbb{P}(S_1 = s_1, \dots, S_T = s_t) \cdot \mathbb{P}(Y_{1\to T} | S_1 = s_1, \dots, S_T = s_t) \\ \implies \mathcal{V}_\theta(Y_{1\to T}) &= \sum_{s_1,\dots,s_T} \pi_{s_1} \prod_{t=2}^{T} p_{s_{t-1}s_t} \prod_{t=1}^{T} f\left(Y_t; \mu_{s_t}, \sigma^2_{s_t}\right) \end{aligned}$$

Cette expression est quasiment impossible à calculer : elle comporte trop de termes. Typiquement, le nombre de termes croît en $M^T$, ce qui est rédhibitoire. Toutefois, Ryden, Terasvirta et Asbryk (1998) introduit une simplification, amorcée par :

$$\mathcal{V}_\theta(Y_{1\to T}) = \sum_{s_T} \mathbb{P}_\theta(Y_T | S_T = s_t, Y_{1\to T-1}) \cdot \mathbb{P}_\theta(Y_{1\to T-1}, S_t = s_T)$$

Le premier terme de cette somme est connu, c'est $f(Y_t; \mu_{s_t}, \sigma^2_{s_t})$. Quant au second, on écrit :

$$\begin{aligned} \mathbb{P}(Y_{1\to t-1}, S_t = s_t) &= \sum_{s_{t-1}} \mathbb{P}(Y_{1\to t-1}, S_t = s_t, S_{t-1} = s_{t-1}) \\ &= \sum_{s_{t-1}} \mathbb{P}(S_t = s_t | S_{t-1} = s_{t-1}, Y_{1\to t-1}) \cdot \mathbb{P}(S_{t-1} = s_{t-1}, Y_{1\to t-1}) \\ \implies \mathbb{P}(Y_{1\to t-1}, S_t = s_t) &= \sum_{s_{t-1}} \mathbb{P}(S_t = s_t | S_{t-1} = s_{t-1}) \cdot \mathbb{P}(S_{t-1} = s_{t-1}, Y_{1\to t-1}) \end{aligned}$$

puisque $S_t$ est, conditionnellement à $S_{t-1}$, indépendant du passé de $Y$. Par récurrence, une expression plus aisée à calculer de la log-vraisemblance apparaît. Soit en effet $F(y; \mu, \sigma^2)$ la matrice diagonale dont le terme d'ordre $m$ est $f(y; \mu_m, \sigma^2_m)$ ($F(Y_t; \mu, \sigma^2)$ est donc la matrice diagonale formée sur le vecteur $G_t$ précédemment défini), alors :

$$\mathcal{V}_\theta(Y_{1\to T}) = \pi' P(Y_1; \mu, \sigma^2) \prod_{t=2}^{T} \left\{ PF(Y_t; \mu, \sigma^2) \right\} \mathbb{1}_M \tag{1.14}$$

Finalement, un algorithme numérique maximise cette vraisemblance.

#### 1.3.2.2 Existence de maxima locaux

La maximisation numérique de la log-vraisemblance bute sur l'existence d'extréma locaux ; il existe en effet au moins un maximum local non global correspondant aux états indistinguables. Considérons le

modèle simple $Y_t = \mu_{S_t} + \sigma \cdot \varepsilon_t$ où $S$ est une chaîne de Markov à 2 états ; les paramètres de ce modèle sont $\theta = (\mu_1, \mu_2, \sigma^2, p_{11}, p_{22}) \in \mathbb{R}^2 \times \mathbb{R}_+^* \times ]0;1[^2 = \Theta$ ; on s'intéresse à un sous-ensemble contraint de paramètres : $\theta^C = (\mu, \mu, \sigma^2, p_{11}, p_{22}) \in \Theta^C = \{(\mu, \mu, \sigma^2, p_{11}, p_{22}) / \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*, (p_{11}, p_{22}) \in [0;1]^2\} \subsetneq \Theta$. Soit $\hat{\theta}^C$ un ensemble d'estimateurs : $\hat{\theta}^C = \hat{\theta}^C(p_{11}, p_{22}) = (\hat{\mu}, \hat{\mu}, \hat{\sigma}^2, p_{11}, p_{22})$ où :

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Y_t \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{\mu})^2$$

Dans le cas où la chaîne $S_t$ est supposée stationnaire (*ie* $S_1$ est distribué selon $\pi$ où $P'\pi = \pi$), $\theta^C$ constitue un[21] maximum local (au sens où $(\nabla\mathscr{L})(\hat{\theta}^C) = 0$) mais pas global, car $\exists \theta^* \in \Theta$ / $\mathscr{L}(\theta^*) > \mathscr{L}(\hat{\theta}^C)$.

Montrons que $\hat{\theta}^C$ constitue un maximum local. Pour $\theta^C \in \Theta^C$, la loi de $Y_t$ ne dépend pas du régime, *ie* $\mathbb{P}_{\theta^C}(Y_t|S_t) = \mathbb{P}_{\theta^C}(Y_t)$ ou $G_t \propto \mathbb{1}_M$ (avec les notations de la partie 1.2.6.1), auquel cas la récurrence des probabilités filtrées (1.9) se simplifie en $S_{t|t} = P'S_{t-1|t-1}$, de sorte que $S_{1|1} = \pi \implies \forall t, S_{t|t} = \pi$, ce qui se transmet ensuite aux probabilités lissées[22] : $S_{t|T} = \pi$. Ces résultats, quoique non surprenants, sont forts : les probabilités filtrées et lissées sont dans le cas général des variables aléatoires, puisqu'elles dépendent de $Y_{1\to T}$ ; or on montre ici que, pour $\theta^C \in \Theta^C$, ces variables aléatoires sont dégénérées, c'est-à-dire réduites à des constantes :

$$\forall y_{1\to T} \in \mathbb{R}^T, \quad \mathbb{P}_{\theta^C}(S_t = m|Y_{1\to t} = y_{1\to t}) = \mathbb{P}_{\theta^C}(S_t = m|Y_{1\to T} = y_{1\to T}) = \pi_m \tag{1.15}$$

Finalement, pour tout $\theta^C \in \Theta^C$ :

$$\begin{aligned}
\mathscr{L}(Y_{1\to T}) &= \sum_{t=1}^{T} \log \sum_{m=1}^{M} f(Y_t; \mu, \sigma^2)\pi_m \\
&= \sum_{t=1}^{T} \log f(Y_t; \mu, \sigma^2) \\
\implies \mathscr{L}(Y_{1\to T}) &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(Y_t - \mu)^2}{\sigma^2}
\end{aligned}$$

et cette expression s'annule pour $\hat{\theta}^C$, de sorte que : $(\nabla\mathscr{L})(\hat{\theta}^C) = 0$. On a donc une infinité de maxima locaux, distincts du maximum global.

#### 1.3.2.3 Paramétrisation

Les algorithmes de maximisation numériques usuels optisent une fonction $f$ définie sur $\mathbb{R}^n$, or $\Theta$ n'a pas cette structure, en deux points :

- les variances sont positives,
- les probabilités de transition présentent deux problèmes :
  - elles sont comprises entre 0 et 1,
  - chaque ligne somme à l'unité.

Pour pallier cela[23], la méthode classique consiste à définir une bijection $h$ : $\begin{array}{ccc} X = \mathbb{R}^n & \to & \Theta \\ x & \mapsto & \theta \end{array}$. Voici un exemple de paramétrisation[24] :

- pour les variances : $\sigma_m^2 = e^x$ dans le cas univarié,
- pour une distribution de probabilités $(p_i)_{i\in[\![1;n]\!]}$ : $x = (\log\frac{p_i}{p_n})_{i\in[\![1;n-1]\!]}$ qui s'inverse en $p_i = \frac{1}{1+\sum_{j=1}^{n-1} e^{x_j}}(e^{x_1}, \ldots, e^{x_{n-}}$

  **reformuler plus clairement**

Avec cette paramétrisation, l'estimateur du maximum de vraisemblance s'écrit :

$$\hat{\theta} = h^{-1}\left(\underset{x\in X}{\operatorname{argmax}}\, \mathscr{L}_{h(x)}(Y_{1\to T})\right)$$

21. En réalité, une infinité, puisque $p_{11}$ et $p_{12}$ peuvent être choisis n'importe où dans $]0;1[$.
22. Par récurrence avec (1.13).
23. Et bien que la littérature ne fasse généralement pas état de cette étape, pourtant capitale.
24. Il en existe bien entendu une infinité.

#### 1.3.2.4 Problèmes numériques

Maximiser numériquement $(1.14)$ tel quel ne produit pas de résultat convaincant. Il faut en effet introduire une modification : il n'est pas rare que les termes de la matrice $F$ soient trop petits, étant donné la précision des logiciels. Pour pallier ce problème, on ne travaille pas sur $PF$ mais $cPF$ où $c$ est une constante qui dépend de la taille. Cette constante, de l'ordre de quelques unités pour des échantillons dont la taille varie entre 100 et 1000, est suffisante : ni trop petite sinon la log-vraisemblance vaut $-\infty$, ni trop grande sinon la log-vraisemblance vaut $+\infty$.

Autre obstacle : l'estimation dépend énormément des probabilités de transition initiales. Un procédé palliatif consiste à estimer le modèle en considérant successivement un grand nombre d'initialisations, en balayant une quadrillage de $\Sigma_M$ par exemple. Le problème est que cette grille est lourde : si l'on considère une grille de $\Sigma_M$ à $p$ points, alors la grille totale est énorme. Le recours au hasard permet de contourner, partiellement, ce problème : on décide de tirer les probabilités initiales au hasard. Les lignes de la matrice de transition initiales sont indépendantes, et chaque ligne est tirée selon une loi uniforme sur le simplexe $\Sigma_M$ (ce qui n'est rien d'autre qu'une loi de Dirichlet, voire annexe). Considérant un grand nombre d'initialisations, nous stockons les résultats, et choisissons celui qui conduit à la log-vraisemblance la plus élevée.

Ces solutions étant implémentées, nous pouvons maximiser la log-vraisemblance.

#### 1.3.2.5 Initialiation des paramètres markoviens

La méthode d'initialisation de la matrice de transition de la chaîne de Markov repose sur les différences de moyennes, elle est donc adaptée au cas de moyennes dépendantes du régime. L'adaptation à d'autres cas est aisée.

L'initialisation consiste à construire $M$ groupes homogènes en termes de variance, *ie* à partitionner l'échantillon $(X_1, \dots, X_n)$ en $\{I_1, \dots, I_M\}$ ($n_m = \#I_m$) de façon à minimiser

$$\sum_m \sum_i (X_i - \bar{X}_m)^2 \qquad \text{où} \qquad \bar{X}_m = \frac{1}{n_m} \sum_{i \in I_m} X_i$$

Etant donnée la décomposition des carrés, cette initialisation consiste, symétriquement, à maximiser la variance entre groupes :

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{m=1}^{M} \sum_{i \in I_m} (X_i - \bar{X}_i)^2 + \sum_m N_m (X_m - \bar{X})^2$$

Une estimation de la matrice de transition est déduite de cette première estimation des états ; la loi initiale est la loi stationnaire associée à cette matrice.

#### 1.3.2.6 Maximisation de la log-vraisemblance par l'algorithme EM

L'algorithme EM procède de façon itérative à la maximisation de la vraisemblance :
- la phase E, pour Expectation, dérive les probabilités filtrées et lissées des paramètres et de l'échantillon
- la phase M, pour Maximisation, dérive de l'étape précédente et de l'échantilon une nouvelle estimation des paramètres

Ces étapes sont répétées jusqu'à convergence, estimée selon les révisions des paramètres. En pratique, le nombre d'itérations à réaliser est relativement faible.

Cet algorithme est aisé à implémenter dans un cas particulier, dit "non-contraint", dans lequel la loi initiale n'est pas nécessairement la loi stationnaire du processus d'état. Dans ce cas précis, les équations normales (obtenues en annulant le gradient de la log-vraisemblance) se simplifient, et les nouveaux paramètres se déduisent simplement des précédents. Krolizg (1997) traite ce sujet de façon extensive.

**Paramètres markoviens** Dans le cas non-contraint, les équations normales permettent d'actualiser :
- les probabilités de transition [25] :

$$p_{ij} = \frac{S^{(2)}_{ij}}{\sum_j S^{(2)}_{ij}} \quad \text{soit} \quad P = S^{(2)} \oslash \left( S^{(2)} J_M \right)$$

25. Les détails sont renvoyés en section 2.1.3.3.2.1, page 29.

où $S^{(2)} = \sum_{t=1}^{T} S_t^{(2)}$ est la matrice $M \times M$ obtenue en sommant les probabilités lissées consécutives, $S_t^{(2)}$ est la matrice $M \times M$ de terme général $(S_t^{(2)})_{ij} = \mathbb{P}(S_{t-1} = i, S_t = j | Y_{1\to T})$ :

$$S_{ij}^{(2)} = \sum_{t=1}^{T} \mathbb{P}(S_{t-1} = i, S_t = j | Y_{1\to T}) \tag{1.16}$$

– la loi initiale $\pi_{S_1} = S_{1|T}$ exploite toute l'information disponible, contenue dans les probabilités lissées.

**Paramètres VAR** On montre **Hamilton (1990) je crois. Vérifier.** que la maximisation de la log-vraisemblance revient à maximiser l'expression suivante :

$$\sum_{t,m} \ln \mathbb{P}_\theta(Y_t | S_t = m) \mathbb{P}(S_t = m | Y_{1\to T})$$

En raison de l'hypothèse de normalité conditionnelle à l'état du processus, cette maximisation est aisée. Quelques lignes de calcul mènent en effet à la conclusion suivante :

$$\forall m, \quad \begin{cases} \hat{\mu}_m &= \sum_t \alpha_{mt} y_t \\ \hat{\sigma}_m^2 &= \sum_t \alpha_{mt} (y_t - \mu_m)^2 \end{cases} \quad \text{où} \quad \alpha_{mt} = \frac{\mathbb{P}(S_t = m | Y_{1\to T})}{\sum_s \mathbb{P}(S_s = m | Y_{1\to T})}$$

#### 1.3.2.7 Procédure complète d'estimation par ML

Il semble pertinent d'appliquer la procédure suivante :
– Obtenir une première initialisation des paramètres en construisant les groupes homogènes
– Améliorer cette estimation avec l'algorithme EM
– Incorporer cette estimation comme initialisation de la maximisation de la log-vraisemblance

#### 1.3.2.8 Propriétés de l'estimateur du maximum de vraisemblance

Pour évaluer la qualité de l'estimation par maximum de vraisemblance, on dispose de deux approches complémentaires :
– des théorèmes statistiques de convergence des estimateurs assurent que, lorsque la taille de l'échantillon augmente, l'estimateur converge vers la vraie valeur, et se rapproche d'une loi gaussienne ;
– en pratique, il convient toutefois d'évaluer pour quelles tailles d'échantillon l'ajustement est correct.

**1.3.2.8.1 Consistence** Leroux (1992) a mis en évidence un ensemble d'hypothèses, vérifiées aisément pour le modèle que nous examinons, sous lesquels l'estimateur du maximum de vraisemblance converge presque sûrement vers la vraie valeur : $\forall \theta, \mathbb{P}_\theta(\hat{\theta}_n \to \theta) = 1$. **Détailler rapidement**.

**1.3.2.8.2 Loi asymptotique** Bickel, Ritov et Ryden (1998) ont établi un jeu d'hypothèse, nécessairement plus restrictif que les conditions de Leroux (1992), sous lesquelles l'estimateur du maximum de vraisemblance vérifie une loi des grands nombres : $\sqrt{n}(\hat{\theta}_n - \theta) \Longrightarrow \mathcal{N}\left(0, I_1(\theta)^{-1}\right)$, où $I_1(\theta)$ est la matrice d'information de Fisher (pour un échantillon de taille 1) et $\Longrightarrow$ désigne la convergence en loi.

**1.3.2.8.3 Validation par simulation** **Détailler un exemple ; on peut aussi en profiter pour mettre en évidence la validité de l'approximation gaussienne qui découle de la section précédente.**

### 1.3.3 Méthodes bayésiennes

**Il est important de rappeler les avantages et inconvénients des méthodes bayésiennes, en général, et dans le cas particulier des MS.**

#### Rappels sur l'estimation bayésienne

Supposer une distribution pour $\theta$, appelée *loi a priori*, et en déduire la *loi a posteriori* : $\theta|Y$. L'estimateur final est souvent la moyenne de la loi a posteriori $\hat{\theta} = \mathbb{E}(\theta|Y)$.

Voir par exemple Kim et Nelson (1999).

## 1.4 Choix de modèles

**Dans quelle mesure faut-il distinguer les deux parties suivantes ? Peut-on réduire les tests remettant en cause la structure même du modèle (qui causent les deux problèmes majeurs (unindentified nuisance parameter under the null, and identically zero scores)) aux tests sur le nombre d'états ?**

### 1.4.1 Choix du nombre d'états

**Existe-t-il des tests ? Présenter rapidement une analyse coût-bénéfice.**

### 1.4.2 Tests

Il existe deux types de tests :
- les tests remettant en cause la structure même du modèle, tels que : le modèle à deux états peut-il se simplifier en un modèle linéaire ?
- les tests ne remettant pas en cause la structure du modèle, tels que $\sigma_1^2 = 1$.

Pour les modèles à changements de régimes markoviens, les tests du premier type présentent deux problèmes majeurs :
- Certains paramètres de nuisance ne sont pas identifiés sous l'hypothèse nulle. Supposons par exemple qu'on souhaite tester le DGP :
  - Hypothèse nulle $(H_0) : Y_t = \mu + \sigma \cdot \varepsilon_t$,
  - Hypothèse alternative $(H_1) : Y_t = \mu_{S_t} + \sigma \cdot \varepsilon_t$, où $S_t$ est une chaîne de Markov stationnaire (indépendante de $\varepsilon_t$) à deux états, de matrice de transition $P$.

  Sous l'hypothèse nulle, les probabilités de transition, qui sont des paramètres de nuisance, ne sont pas identifiées : soient $\theta = (\mu_1, \mu_2, \sigma^2, p_{11}, p_{22})$ les paramètres du modèle non contraint, alors $\mathscr{L}_{(\mu,\mu,\sigma^2,p_{11},p_{22})}(Y_{1\to T})$ ne dépend pas de $p_{11}$ et $p_{22}$.
- Le score est nul sous l'hypothèse nulle [26] : c'est moins immédiat, mais nous avons déjà montré (cf section 1.3.2.2, page 18) que le modèle contraint avec tous les états identiques constituait un maximum local mais non global.

**Continuer : présenter les bons tests, qui sont me semble-t-il dans Carrasco, Hu et Potterba. Dans quelle mesure est-il nécessaire de présenter les étapes successives ?**

## 1.5 Applications

**Idée : comparer les estimations sur données trimestrielles, et sur données mensuelles. Et, éventuellement, travailler comme Hamilton et Chauvet sur les estimations sur les premières versions du PIB.**

### 1.5.1 Données macroéconomiques

#### 1.5.1.1 Produit intérieur brut américain

Hamilton (1989) a appliqué en premier lieu la méthode des processus markoviens à changements de régime aux données de comptabilité nationale trimestrielles américaines. Il ne pensait bien entendu pas que le processus de génération des données (data generating process) était effectivement celui-ci. Toutefois, un tel modèle offre un cadre très pratique pour la détection des cycles. Comme nous allons le voir dans cette application, une modèlisation très simple, peu sophistiquée, permet de retrouver relativement précisément la datation du cycle américain fournie par le Business Cycle Dating Committee du National Bureau of Economic Research.

**1.5.1.1.1 Les données** Nous travaillons sur les données de comptabilité nationale produites par le Bureau of Economic analysis (l'institut national statistique américain). Précisément, nous nous intéressons à la série du PIB [27] publiée le 29 mars 2012. La série couvre 1947 à 2011T4.

Comme le BEA dans ses publications, nous transformons les données en niveau en évolutions annualisées [28]. Nous disposons donc de 259 points ; toutefois, comme les derniers points peuvent être soumis à

26. Soit $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^* \times ]0;1[ \times ]0;1[$ l'ensemble des valeurs possibles pour $\theta = (\mu_1, \mu_2, \sigma^2, p_{11}, p_{22})$ ; considérons le sous-ensemble $\Theta^c = \{(x, x) \,/\, x \in \mathbb{R}\} \times \mathbb{R}_+^* \times ]0;1[ \times ]0;1[$ : le score est nul sur tout $\Theta^c$.

27. Produit intérieur brut ; série chaînée, corrigée des variations saisonnières et des jours ouvrables.

28. Au lieu d'étudier $X_t$ ou $100\frac{X_t - X_{t-1}}{X_{t-1}}$, le BEA publie des évolutions annualisées : $100\left(\frac{X_t}{X_{t-1}}\right)^4 \approx 4 \cdot 100 \cdot \frac{X_t - X_{t-1}}{X_{t-1}}$.

révision, nous restreignons l'estimation à [1947T2 ;2010T4] (soit 255 points). Sur cette période, la moyenne est 3,3 % et l'écart-type 4,1 %.

**1.5.1.1.2 Les résultats** L'estimation d'un processus switching en moyenne et en variance produit les paramètres suivants :

$$\hat{\mu} = \begin{pmatrix} -0,8\,\% \\ 4,4\,\% \end{pmatrix} \quad \hat{\sigma} = \begin{pmatrix} 3,8\,\% \\ 3,5\,\% \end{pmatrix} \quad \hat{P} = \begin{pmatrix} 0,75 & 0,25 \\ 0,07 & 0,93 \end{pmatrix}$$

Deux états sont estimés :

- un état de contraction de l'économie, durant lequel le PIB diminue de 0,8 % (en moyenne annuelle) chaque trimestre,
- un état de croissance relativement élevée durant lequel le PIB américain augmente de 4,4 % (en moyenne annuelle) chaque trimestre.

Les variances des deux états sont quasiment identiques ; il n'est pas étonnant que ces variances soit inférieures à la variance agrégée, en raison de l'écart entre les moyennes dans les deux états. L'état de récession est très peu persistant : les récessions durent en moyenne 4 trimestres, selon la modélisation ; à l'inverse, les phases de croissance durent en moyenne 12,5 trimestres.

L'intervalle de confiance sur les estimations de moyenne est large : à 95 %, pour $\mu_1$ [-2,2 % ; 0,9 %], de sorte qu'il est préférable de qualifier cette croissance de faible ou nulle. L'intervalle de confiance sur $\mu_2$ est plus réduit : [3,9 % ; 5,1 %] : il s'agit bien d'un état de croissance forte. **Actualiser ces chiffres**

Toutefois, le résultat le plus intéressant est le suivant : en dépit de la simplicité de ce modèle, le diagnostic conjoncturel issu des probabilités lissées rejoint de près les cycles identifiés par le NBER, comme le résume le tableau 1.1 ; le graphique 1.5 représente les probabilités filtrées (trait bleu), lissées (trait rouge) et fonce les états lissés de récession. Pour le construire, un trimestre est déclaré en récession si sa probabilité lissée de récession (*ie* de l'état 1) excède 50 %. Les chiffres romains indiquent les trimestres.

| Cette datation | | | Datation NBER | | |
|---|---|---|---|---|---|
| Début | Fin | Trim | Début | Fin | Trim |
| 1948 III | 1949 IV | 6 | 1948 IV | 1949 IV | 5 |
| 1953 II | 1954 II | 5 | 1953 II | 1954 II | 5 |
| 1955 IV | 1958 II | 11 | 1957 II | 1958 II | 4 |
| 1960 II | 1961 I | 4 | 1960 II | 1961 I | 4 |
| 1969 II | 1970 IV | 7 | 1969 IV | 1970 IV | 5 |
| 1973 III | 1975 II | 8 | 1973 IV | 1975 I | 6 |
| 1979 I | 1980 III | 7 | 1980 I | 1980 III | 3 |
| 1981 II | 1982 IV | 7 | 1981 III | 1982 IV | 4 |
| 1989 IV | 1991 IV | 9 | 1990 III | 1991 I | 3 |
| 2000 II | 2003 I | 11 | 2000 I | 2001 IV | 8 |
| 2006 II | 2009 IV | 11 | | | |

TABLE 1.1 – Récessions du PIB

Toutes les récessions sont détectées, mais leur longueur est généralement sur-estimée.

Sur la figure 1.6, on a représenté les densités du modèle :

- en noir : la densité empirique des 239 évolutions trimestrielles (en rythme annuel), calculées avec une fenêtre de 1.5,
- en rouge : la gaussienne de l'état bas. En trait pointillé la gaussienne non pondéré, et en trait plein la gaussienne pondérée par la probabilité non conditionnelle,
- en bleu : comme en rouge, mais pour l'état bas,
- en violet : la somme de rouge et bleu, c'est-à-dire la densité du mélange de gaussienne.

Les moyennes sont indiquées par les traits pointillés verticaux : il est difficile de distinguer la moyenne empirique de la moyenne du mélange.

Un modèle à trois états fournit aussi des résultats intéressants :

$$\hat{\mu} = \begin{pmatrix} -2,8\,\% \\ 2,9\,\% \\ 5,4\,\% \end{pmatrix} \quad \hat{\sigma} = \begin{pmatrix} 3,0\,\% \\ 1,7\,\% \\ 4,5\,\% \end{pmatrix} \quad \hat{P} = \begin{pmatrix} 0,63 & 0,37 & 0 \\ 0,08 & 0,74 & 0,18 \\ 0 & 0,19 & 0,81 \end{pmatrix}$$

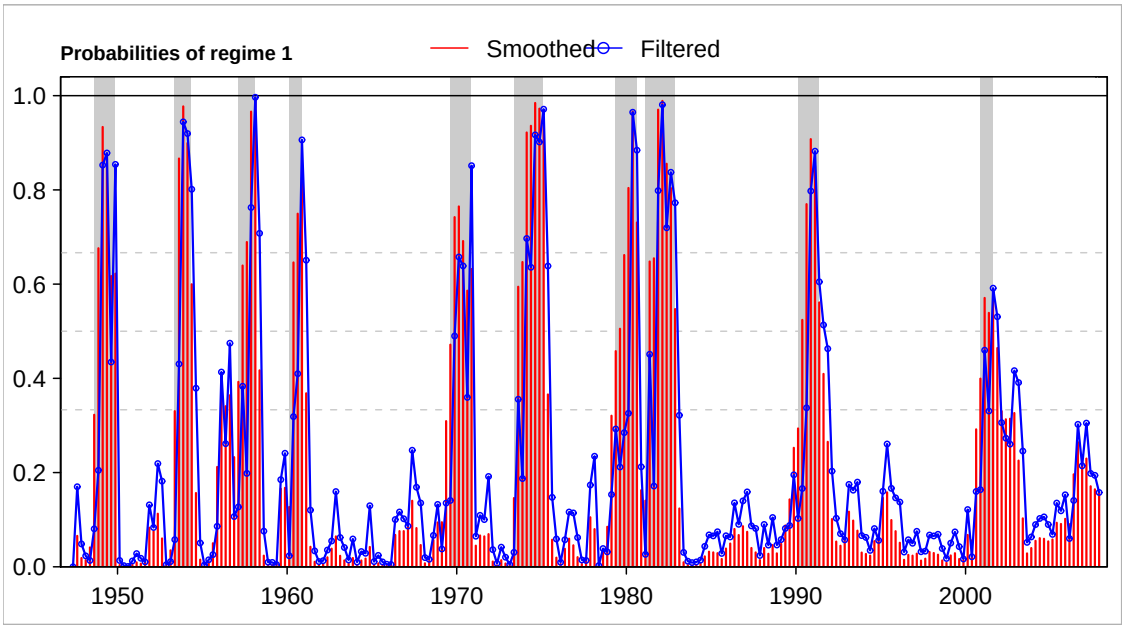


FIGURE 1.5 – Datation du PIB américain

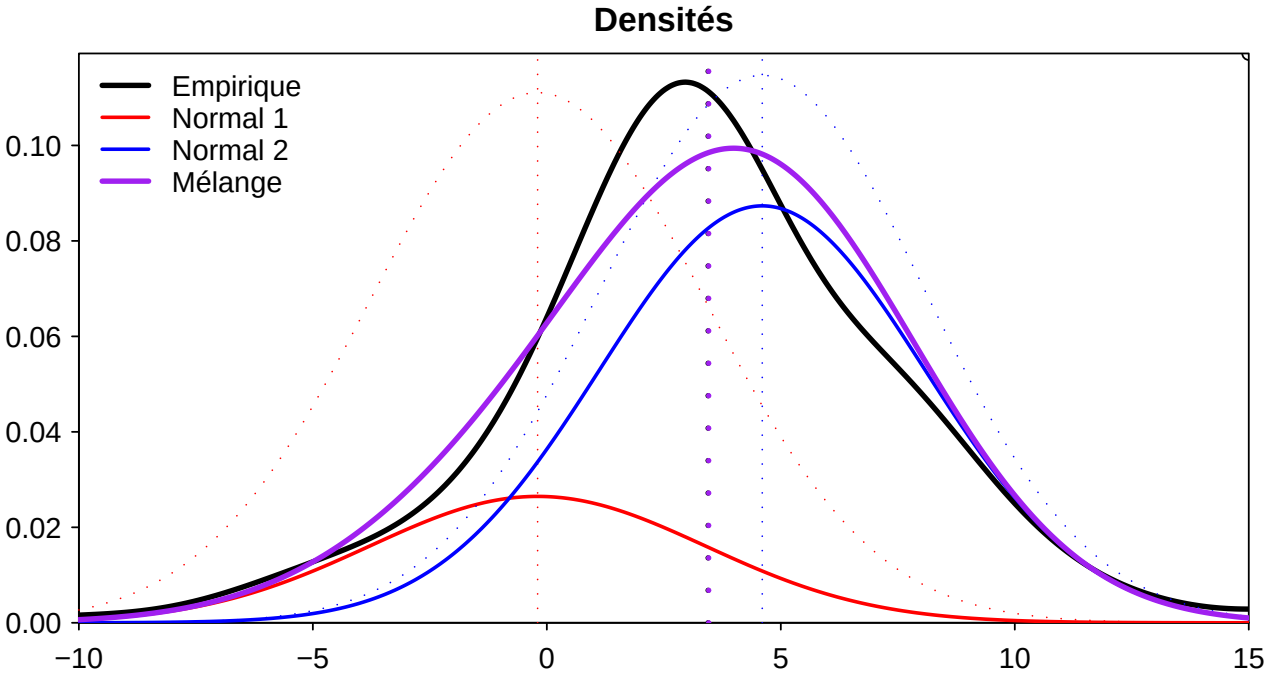


FIGURE 1.6 – Mélange de densités

Les intervalles de confiance sur la moyenne sont disjoints : $[-4,7\,\%; -0,9\,\%]$, $[2,3\,\%; 3,4\,\%]$ et $[3,7\,\%; 7,0\,\%]$. Les trois états sont aisément interprétables :

- Le premier état correspond à une contraction nette de l'économie, qui caractérise deux crises économiques : la récession due à la crise financière de 2007-2008 et le premier choc pétrolier ;
- le second correspond à une croissance moyenne, comme on en connaît depuis les chocs pétroliers des années 1970 ;
- le troisième état correspond à une croissance forte, comme l'économie américaine en connaissait durant les Trente Glorieuses.

Augmenter le modèle de 2 à 3 états améliore-t-il nettement l'ajustement ? La vraisemblance augmente nécessairement, mais le nombre de paramètres estimés aussi. Pour un modèle à $M$ états, il y a $M^2 - 2M + 1$ paramètres, de sorte qu'augmente de 2 à 3 double le nombre d'inconnues (de 7 à 14). Des critères de log-vraisemblance pénalisée peuvent être mobilisés pour éclairer cette interrogation. Selon le critère d'Akaike [29], l'amélioration de la log-vraisemblance justifie la complexification du modèle : on privilégierait plutôt la modélisation à trois états.

#### 1.5.1.2 Inflation

### 1.5.2 Données financières

Appliquons cette modélisation à la série des returns de l'index SP500.

29. Étants donnés deux modèles imbriqués, privilégier celui pour lequel AIC $= -2\log\mathscr{L} + 2k$ (où $k$ est le nombre de paramètres (libres)) est le petit.

# Chapitre 2

# Variations sur le thème

Ce chapitre généralise le modèle précédent, dans plusieurs directions :
- prise en compte d'une dynamique auto-régressive, dépendante du régime ou non. Deux extensions sont envisageables :
  - Hamilton (1989 ?) propose :

$$Y_t = \mu_{S_t} + \sum_{j=1}^{p} \Gamma_{j,S_t}(Y_t - \mu_{S_{t-j}}) + \varepsilon_t$$

    Krolzig (1997) qualifie ce modèle de Markov Switching in Mean (MSM) : le processus s'ajuste directment à un changement de régime.
  - Mais on peut aussi envisager le Markov Switching with Intercept (MSI) :

$$Y_t = \mu_{S_t} + \sum_{j=1}^{p} \Gamma_{j,S_t} Y_t + \varepsilon_t$$

    Ces deux processus ne partagent pas une dynamique commune.
- prise en compte d'exogènes, dépendantes du régime ou non. Dans le cas général, on peut écrire :

$$Y_t = A_{S_t} X_t + BZ_t + \varepsilon_t$$

  où l'influence des variables $X$ sur $Y$ dépend du régime tandis que l'impact de $Z$ sur $Y$ non.
- dynamique alternative du processus d'état $(S_t)_t$, qui n'est plus nécessairement markovien.

## 2.1 Le modèle MSI

Dans cette section, nous présentons une généralisation du modèle simple présenté au chapitre précédent. Si les calculs sont plus complexes, ce cas est fondamentalement proche deu précédent. Sous sa forme la plus générale, on s'intéresse au modèle suivant :

$$Y_t = \mu_{S_t} + \sum_{j=1}^{p} \Gamma_{j,S_t} Y_{t-j} + \mathcal{A}_{S_t} X_t + BZ_t + \Sigma_{S_t}^{1/2} \varepsilon_t \tag{2.1}$$

Les variables $X_t$ et $Z_t$ sont $x + z$ exogènes. C'est un modèle très général, au sens où :
- les variable sont multivariées : $Y_t \in \mathbb{R}^n$
- les matrices auto-régressives dépendent du processus d'état,
- les coefficients de certaines exogènes dépendent du processus d'état

### 2.1.1 Stationnarité

**Outre de la biblio, il faut décrire ici l'aspect original des markov switching : on peut avoir des $|\phi| > 1$. C'est Franck et Zakoïan, je crois. Vérifier aussi Timmerman.**

### 2.1.2 Intégration de termes MA

Dans le modèle $(2.1)$, nous avons intégré une dynamique auto-régressive, mais il est aussi possible de tenir compte de termes MA :

$$Y_t = X_t A_{S_t} + Z_t B + \varepsilon_t \quad \text{où} \quad \varepsilon_= \eta_t + \sum_{j=1}^{q} \Theta_{j,S_t} \eta_{t-j}$$

Dans ce cas, ce n'est qu'une façon particulière de modéliser la matrice de variance de $\varepsilon_t$.

Pour l'estimation ML par EM (cf infra), ça ne change pas grand chose :
- phase E : le processus inobservé est augmenté de $(\eta_t)_t$. Voici comment adapter la phase E :
  - Pour le calcul des probabilités filtrées, l'introduction des ordres MA ne fait qu'augmenter la variance.
  - ayabt estimé ensuite les probabilités lissées, on en déduit une estimation de $\varepsilon_t$, puis de $\eta_t$.
- phase M : adaptation à la marge assez immédiate, pour estimer en plus les $(\Theta_{js})_{js}$.

### 2.1.3 Estimation

Il est très utile de réduire le modèle avant de l'estimer.

#### 2.1.3.1 Réduction du modèle général

**2.1.3.1.1 Modèle général** L'écriture suivante englobe un grand nombre de modèles :

$$Y_t = \mu_{S_t} + \sum_{j=1}^{p} \Gamma_{j,S_t} Y_{t-j} + A_{S_t} X_t + B Z_t + \Sigma_{S_t}^{1/2} \varepsilon_t$$

**2.1.3.1.2 Réduction** Quitte à inclure la moyenne et les variables pré-déterminées (retards de la variable dépendante) dans les exogènes $X_t$ ou $Z_t$, le modèle précédent se réduit en :

$$Y_t = A_{S_t} X_t + B Z_t + \Sigma_{S_t}^{1/2} \varepsilon_t$$

où $X_t$ comporte $x$ variables et $Z_t$ $z$. Par conséquent, $\forall m$, $A_m$ est une matrice $n \times x$, et $B : n \times z$. Ce modèle comporte donc $Mnx + nz$ paramètres pour la moyenne et $Mn(n-1)/2$ paramètres pour la variance.

Détaillons le cas des retards. On peut écrire :

$$\sum_{j=1}^{p} \Gamma_{j,S_t} Y_{t-j} = A_{S_t} X_t$$

où $A : n \times (npM)$ est de la forme : $A_m = e_m' \otimes (\Gamma_{1m} \cdots \Gamma_{pm})$ et

$$X_t = \left( \mathbb{1}_M \otimes \begin{pmatrix} Y_{t-1} & \dots & Y_{t-p} \end{pmatrix}' \right)$$

**2.1.3.1.3 Factorisation** Il est utile de factoriser l'espérance conditionnelle aux régimes sous la forme :

$$\mathbb{E}(Y_t | S_t = \mu, Y_{1 \to t-1}) = U_{mt} \gamma \tag{2.2}$$

où $U_{mt} = (e_m' \otimes X_t' \, Z_t') \otimes \mathrm{Id}_n : n \times [n(Mx+z)]$ et $\gamma \in \mathbb{R}^{n(Mx+z)}$ sont composés de deux blocs, correspondant :
- aux endogènes dont l'impact sur la variable dépendante change avec le régime, $X_t$,
- aux autres endogènes $Z_t$

La matrice $U_{mt}$ est de grande dimension. Envisageons un cas presque standard :
- le nombre de régimes, $M$, s'élève à 2 ou 3,
- le nombre de retards, $p$, est limité à 4,
- le nombre de variables dépendante, $n$, est borné par 4 ou 5,
- le nombre d'exogènes, est limité : mettons (avant réduction) $x = 2$ et $z = 5$ (cas des dummies temporelles).

Après réduction, la matrice $X_t$ comprend les deux variables initiales, la constante, ainsi que $npM = 4 \cdot 4 \cdot 3 = 48$ variables pour les retards. Finalement, $x = 51$. La matrice $U_{mt}$ comporte donc $n\,[M(npM) + z] = 4[3 \cdot 48 + 10] = 616$ colonnes : néanmoins, un très grand nombre de cases sont vides.

#### 2.1.3.2 Estimation par maximum de vraisemblance

#### 2.1.3.3 Estimation par l'algorithme EM

Comme déjà vu en partie 1.3.2.6, l'algorithme EM enchaîne les phases E et M jusqu'à convergence.

**2.1.3.3.1 Phase E de l'algorithme EM** La phase E est identique à celle présentée en section 1.3.2.6 : il faut uniquement adapter le vecteur $G_t$ pour le calcul des probabilités filtrées, cf équation $(2.2)$.

**2.1.3.3.2 Phase M de l'algorithme EM** Étant donnée l'inférence sur le processus caché réalisée dans la phase E, qui se traduit par la connaissance des probabilités filtrées et lissées, la phase M consiste en l'estimation des paramètres par maximisation de la log-vraisemblance, ce qui ne pose pas de difficultés car nous disposons d'une expression analytique des estimateurs.

**2.1.3.3.2.1 Estimation des paramètres markoviens** Une fois dérivée la log-vraisemblance (par rapport à la matrice de transition et à la loi initiale), les conditions du premier ordre s'annulent aisément. Pour fixer les idées, prenons pour $\theta_M$ un élément de la matrice de transition : $\theta_M = p_{ij}$. On commence par décomposer la vraisemblance sur les états :

$$\mathscr{V} = \mathbb{P}(Y_{1\to T}) = \int \mathbb{P}(Y_{1\to T}|S_{1\to T})\mathbb{P}(S_{1\to T})dS_{1\to T}$$

Par conséquent, le score s'écrit ($\mathscr{L}$ est la log-vraisemblance, $\log \mathscr{V}$) :

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \theta_M} &= \int \frac{\mathbb{P}(Y_{1\to T}|S_{1\to T})}{\mathbb{P}(Y_{1\to T})} \frac{\partial \mathbb{P}(S_{1\to T})}{\partial \theta_M} dS_{1\to T} &(2.3)\\
&= \int \frac{\mathbb{P}(Y_{1\to T}|S_{1\to T})\mathbb{P}(S_{1\to T})}{\mathbb{P}(Y_{1\to T})} \frac{\partial \ln \mathbb{P}(S_{1\to T})}{\partial \theta_M} dS_{1\to T} &(2.4)\\
\Longrightarrow \frac{\partial \mathscr{L}}{\partial \theta_M} &= \int \mathbb{P}(S_{1\to T}|Y_{1\to T}) \frac{\partial \ln \mathbb{P}(S_{1\to T})}{\partial \theta_M} dS_{1\to T} &(2.5)
\end{aligned}$$

Or $\ln \mathbb{P}(S_{1\to T}) = \ln \mathbb{P}(S_1) + \sum_{t=2}^{T} \ln \mathbb{P}(S_t|S_{t-1})$, d'où :

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \theta_M} &= \int \mathbb{P}(S_{1\to T}|Y_{1\to T}) \frac{\partial \ln \mathbb{P}(S_1)}{\partial \theta_M} + \sum_{t=2}^{T} \int \mathbb{P}(S_{1\to T}|Y_{1\to T}) \frac{\partial \ln \mathbb{P}(S_t|S_{t-1})}{\partial \theta_M} dS_{1\to T}\\
\Longrightarrow \frac{\partial \mathscr{L}}{\partial \theta_M} &= \sum_s \mathbb{P}(S_1 = s|Y_{1\to T}) \frac{\partial \ln \mathbb{P}(S_1 = s)}{\partial \theta_M} + \sum_{t=2}^{T} \int \mathbb{P}(S_t, S_{t-1}|Y_{1\to T}) \frac{\partial \ln \mathbb{P}(S_t|S_{t-1})}{\partial \theta_M} dS_t, S_{t-1} \quad (2.6)
\end{aligned}$$

Le membre de droite de cette dernière équation sépare l'influence des deux paramètres markoviens que sont la loi initiale et les transitions entre états :

- le premier terme est relatif à la loi initiale : si la loi initiale n'est pas la loi stationnaire associée aux transition markoviennes (*ie* si $S_1$ n'est pas distribuée selon $\pi$ où $P'\pi = \pi$), alors la matrice de transition n'a aucun impact sur ce terme,
- le second terme est affecté uniquement par les transitions.

Et donc :

$$\frac{\partial \mathscr{L}}{\partial p_{ij}} = \sum_{t=2}^{T} \sum_{m_1,m_2} \mathbb{P}(S_{t-1} = m_1, S_t = m_2|Y_{1\to T}) \frac{\partial \ln p_{m_1 m_2}}{\partial p_{ij}} = \frac{1}{p_{ij}} \sum_{t=2}^{T} \mathbb{P}(S_{t-1} = i, S_t = j|Y_{1\to T})$$

Ainsi, la dérivée partielle fait intervenir les probabilités lissées consécutives d'ordre deux.

Le lagrangien s'écrit ensuite $\mathscr{L} - \lambda'(P\mathbb{1}_M - \mathbb{1}_M)$ ; des conditions du premier ordre dérivent : $p_{ij} \propto \sum_t \mathbb{P}(S_{t-1} = i, S_t = j|Y_{1\to T})$, d'où en normalisant :

$$p_{ij} = \frac{S_{ij}^{(2)}}{\sum_j S_{ij}^{(2)}} \quad \text{soit} \quad P = S^{(2)} \oslash \left(S^{(2)} J_M\right)$$

où $S^{(2)} = \sum_{t=1}^{T} S_t^{(2)}$ est la matrice $M \times M$ obtenue en sommant les probabilités lissées consécutives, $S_t^{(2)}$ est la matrice $M \times M$ de terme général $(S_t^{(2)})_{ij} = \mathbb{P}(S_{t-1} = i, S_t = j|Y_{1\to T})$.

**2.1.3.3.2.2 Estimation des paramètres VAR** Les paramètres VAR $\theta_Y$ sont déterminés en maximisant la log-vraisemblance du modèle réduit (cette écriture résulte directement de la transposition des écritures de la section 2.1.3.3.2.1 au cas des paramètres VAR) :

$$\frac{\partial \ln V}{\partial \theta_Y} = \sum_{m,t} \frac{\partial \ln \mathbb{P}(Y_t|S_t = m, Y_{1\to t-1})}{\partial \theta_Y} S_{t|T}(m) \tag{2.7}$$

où

$$\left\{ \begin{array}{rcl} \mathbb{E}(Y_t|S_t = m, Y_{1\to t-1}) & = & U_{mt}\gamma \\ \mathbb{V}(Y_t|S_t = m, Y_{1\to t-1}) & = & \Sigma_m \end{array} \right.$$

$U_{mt}$ est une matrice de dimension : $n \times n(Mx + z)$ :

$$U_{mt} = \begin{pmatrix} e_m' \otimes X_t' & Z_t' \end{pmatrix} \otimes \mathrm{Id}_n$$

Elle est donc essentiellement constituée de 0, puisque, sur $n^2(Mx+z)$ valeurs, seulement $n(x+z)$ sont non nulles. Correspondant à cette matrice, le paramètre $\gamma$ est composé des vecteurs $(A_m)_m$ et $B$ :

$$\gamma = \begin{pmatrix} A_1 \\ \vdots \\ A_M \\ B \end{pmatrix} \in \mathbb{R}^{n(Mx+z)}$$

Avec ces notations, $(2.7)$ s'écrit

$$\begin{aligned} & \sum_{m,t} \frac{\partial}{\partial \theta_Y} \left\{ -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_m| - \frac{1}{2}(Y_t - U_{mt}\gamma)'\Sigma_m^{-1}(Y_t - U_{mt}\gamma) \right\} S_{t|T}(m) \\ = \quad & c' - \frac{1}{2}\sum_m frac\partial\partial\theta_Y \left\{ \ln|\Sigma_m| \cdot T_m + \sum_t (Y_t - U_{mt}\gamma)' S_{t|T}(m) \Sigma_m^{-1} (Y_t - U_{mt}\gamma) \right\} \end{aligned}$$

où $T_m = \sum_t S_{t|T}(m)$. On vectorise alors la variable dépendante et les régresseurs :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \in \mathbb{R}^{nt} \qquad U_m = \begin{pmatrix} U_{m1} \\ \vdots \\ U_{mT} \end{pmatrix} : nT \times n(Mx+z)$$

Soit enfin $\Xi_m = \mathrm{diag}\left((S_{t|T}(m))_t\right) : T \times T$, on peut alors résoudre le modèle MCQG dans le cas non-contraint :

$$\boxed{\hat{\gamma}_{NC} = \left( \sum_m U_m'(\Xi_m \otimes \Sigma_m^{-1}) U_m \right)^{-1} \sum_m U_m'(\Xi_m \otimes \Sigma_m^{-1}) Y} \tag{2.8}$$

et l'estimation de la matrice de variance :

$$\boxed{\hat{\Sigma}_m = \frac{1}{T_m} \sum_t S_{t|T}(m)(Y_t - U_{mt}\hat{\gamma}_{NC})(Y_t - U_{mt}\hat{\gamma}_{NC})'} \tag{2.9}$$

En pratique, comme les calculs de $\hat{\gamma}$ et $\hat{\Sigma}_m$ dépendent l'un de l'autre, on procéde par itération ; par exemple :

- on initialise $\hat{\gamma}$ en supposant que la variance ne dépend pas de l'état : les $\sigma_m^2$ se simplifient alors dans $(2.8)$, et $\hat{\gamma}$ est simplement l'estimateur des moindres carrés (il suffit donc de prendre $\Sigma_m^2 = \mathrm{Id}_n$ dans $(2.8)$) ; on obtient donc $\hat{\gamma}^{(1)}$ ;
- on insère $\hat{\gamma}^{(1)}$ dans $(2.9)$ pour estimer $\Sigma_m^{(1)}$ ;
- on recalcule $\hat{\gamma}$ avec les nouvelles estmations de $\Sigma_m$, etc jusqu'à convergence (*ie* stabilisation de $\hat{\gamma}$ et $\hat{\Sigma}_m$), qui intervient en pratique très rapidement (moins de 5 itérations).

Dans le cas univarié ($n = 1$), les équations $(2.8)$ et $(2.9)$ s'écrivent :

$$\hat{\gamma} = \left( \sum_{mt} U_{mt}' U_{mt} \frac{S_{t|T}(m)}{\sigma_m^2} \right)^{-1} \sum_{mt} U_{mt}' Y_t \frac{S_{t|T}(m)}{\sigma_m^2} \qquad \text{et} \qquad \hat{\sigma}_m^2 = \frac{1}{T_m} \sum_t S_{t|T}(m)(Y_t - U_{mt}\hat{\gamma})^2$$

Des modélisations supplémentaires de la matrice de variance sont envisageables, il faut adapter l'estimation en conséquence.

Le modélisateur peut souhaiter introduire des contraintes particulières sur $\theta_Y$. On suppose ici que ces contraints s'expriment sous la forme $R\gamma = r$ (cette forme n'englobe pas toutes les contraintes possibles). L'estimation des variances est inchangée, mais il faut adapter l'estimation des paramètres de moyenne $\gamma$, à l'image de l'estimation sous contrainte dans les MCG.[1] L'estimateur contraint s'écrit :

$$\hat{\gamma}_C = \hat{\gamma}_{NC} + \left(\sum_m U_m'(\Xi_m \otimes \Sigma_m^{-1})U_m\right)^{-1} R' \left[R\left(\sum_m U_m'(\Xi_m \otimes \Sigma_m^{-1})U_m\right)^{-1} R'\right]^{-1} (r - R\hat{\gamma}_{NC}) \tag{2.11}$$

### 2.1.4 Exemples d'application

## 2.2 Le modèle MSM-VAR

Hamilton (1989) considère non pas une modélisation de type $(2.1)$ mais plutôt :

$$\Phi_{S_t}(L)\left[Y_t - (\mathcal{A}_{S_t}X_t + BZ_t)\right] = \varepsilon_t$$

Cette variation apparemment innocente est en réalité bien plus complexe. La difficulté principale réside dans la dépendance de $Y_t$ à $S_t$ certes, mais aussi et de façon directe à $S_{t-1}, \ldots, S_{t-p}$. Dans l'estimation, il faut donc remplacer la matrice de transition $P$ de taille $M$ par une matrice de transition augmentée $\tilde{P}$, de taille $M^{p+1}$.

## 2.3 Les modèles semi-markoviens

Jusqu'à présent, on a modélisé le processus inobservé $(S_t)_t$ d'une façon très sommaire, en supposant que c'était une chaîne de Markov homogène. Le temps de séjour dans un état suit alors nécessairement une loi géométrique. Toutefois, d'autres modélisations sont envisageables en théorie, et peuvent être plus adaptées aux phénomènes modélisés en pratique. Les modèles semi-markoviens procèdent de la façon suivante : ils modélisent le processus $(T_n, D_n)_{n\in\mathbb{N}^*}$ et en déduisent la loi de $(S_t)_t$.

1. Rappel : on cherche à minimiser $||Y - X\beta||_\Sigma^2$ sous la contrainte $R\beta = r$. Le lagrangien associé s'écrit :

$$\mathcal{L} = (Y - X\beta)'\Sigma^{-1}(Y - X\beta) - 2\lambda'(R\beta - r)$$

La dérivée partielle en $\beta$ est donc :

$$\frac{\partial \mathcal{L}}{\partial \beta} = 2X'\Sigma^{-1}X\beta - 2X'\Sigma^{-1}Y - 2R'\lambda$$

Les conditions du premier ordre fournissent donc : $\hat{\beta} = \hat{\beta}_{NC} + (X'\Sigma^{-1}X)^{-1}R'\lambda$ où $\hat{\beta}_{NC} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ est l'estimateur des MCG non contraint. Le multiplicateur $\lambda$ découle directement de la contrainte $R\beta = r$ :

$$\lambda = \left[R(X'\Sigma^{-1}X)^{-1}R'\right]^{-1}(r - R\hat{\beta}_{NC})$$

de sorte que :

$$\boxed{\hat{\beta}_C = \hat{\beta}_{NC} + (X'\Sigma^{-1}X)R'\left[R(X'\Sigma^{-1}X)^{-1}R'\right]^{-1}(r - R\hat{\beta}_{NC})} \tag{2.10}$$

# Chapitre 3

# Modèles markoviens qualitatifs

Dans les deux premiers chapitres, le processus modélisé était quantitatif : il s'agissait typiquement de variables macro-économiques, ou financières, continues. Grégoir et Lenglart (1998, 2000) dérivent d'enquête de conjoncture le signe de l'innovation : s'il est positif, alors l'activité se porte mieux qu'anticipé, en résumé. Dans ce cas, ils transforment les variables quantitatives initiales en variables binomiales. Pour utiliser un processus à changement de régime markovien, il faut adapter la théorie développée précédemment dans un cadre quantitatif.

La première partie sera consacrée à l'adaptation au cas qualitatif : outre la généralisation par Baron et Baron (2002) au cas polytomique du modèle binomial de Grégoir et Lenglart, nous montrerons une spécificité intéressant du cas discret. L'estimation, envisagée dans la seconde partie, ne présente pas de difficultés particulières. Nous appliquerons enfin ces travaux à l'enquête Industrie.

## 3.1 Modèle markovien qualitatif

### 3.1.1 Adaptation du cas quantitatif

Comme on modélise des variables binomiales $Y_t \in \{0,1\}^p$, l'idée consiste simplement à remplacer les densités conditionnelles par des probabilités conditionnelles :

$$Y_t|S_t = m \rightsquigarrow \mathcal{N}(\mu_m, \Sigma_m) \qquad \longrightarrow \qquad Y_t|S_t = m \rightsquigarrow \mathcal{B}(A_{m\cdot})$$

où $A_{m\cdot} \in [0;1]^p$. La dépendance entre variables, définie par les termes extra-diagonaux de la matrice de variance $\Sigma_m$ dans le cas quantitatif, est traitée de façon définitive **affiner le mot** : conditionnellement au processus d'état, on suppose les variables indépendantes. Par conséquent :

$$\mathbb{P}\left((Y_t^1, \ldots, Y_t^p) = (y_t^1, \ldots, y_t^p)|S_t = m\right) = \prod_{i=1}^{p} A_{mi}^{\mathbb{1}_{y_t^i=1}} (1 - A_{mi})^{\mathbb{1}_{y_t^i=0}} \tag{3.1}$$

On généralise aisément cette formule du cas binomial au cas multinomial.

Les paramètres de ces modèles sont donc distingués entre :
- les paramètres markoviens :
  - la matrice de transition de la chaîne de Markov,
  - les probabilités initiales, si besoin.
- les probabilités conditionnelles, regroupées dans la matrice $A : M \times p$, où $M$ est le nombre d'état de la CMH sous-jacente :

$$A_{ij} = \mathbb{P}(Y_t^j = 1|S_t = i)$$

Bien entendu, ce modèle est pertinent si les lignes de $A$ sont différentes, c'est-à-dire si les réponses des endogènes différent suivant l'état du régime sous-jacent.

### 3.1.2 Une intéressante spécificité de la modélisation qualitative

Outre la chaîne de Markov habituelle (notée $Z_t$), Grégoir et Lenglart (1998) introduisent un processus supplémentaire, renseignant sur le degré d'information de l'observation courante : le processus $W_t$ est aussi

une chaîne de Markov, à deux états. Le processus caché est donc $S_t = (Z_t, W_t)$, à $2M$ états, dont la matrice de transition est $P^S = P^Z \otimes P^W$, avec le classement suivant des états :

$$\begin{array}{l} S_t = 1 \Longleftrightarrow Z_t = 1 \text{ et } W_t = \text{non} \\ S_t = 2 \Longleftrightarrow Z_t = 1 \text{ et } W_t = \text{oui} \\ S_t = 3 \Longleftrightarrow Z_t = 2 \text{ et } W_t = \text{non} \\ S_t = 4 \Longleftrightarrow Z_t = 2 \text{ et } W_t = \text{oui} \end{array}$$

L'absence d'information, signalée par l'état $\{W = \text{non}\}$ se traduit concrètement par :

$$\forall m, \quad \mathbb{P}(Y_t = i | Z_t = m, W_t = \text{non}) \text{ ne dépend pas de } i$$

Or (dans le cas binomial) :

$$\mathbb{P}(Y_t = 0 | Z_t = m, W_t = \text{non}) + \mathbb{P}(Y_t = 1 | Z_t = m, W_t = \text{non}) = 1$$

de sorte que :

$$\mathbb{P}(Y_t = 0 | Z_t = m, W_t = \text{non}) = \mathbb{P}(Y_t = 1 | Z_t = m, W_t = \text{non}) = \frac{1}{2}$$

Certaines cases de la matrice $A$ sont donc imposées : ce ne sont pas des paramètres de l'algorithme.

Cette facilité de modélisation est propre au cas discret : elle n'a pas d'équivalent dans le cas continu.

## 3.2 Estimation

Outre le problème déjà abordé d'identifiabilité (cf section 1.3.1), le calcul de la vraisemblance et son optimisation ne pose pas de problèmes. Un algorithme EM est même disponible.

### 3.2.1 Calcul de la vraisemblance

La vraisemblance est un sous-produit du calcul des probabilités filtrées. Dans le filtre BLHK présenté en section 2.1.3.3.1, la seule adaptation nécessaire consiste à remplacer $G_t = (\mathbb{P}(Y_t | S_t = m))_m$ par $(3.1)$.

### 3.2.2 Algorithme EM

L'algorithme EM alterne entre les phases E et M jusqu'à convergence, qui intervient généralement assez rapidement.

#### 3.2.2.1 Phase Expectation

Il ne s'agit que du filtre BLHK : on construit l'inférence (filtrée et lissée) sur l'état du processus sous-jacent à l'aide des paramètres estimés $\hat{\theta}$ à l'étape précédente et des observations $Y$.

#### 3.2.2.2 Phase Maximization

**3.2.2.2.1 Estimation des paramètres markoviens** Comme dans le cas quantitatif (cf section 2.1.3.3.1 page 29).

**3.2.2.2.2 Estimation de la matrice $A$** Le calcul du score, cf $(2.7)$ dans le cas continu, s'adapte à l'identique :

$$\frac{\partial \mathscr{L}}{\partial A_{mi}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \frac{\partial \ln \mathbb{P}(Y_t | S_t = m)}{\partial A_{mi}} \mathbb{P}(S_t = m | Y_{1 \to T})$$

Or : $\ln \mathbb{P}(Y_t | S_t = m) = \sum_{i / Y_{it} = 1} \ln A_{mi} + \sum_{i / Y_{it} = -1} \ln(1 - A_{mi})$. D'où :

$$\frac{\partial \mathscr{L}}{\partial A_{mi}} = \sum_{t=1}^{T} \mathbb{P}(S_t = m | Y_{1 \to T}) \left[ \frac{\mathbb{1}_{Y_{it}=1}}{A_{mi}} - \frac{\mathbb{1}_{Y_{it}=-1}}{1 - A_{mi}} \right]$$

Soit alors $T_i = \{t \in [\![1; M]\!] / Y_{it} = 1\}$ et $\bar{T}_i = [\![1; M]\!] \backslash T_i$

$$\begin{cases} G^1_{mi} &= \sum_{t \in T_i} \mathbb{P}(S_t = m | Y_{1\to T}) \\ G^0_{mi} &= \sum_{t \in \bar{T}_i} \mathbb{P}(S_t = m | Y_{1\to T}) &= \sum_{t/Y_{it}=0} \mathbb{P}(S_t = m | Y_{1\to T}) \\ G_{mi} &= G^0_{mi} + G^1_{mi} &= \sum_{t=1}^{T} \mathbb{P}(S_t = m | Y_{1\to T}) \end{cases}$$

Avec ces notations :

$$\frac{\partial \ln V}{\partial A_{mi}} = \frac{G^1_{mi}}{A_{mi}} - \frac{G^0_{mi}}{1 - A_{mi}} \tag{3.2}$$

Des conditions du premier ordre on déduit sans peine :

$$\hat{A}_{mi} = \frac{G^1_{mi}}{G^0_{mi} + G^1_{mi}} = \frac{G^1_{mi}}{G_{mi}} \tag{3.3}$$

Finalement :

$$\boxed{\hat{A}_{mi} = \frac{\sum_{t/Y_{it}=1} \mathbb{P}(S_t = m | Y_{1\to T})}{\sum_{t=1}^{T} \mathbb{P}(S_t = m | Y_{1\to T})}} \tag{3.4}$$

Par conséquent, $\hat{A}_{mi}$ n'est autre que la version empirique de l'expression théorique de $A_{mi}$ (**??**), où la moyenne utilise les probabilités lissées comme pondérations. En effet :

$$\hat{A}_{mi} = \sum_{t=1}^{T} \alpha_{mt} \mathbb{1}_{Y^i_t = 1} \quad \text{où } \alpha_{mt} = \frac{\mathbb{P}(S_t = m | Y_{1\to T})}{\sum_{\tau=1}^{T} \mathbb{P}(S_\tau = m | Y_{1\to T})}$$

Par ailleurs, il est immédiat que : $\forall i, m, A_{mi} \in [0; 1]$.

L'adaptation de Hamilton (1996) à $(3.2)$ facilite le calcul du score, et donc de la variance (asymtotique) de $\hat{A}_{mi}$ :

$$\frac{\partial^2 \ln V}{\partial A^2_{mi}} = -\frac{G^1_{mi}}{A^2_{mi}} - \frac{G^0_{mi}}{(1 - A_{mi})^2}$$

de sorte que, après quelques lignes et en remplaçant $A_{mi}$ par son expression $(3.3)$ :

$$-\frac{\partial^2 \mathscr{L}}{\partial A^2_{mi}} = \frac{(G_{mi})^3}{G^0_{mi} G^1_{mi}}$$

Par conséquent, la variance de $\hat{A}_{mi}$ peut être approchée par :

$$\hat{\mathbb{V}}(\hat{A}_{mi}) = \frac{G^0_{mi} G^1_{mi}}{(G_{mi})^3} = \frac{\hat{A}_{mi}(1 - \hat{A}_{mi})}{G_{mi}} \tag{3.5}$$

Le point remarquable de $(3.5)$ est sa ressemblance avec la variance classique d'une proportion : dans le cas iid bernoullien, $\hat{p} = \bar{X}_n \Longrightarrow \mathbb{V}(\hat{p}) = \frac{p(1-p)}{n}$.

## 3.3 Applications

# Chapitre 4

# Switching dynamic factor models

**Rappel sur les modèles à facteurs**

Les modèles à facteurs sont utilisés en analyse de la conjoncture pour extraire une (ou plusieurs) composante(s) commune(s) à plusieurs séries temporelles. Voir Stock et Watson (1989). Soit $(Y_{it})_{i\in[\![1;n]\!],t\in[\![1;T]\!]}$ les $n$ séries temporelles, alors on cherche à décomposer :

$$Y_{it} = \sum_{j=1}^{q} \lambda_{ij}(L)F_{jt} + u_{it}$$

où $F_j = (F_{jt})_{t\in[\![1;T]\!]}$ est le facteur $j$. Pour des raisons d'identification, on se restreint fréquemment à l'extraction d'un seul facteurs (*ie* $q = 1$). Le terme $u_i$ est la composante spécifique de la série $i$. $\lambda_{ij}$ est un polynôme en $L$ (éventuellement restreint à la constante), qui détermine l'impact du facteur commun $j$ sur la série $i$.

L'estimation de ces modèles dépend de la structure probabiliste imposée, il existe trois méthodes :
- la technique naturelle : ACP,
- l'analyse factorielle statique consiste à maximiser la log-vraisemblance, en l'absence d'hypothèse sur la structure probabiliste des facteurs,
- l'analyse factorielle dynamique repose sur une modélisation des facteurs. Typiquement, Doz et Lenglart (19 ? ?) suppose que le facteur suit un processus ARMA

## 4.1 Une modélisation simplifiée

### 4.1.1 Présentation

Les modèles à facteurs ont été introduits pour résumer l'information à un ensemble de variables ; en tant que tel, on peut considérer que ces modèles procédent de techniques de réduction de la dimension. On suppose les processus $(Y_{it})_{it}$ reliés à un facteur caché, $(F_t)_t$, qui fait l'objet d'une modélisation MS propre. Concrètement, la modélisation est donc double :

**Équation d'observation**

$$Y_{it} = \lambda_i F_t + \varepsilon_{it} \tag{4.1}$$

**Équation d'état** Le facteur suit un modèle à changement de régime markovien, par exemple celui-ci :

$$F_t = \mu_{S_t} + \varepsilon_t \tag{4.2}$$

Ici, le facteur ne sert qu'à résumer un ensemble de variables, $Y_t = (Y_{1t}, \dots, Y_{nt}) \in \mathbb{R}^n$, à un facteur univarié $F_t \in \mathbb{R}$, et c'est ce facteur que l'on modélise sous forme à l'aide d'un modèle à changement de régime. L'écriture $(4.1)$ et $(4.2)$ est bien entendu simplifiée au maximum, de nombreux enrichissements sont possibles :
- dans l'équation d'observation : la loi du résidu $\varepsilon$ peut-être plus complexe qu'un simple bruit blanc,
- dans l'équation d'évolution du facteur sous-jacent : on peut introduire des retards de $F_t$, des exogènes, etc. comme dans n'importe quelle modélisation envisagée au chapitre 2.

### 4.1.2 Estimation

**Je ne suis pas certain de la pertinence de l'ordre des sous-parties.**

#### 4.1.2.1 Maximum de vraisemblance

**Écrire la log-vraisemblance** et la maximiser numériquement.

#### 4.1.2.2 Algorithme EM

**Voir si on peut en trouver un simplement.**

#### 4.1.2.3 Procédure en deux étapes

Étant donné que les régimes ne rétro-agissent pas sur le processus $Y_t$, on peut envisager une procédure d'estimation en deux étapes :

- estimation du facteur à partir de $Y$. Différentes méthodes sont envisageables, qui renvoient à différentes relations entre $Y$ et $F$ :
  - analyse en composantes principales : la dimension temporelle n'est pas exploitée,
  - analyse factorielle,
  - analyse dynamique : cf Stock et Watson ou Doz et Lenglart.
- estimation de la dynamique du facteur, à l'aide des techniques des chapitres précédents.

### 4.1.3 Application

**Il me semble qu'un papier de Benoît fait ça.**

Ces modéles constituent une première étape vers les vais modèles factoriels à changement de régimes markoviens, que nous étudions maintenant.

## 4.2 Modèles à facteurs à sauts markoviens

Ici, on suppose le facteur commun suit une dynamique markovienne. Typiquement, Nguiffo-Boyom (2006) étudie le modèle :

$$\left\{\begin{array}{ccc} Y_{it} & = & \Phi_i^1(L)F_t + \varepsilon_{it} \\ \Phi_i^2(L)\varepsilon_{it} & = & u_{it} \\ \Phi^3(L)F_t & = & \mu_{S_t} + \eta_t \end{array}\right.$$

où $(\forall i)$ $(u_{it})_t$ et $eta_t$ sont des bruits blancs non corrélés, et $\Phi$ sont des polynômes retards :

$$\left\{\begin{array}{rcl} \Phi_i^1(L) & = & \phi_{i0}^1 + \phi_{i1}^1 L + \cdots + \phi_{ip_i^1}^1 L^{p_i^1} \\ \Phi_i^2(L) & = & 1 + \phi_{i1}^2 L + \cdots + \phi_{ip_i^2}^2 L^{p_i^2} \\ \Phi^3(L) & = & 1 + \phi_1^3 L + \cdots + \phi_{p^3}^3 L^{p^3} \end{array}\right.$$

# Bibliographie


[AF02] Jacques Anas and Laurent Ferrara. Un indicateur d'entrée et sortie de récession : Application aux etats-unis. Technical Report 58, Centre d'Observation Economique, 2002.

[BC09] Muriel Barlet and Laure Crusson. Quel impact des variations du prix du pétrole sur la croissance française ? *Économie et Prévision*, 2, 2009.

[Ber92] Jan Beran. Statistical methods for data with long-range dependence. *Statistical Science*, 7 :404–416, 1992.

[Ber94] Jan Beran. *Statistics for Long-Memory Processes*. Chapman-Hall, 1994.

[BNP03] Robert Breunig, Serinah Najarian, and Adrian Pagan. Specification testing of markov switching models. *Oxford Bulletin of Economics and Statistics*, 65(s1) :703–725, December 2003.

[BSM03] B. Bellone and D. Saint-Martin. Detecting turning points with many predictors through hidden markov models. Technical report, ?, 2003.

[Cha98] Marcelle Chauvet. An econometric characterization of business cycle dynamics with factor structure and regime switching. *International Economic Review*, 39(4) :969–96, November 1998.

[DLW93] Francis X. Diebold, Joon-Haeng Lee, and Gretchen C. Weinbach. Regime switching with time-varying transition probabilities. Working Papers 93-12, Federal Reserve Bank of Philadelphia, 1993.

[DR10] Olivier Damette and Zorah Rabah. La datation du cycle français : une approche probabiliste. *Revue française d'économie*, XXIV :136–163, April 2010.

[EH89] Charles Engel and James D. Hamilton. Long swings in the exchange rate : Are they in the data and do markets know it ? NBER Working Papers 3165, National Bureau of Economic Research, November 1989.

[EMW07] Charles Engel, Nelson C. Mark, and Kenneth D. West. Exchange rate models are not as bad as you think. NBER Working Papers 13318, National Bureau of Economic Research, Inc, August 2007.

[Fer03] Laurent Ferrara. A three-regime real-time indicator for the us economy. *Economics Letters*, 81 :373–378, 2003.

[Fil98] Andrew J. Filardo. Choosing information variables for transition probabilities in a time-varying transition probability markov switching model. Technical report, 1998.

[FZ01] Christian Francq and J.-M. Zakoïan. Stationnarity of multivariate markov-switching arma models. *Journal of Economics*, 102 :329–364, 2001.

[Gar98] René Garcia. Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, 39(3) :763–88, August 1998.

[GH04] Granger and Huyng. Occasional structural breaks and long memory with an application to the s&p 500 absolute stock returns. *Journal of Empirical Finance*, 11 :399–421, 2004.

[GJ01] C. Gourieroux and J. Jasiak. Memory and infrequent breaks. *Economics Letters*, 70 :29–41, 2001.

[GL00] S. Grégoir and F. Lenglart. Measuring the probability of a business cycle turning point by using a multivariate qualitative hidden markov model. *Journal of Forecasting*, 19 :81–102, 2000.

[GR03] Dominique Guégan and Stéphanie Rioublanc. Study of regime switching models. do they provide long memory behavior ? an empirical approach. Technical Report 13, IDHE MORA, 2003.

[GT99] Granger and Teräsvirta. A simple nonlinear time series model with misleading linear properties. *Economics Letters*, 62 :161–165, 1999.

[Ham89] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 1989.

[Ham90] James D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, pages 39–70, 1990.

[Ham94] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[Ham96] James D. Hamilton. Specification testing in markov-switching time-series models. *Journal of Econometrics*, 70(1) :127–157, January 1996.

[Ham05] James D. Hamilton. What's real about the business cycle ? *Review*, (Jul) :435–452, 2005.

[Han92] Bruce E Hansen. The likelihood ratio test under nonstandard conditions : Testing the markov switching model of gnp. *Journal of Applied Econometrics*, 7(S) :S61–82, Suppl. De 1992.

[Kim94] C.-J. Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60 :1–22, 1994.

[Kla05] Franc Klaassen. Long swings in exchange rates : Are they really in the data ? *Journal of Business & Economic Statistics*, 23 :87–95, January 2005.

[KMP05] Chang-Jin Kim, James Morley, and Jeremy Piger. Nonlinearity and the permanent effects of recessions. *Journal of Applied Econometrics*, 20(2) :291–309, 2005.

[KN99] C.-J. Kim and C.R. Nelson. Has the u.s. economy become more stable ? a bayesian approach based on a markov-switching model of the business cycle. *Review of Economics and Statistics*, november 1999.

[Kro97] H.-M. Krolzig. *Markov Switching Vector Autoregressions*. Springer Verlag, 1997.

[MR83] Richard A. Meese and Kenneth Rogoff. Empirical exchange rate models of the seventies : Do they fit out of sample ? *Journal of International Economics*, 14(1-2) :3–24, February 1983.

[Per02] Corinne Perraudin. La prise en compte de ruptures dans l'évolution des variables économiques : les modèles à changements de régimes. Technical report, SAMOS-MATISSE et EUREQua, october 2002.

[R D04] R Development Core Team. *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[RL05] Ray and Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics*, 33 :2042–2065, 2005.

[RR97] Jennie E Raymond and Robert W Rich. Oil and the macroeconomy : A markov state-switching approach. *Journal of Money, Credit and Banking*, 29(2) :193–213, May 1997.

[RTA98] Rydén, Teräsvirta, and Asbrik. Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics*, 13 :217–244, 1998.

[Sic94] Daniel E Sichel. Inventories and the three phases of the business cycle. *Journal of Business & Economic Statistics*, 12(3) :269–77, July 1994.

[Wat05] Mark W. Watson. Commentary on "what's real about the business cycle ?". *Review*, (Jul) :453–458, 2005.

# Annexe A

# Liste de modèles

**Cette partie vous a un air de Prévert...**

## A.1 Modèle MSVAR

Il s'agit du modèle détaillé dans ce document

### A.1.1 Modèle simple

Il s'agit du modèle sans terme exogène ni retard, mais avec une moyenne et éventuellement une variance dépendant du régime :

$$Y_t = \mu_{S_t} + \Sigma_{S_t}^{1/2}\varepsilon_t$$

#### A.1.1.1 Hamilton (1988)

#### A.1.1.2 Chauvet et Hamilton (2005)

appliquent ce modèle à la série trimestrielle de PIB américain pour détecter les points de retournement conjoncturel.

### A.1.2 Modèle général

Il s'agit de

$$Y_t = \sum_{i=1}^{p} \Gamma_{j,S_t} Y_{t-j} + A_{S_t} X_t + BZ_t + \Sigma_{S_t}^{1/2}\varepsilon_t$$

où $X_t$ et $Z_t$ sont des variables exogènes.

Krolzig (2003) l'estime.

#### A.1.2.1 Kim et Nelson (1999)

Ils estiment le modèle $\Phi(L)(Y_t - \mu_{S_t}) = \varepsilon_t$ où $S_t = (S_t^1, S_t^2)$ est l'union de deux processus markoviens. Le second détermine la variance, et le premier la moyenne. Comme les auteurs cherchent à mettre en évidence une rupture dans la variance de $Y$, ils utilisent pour $S_t^2$ une modélisation particulière, sa matrice de transition est dégénérée, au sens où l'état 2 est absorbant[1]. Ils estiment leur modèle par une méthode bayésienne.

### A.1.3 Variation Moyenne-constante

Le modèle développé dans Hamilton (1989) est une version particulière de :

$$Y_t - \mu_{S_t} = \sum_{j=1}^{p} \Gamma_{j,S_t}(Y_{t-j} - \mu_{S_{t-j}}) + \Sigma_{S_t}^{1/2} \cdot \varepsilon_t$$

Ce dernier modèle généralise celui de Hamilton (1989) dans cinq directions :

1. La matrice de transition est de la forme $\begin{pmatrix} p_1 & q_1 \\ 0 & 1 \end{pmatrix}$

– il est multivarié,
– le nombre d'états $M$ n'est pas contraint à 2,
– le nombre de retards $p$ est libre,
– la transmission de $Y_{t-j} - \mu_{S_{t-j}}$ dépend du régime,
– et la variance du bruit dépend du régime.

Ce modèle dit *en moyenne* s'oppose au modèle dit *en constante* ; il présente une différence fondamentale : la moyenne dans l'état $j$ est $\mu_j$, et le processus s'ajuste rapidement aux changements de régimes.

Estimer de ce modèle n'est pas évident, au moins numériquement. Une bonne méthode consiste à transformer la chaîne de Markov $S_t$ en une chaîne $\bar{S}_t = (S_t, \ldots, S_{t-p}) \in [\![1; M]\!]^{p+1}$, dont la matrice de transition (de taille $M^{p+1}$) comporte un grand nombre de termes nuls : sur chacune des $M^{p+1}$ lignes, seuls $M$ termes (au plus) sont non nuls.

Krolzig (2001, 2003) applique cette modélisation à la datation d'un cycle des affaires pour la zone Euro (2 états, 1 retard, qq dummies). Néanmoins, il précise [2] qu'un modèle en constante est plus plausible (le terme, en anglais, est le sien) qu'un modèle en moyenne, car l'ajustement n'est pas immédiat.

### A.1.4 Krolzig (2003)

estime un tel modèle sur les indices de production industrielle (IPI), trimestriels, de plusieurs pays de la zone Euro. Modèle :

$$\Delta Y_t = \mu_{S_t} + A_1(\Delta Y_{t-1} - \mu_{S_{t-1}}) + \Sigma^{1/2} u_t$$

2 états. K a pas confiance car 1) limited data quality 2) manque de robustesse (typiquement : modif si on change les pays).

### A.1.5 Time-varying transition probabilities

Introduit simultanément par Filardo (1993) et Diebold, Lee et Weinbach (1994) **Selon Raymond et Rich (1997)**, ces modélisations autorisent les probabilités de transition à varier en fonction du temps et notamment à dépendre de variables exogènes. Voir aussi Filard (1998), ou Kim, Piger, Startz (2004).

Typiquement, dans un modèle à deux états, Diebold, Lee et Weinbach (1994) supposent que $p_{ii}(t) = \Phi(z_t'\beta_i)$, $\Phi$ est une fonction de répartition (logistique). Ils estiment le modèle $Y_t = \mu_{S_t} + \sigma_{S_t}\varepsilon_t$ sur données simulées, avec un algorithme EM adapté aux time-varying transition probabilities.

**Papiers avec applications ?**

### A.1.6 Modèle linéaire

Modèle du type $Y_t = X_t'\beta_{S_t} + \varepsilon_t$ où $\varepsilon_t$ est ARMA standard ?

## A.2 Modèles à facteurs

Voir le chapitre 4.

### A.2.1 Chauvet (1998)

estime le modèle (simple) :

$$Y_{it} = \gamma_i c_t + \varepsilon_{it}$$

où $\Phi(L)(c_t - \mu_{S_t}) = u_t$ est un BB, la chaîne $S_t$ comprend deux états. Les composantes spécifiques sont des AR : $\Theta(L)\varepsilon_{it} = \eta_{it}$.

### A.2.2 Chauvet et Piger (2005)

compare deux méthodes de datation du cycle américain :
– la méthode de Bry-Boschan
– un modèle à facteur à sauts markoviens

Ils travaillent sur quatre variables mensuelles :

2. Deux derniers paragraphes en page 5.

– emploi,
– IPI,
– manufacturing and trade sales,
– personal income (**=RDB ?**).

Néanmoins, l'apport principal de ce papier réside dans la base de données en temps réel (autrement dit, millésimée) utilisée pour la comparaison de ces règles. Ils concluent de leur étude que ces modèles obtiennent de bonnes performances, et qu'ils ont un grand avantage : le diagnostic arrive beaucoup plus rapide que le communiqué du NBER.

### A.2.3 Chauvet et Hamilon (2005)

La seconde partie fait un Switching DFM.

### A.2.4 Nguiffo-Boyom (2006)

applique cette méthodologie aux données françaises des enquêtes de conjoncture.

## A.3 Modèles qualitatifs

### A.3.1 Grégoir et Lenglart (1998,2000)

Première apparition des modèles qualitatifs. Plus de détails dans la section A.6.4.

### A.3.2 Baron et Baron (2002)

D'un point de vue théorique, il s'agit essentiellement d'une légère adaptation de Grégoir et Lenglart (1998,2000) au cas multinomial. En pratique, Baron et Baron (2002) récusent la réduction initiale des séries quantitatives au signe de l'innovation par le biais d'une modélisation régressive, et lui préfère le signe du glissement sur deux mois.

### A.3.3 Modèle DEREC

Modèle développé par l'Insee, et employé pour construire des indicateurs de retournement à partir des enquêtes de conjoncture. **Quelles améliorations apporte DEREC par rapport à GL et BB ?**

### A.3.4 Bouabdallah et Tselikas (2007)

[ Trésor-Éco n° 16 de 2007.] Application des modèlisations qualitatives : ils montrent qu'ajouter des variables financières enrichit peu l'inférence sur la datation. Bellone et Gauthier (2007 ?) concluaient à l'inverse, il me semble.

## A.4 De nombreux autres

Guerrero (2002) décompose le PIB US en un terme structurel (MS intégré) et un terme transitoire. Million et Guerrero sur courbe de Philips **vérifier et regarder.**

Bessec et Bouabdallah sur décomposition erreur de prévision.

## A.5 A ajouter

Raymond et Rich (1997) et Barlet et Crusson (2009) sur pétrole, par exemple.

## A.6 Résumés de quelques papiers

### A.6.1 Hamilton (1989, 1990)

Ces deux papiers sont clairement liés : Hamilton (1989) présente[3] la modélisation et ses implications macro-économiques tandis que Hamilton (1990) focalise sur les aspects statistiques (estimation, tests).

Hamilton (1989) (équation (4.3) page 367) propose le modèle suivant :

$$Y_t = \alpha_{S_t} + \varepsilon_t \quad \text{où} \quad \varepsilon_t \rightsquigarrow \text{AR}(p) \quad : \quad \varepsilon_t = \sum_{j=1}^{p} \phi_j \varepsilon_{t-j} + \eta_t \tag{A.1}$$

avec les hypothèses habituelles de stationarité sur le polynôme $\Phi(L) = 1 - \phi_1 L - \cdots \phi_p L^p$. Hamilton (1990) le ré-écrit sous la forme :

$$Y_t = \alpha_{S_t} + \sum_{j=1}^{p} \phi_j (Y_{t-j} - \alpha_{S_{t-j}}) + \eta_t \tag{A.2}$$

Ces deux modélisations sont strictement équivalentes :

– $(A.2)$ se déduit de $(A.1)$ par $\varepsilon_{t-j} = Y_t - \alpha_{S_{t-j}}$ :

$$Y_t = \alpha_{S_t} + \varepsilon_t = \alpha_{S_t} + \sum_{j=1}^{p} \phi_j \varepsilon_{t-j} + \eta_t = \alpha_{S_t} + \sum_{j=1}^{p} \phi_j (Y_{t-j} - \alpha_{S_{t-j}}) + \eta_t$$

– Réciproquement, soit $\varepsilon_t = Y_t - \alpha_{S_t}$, alors $(A.2)$ implique que $\varepsilon_t$ est un AR(p).

Hamilton (1989) estime le modèle sur le PNB[4] réel[5] américain (sur le taux d'évolution annualisé[6], mesuré par $4 \cdot 100 \Delta \log \text{PNB}$) sur la période 1951T2 à 1984T4. La chaîne cachée comporte deux états : expansion et récession. Comme attendu, l'état de récession est moins persistent, et tous les coefficients ont les bons signes.

Entre autres points intéressants, Hamilton (1990) implémente un algorithme EM pour ces modèles.

### A.6.2 Rabault (1993)

L'auteur reprend la modélisation de Hamilton, et l'applique à la croissance du PIB de six pays (USA, UK, France, Allemagne, Japon, Italie), sur la période 1960T1 à 1990T3 (estimation par maximum de vraisemblance, avec l'algorithme de Kitagawa de calcul des probabilités filtrées). Il s'intéresse à plusieurs paramètres influant sur la nature du cycle :

– la longueur des cycles ;
– l'asymétrie de la distribution est mesurée par $e = |p_{11} - p_{22}|$ : il ne parvient pas à rejeter l'hypothèse nulle, correspondant à $e = 0$, de symétrie des cycles ;
– la décomposition de la variance
– l'impact des chocs, en distinguant :
  – l'impact de la dynamique auto-régressive, mesurée par $\frac{1}{\Phi(1)}$ ;
  – l'impact de la dynamique markovienne, mesurée par $(\mu_2 - \mu_1)\frac{1-p_{11}-p_{22}}{2-p_{11}-p_{22}}$.

Il compare enfin la datation obtenue avec celles disponibles. **compléter**.

3. La lecture de cet article est vivement recommandée, en particuler les sections finales, d'une grande rigueur : Hamilton y montre que ce modèle est trés intéressant pour l'analyse. Il est (trop) rare que la justification d'une modélisation soit si détaillée.

4. Le Produit National Brut (PNB=GNP) diffère du Produit Intérieur Brut (PIB=GDP) : le PIB mesure la richesse créée sur le territoire national, tandis que le PNB mesure la richesse créée par les nationaux. Typiquement, l'activité d'une société française en Allemagne compte dans le PNB mais pas dans le PIB. Entre les deux figurent les salaires (précisément : le compte d'exploitation, majoritairement des salaires) et les revenus de la propriété entre pays. Ces concepts sont proches pour beaucoup de pays, mais peuvent différer. Par exemple, une forte partie de la valeur ajoutée en Irlande est réalisée par des sociétés étrangères : le PNB est donc plus faible que le PIB. Pour les USA, l'écart médian entre les taux d'évolution du PIB et du PNB est inférieur à 0.05 points de PIB trimestriel (non annualisé).
Ce choix du PNB plutôt que du PIB n'est pas motivé dans le papier. Hamilton construit sur ce modèle une datation du business cycle, comparée à la référence que constitue la datation du NBER, mais l'indicateur macroéconomique principal de cette dernière est le PIB, et non pas le PNB.

5. A l'époque, le concept de volume était du prix de l'année 1982 ; maintenant, il s'agit de prix chaînés.

6. *ie* multiplié par 4.

### A.6.3 Rydén, Teräsvirta, and Åsbrink (1998)

Les auteurs cherchent à montrer que les modèles à changement de régime sur la variance sont à même de reproduire certains faits stylisés de la distribution des returns[7] d'indices financiers. On avait en effet constaté à cette époque que la valeur absolue du returns semblait suivre partiellement[8] une loi exponentielle, au sens où la distribution empirique présentait les propriétés suivantes :

- moyenne=écart-type,
- skewness=2,
- kurtosis=9.

Á l'aide d'un modèle à deux états[9], RTA montre que certaines combinaisons de paramètres de Markov switching permettent de retrouver ces propriétés. Ces résultats sont synthétisés dans la figure A.1 :

- dans le panneau de gauche, on a représenté :
  - en bleu les paramètres de moyenne et de variance assurant que la moyenne égale l'écart-type,
  - en rouge les paramètres de moyenne et de variance assurant que le skewness s'approche de 2,
  - en bleu les paramètres de moyenne et de variance assurant que le kurtosis valle 9.

  Systématiquement, les courbes plus claires délimitent la zone $\pm$ 20 %. Idéalement, on voudrait satisfaire exactement les trois contraintes : comme ces trois courbes ne se rencontrent pas exactement, ce n'est pas possible. Par conséquent,
- on a représenté dans le panneau de droite les zones où la somme des écarts (relatifs) s'écarte de la situation optimale.

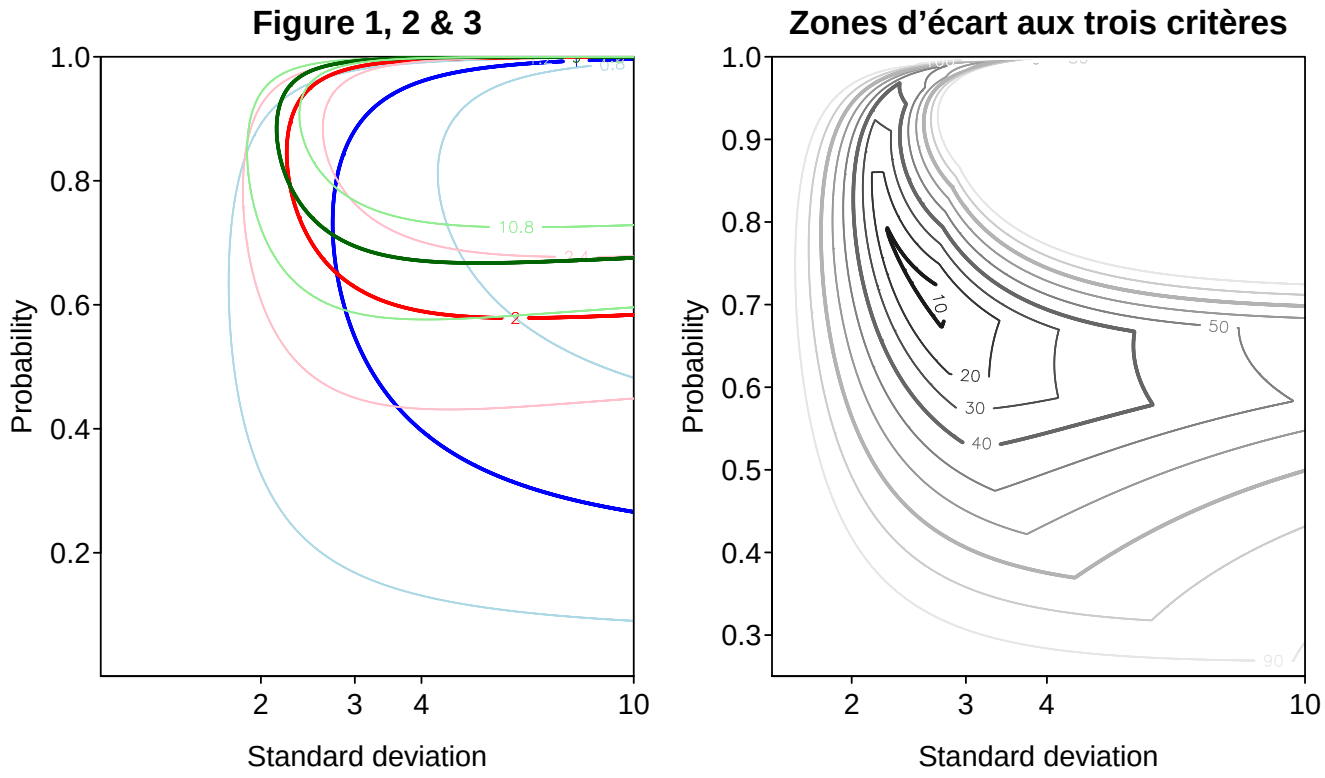


FIGURE A.1 – Paramètres approchant une loi exponentielle

Il existe donc une zone autour de $\pi = 0.7$ et $\sigma = 3$ où les trois propriétés sont approximativement satisfaites. Une fois cela constaté, il convient de vérifier que l'estimation du modèle produit des paramètres compatibles avec ce constat : les auteurs montrent qu'en corrigeant les outliers, les paramètres estimés garantissent bien cela.

Ils proposent quelques points méthodologiques intéressants :

7. Taux d'évolution, quotidiens par exemple.
8. Partiellement, car il suffit de représenter la densité empirique (même avec une fenêtre assez large) et la densité théorique pour constater un écart, qu'un test statistique d'adéquation confirme sans équivoque.
9. Précisément, ils étudient la densité : $\forall x \in \mathbb{R}, \quad f(x) = p \cdot f_1(x) + (1-p) \cdot f_2(x)$ où $f_1 = \mathcal{N}(0,1)$ et $f_2 = \mathcal{N}(0, \sigma^2)$. Soit $X$ de loi $f$et $H = |X|$, on calcule aisément la moyenne, variance, skewness et kurtosis de $H$ à l'aide des moments absolus (non centrés) suivants : $\mathbb{E}(H^n) = c_n \left[p + (1-p)\sigma^n\right]$, où $c_1 = \sqrt{2/\pi}$, $c_2 = 1$, $c_3 = 4(2\pi)^{-1/2}$, $c_4 = 3$.

– un test sur le nombre d'états du modèle,
– **d'autres ? voir**

### A.6.4 Grégoir et Lenglart (1998)

Grégoir et Lenglart sont les premiers à introduire les modèles markoviens qualitatifs, qu'ils motivent un argument patrique : ils cherchent à répliquer l'approche du conjoncturiste. Celui-ci aurait en tête une modélisation des séries d'intérêt, et modifierait son appréciation de la conjoncture au vu du signe de l'innovation [10] de la donnée récente.

La transposition du raisonnement quantitatif au cas binomial (anticipation positive ou négative) ne présente pas de difficultés. Toutefois, Grégoir et Lenglart, constatant l'échec de leur modèle, l'améliore en introduisant une nouvelle chaîne de Markov, renseignant sur la pertinence de chaque innovation en temps réel. Avec cette approche, les résultats sont bien plus probants, et ils dérivent une datation de la conjoncture française à l'aide de l'enquête Industrie.

### A.6.5 Krolzig (2003)

Krolzig (2003) estime plusieurs modèles sur le PIB de plusieurs pays de la zone Euro :
– sur le taux de croissance trimestriel PIB réel (données OCDE, de 1973T3 à 2002T1), il estime le modèle (section 5.3.1 du papier) :

$$\Delta Y_t = \mu_{S_t} + A_1 \Delta Y_{t-1} + \Delta Y_{t-4} + \Sigma_{S_t}^{1/2} u_t \tag{A.3}$$

avec 3 états et $\Sigma_m = \sigma_m^2 \mathrm{Id}$, où $Y_t$ est constitué des PIB des pays DE, FR, IT, ES, NL et AT. (Les lag 2 et 3 ne sont pas statistiquement non nuls). Un phénomène de rattrapage des pays méditerranées justifie l'introduction du troisième état.
– sur le taux de croissance trimestriel du PIB réel (données Eurostat, 1980T3 à 2002T1), corrigé par ses soins : ajustement saisonnier avec X12-ARIMA, puis détection et intervention sur les points aberrants avec le logiciel Gets. Le modèle estimé est proche de $(A.3)$ :

$$\Delta Y_t = \mu_{S_t} + A_1 \Delta Y_{t-1} + \Sigma_S^{1/2} u_t$$

Les pays sont DE, FR, IT, ES, NL, BE et FI.
– sur l'IPI (section 5.3.2 de son papier, résumé rapidement en section A.1.4 de ce document)

Il poursuit en testant l'existence d'un cycle commun à la zone euro, ce qui se ramène**selon lui, et je ne comprends pas comment il parvient à cette interprétation** au test $\mu_{i1} = \mu_{i2}$ pour tous les pays $i$. Si la croissance est la même (autrement dit si on peut rejeter l'existence de deux états), alors il est légitime de parler d'un cycle. Comment il répond par la positive **à l'aide d'un test du rapport de vraisemblance, a-t-il le droit ? est-ce loisible ici ?**, il conclut par une datation du cycle européen (cf table 9 ; la table 10 compare avec d'autres datations proposées dans la littérature). Globalement, voici les épisodes récessifs qui se dégagent :
– la première suit le premier choc pétrolier de 1973-74 : elle s'étend sur la période 1974T2(3 ?)-1975T1(3 ?),
– la deuxième est la récession Volcker (politique de désinflation par hausse des taux Fed) : 1980T1/1981T1-1982T4/1983T1,
– la troisième takes place au début des années 1990 : 1990T2/1992T2-1993T1/1993T3, elle résulte probablement des mesures fiscales et monétaires dues à la réunification allemande,
– enfin, il est probable qu'une récession ait débuté en 2001T2.

### A.6.6 Kim, Morley et Piger (2005)

Ce papier enrichit la modélisation initiale de Hamilton en intégrant explicitement la possibilité d'un rebond (*bounce-back*) du PIB après une récession. Précisément, les auteurs considèrent le modèle MSM enrichi suivant :

$$\Phi(L)\left(Y_t - \mu_{S_t} - \lambda \sum_{i=1}^{m} \mathbb{1}_{S_{t-i}=1}\right) = \varepsilon_t \tag{A.4}$$

où l'état 1 désigne la récession et l'état 2 la phase d'expansion : $\mu_1 < \mu_2$. Avec cette modélisation, si $S_{t-m} = \cdots = S_{t-1} = 1$ et $S_t = 2$, alors (en oubliant le polynôme auto-régressif $\Phi(L)$) $Y_t = \mu_2 + \lambda m + \varepsilon_t$,

10. En pratique, les anticipations sont identifiées à l'aide d'une modélisation auto-régressive.

$Y_{t+1} = \mu_2 + \lambda(m-1) + \varepsilon_{t+1}$, etc jusqu'à $Y_{t+m} = \mu_2 + \varepsilon_{t+2}$ : la croissance est donc majorée au sortir de la récession, cette majoration décroît linéairement sur $m$ périodes.

Kim, Morley et Piger estiment leur modèle[11] sur la log-croissance (non annualisée) du PIB américain sur la période 1947T1 à 2003T1. Ils retiennent $m = 6$, ce qui indique une reprise importante durant un an et demi), et aucun ordre auto-régressif ($\Phi(L) = 1$). Ils estiment la majoration $\lambda$ 0,3 point. Comme $\mu_2 = 0,8\,\%$, la croissance après la récession vaut 2, % le trimestre de sortie de la récession, puis 2,4 % le trimestre suivant, puis 2,1 %, puis 1,8 %, puis 1,5 %, puis 1,2 % ; le septième trimestre, la croissance a retrouvé son niveau normal de 0,8 %.

**compléter**

### A.6.7 Chauvet et Hamilton (2005)

Sans contenu particulièrement original[12], ce papier présente l'avantage d'une grande pédagogie. Il est structuré en deux parties :

- La première partie correspond fondamentalement aux travaux fondateurs de Hamilton : les auteurs commencent par introduire les mélanges de gaussiennes à l'aide d'une analyse empirique du PIB US basée sur la datation NBER ; le modèle à changement de régime markovien présenté au chapitre 1 (avec variance égales) en est la conclusion naturelle.
- La seconde partie semble reprise des travaux de Chauvet : une fois la modélisation simple introduite, il ne reste plus pour répliquer économétriquement l'approche du NBER de datation des cycles économiques qu'à introduire une analyse multivariée, en mêlant modèles à facteurs et modèles à changements de régimes markoviens. Typiquement, le modèle introduit est (cf eq (14), (15) et (16) en page 19) :

$$\left\{ \begin{array}{rcl} Y_{it} & = & \lambda_i F_t + u_{it} \text{ et } u_{it} = \theta_i u_{i,t-1} + \varepsilon_{it} \\ F_t & = & \mu_{S_t} + \phi F_{t-1} + \eta_t \end{array} \right. \tag{A.5}$$

  Les auteurs détaillent l'estimation du modèle, et les résultats. Ils concluent que leur modèle autorise une détection avec très peu de retard de l'économie américaine.

Ce papier présente toutefois deux nouveautés :

- La règle d'inférence sur l'état caché est plus simple qu'une dérivation rapide à partir des probabilités filtrées ou lissées.
- Une analyse en temps réel les autorise à conclure que les données du trimestre $t$ ne peuvent être utilisées pour diagnostiquer l'état de la conjoncture en $t$, mais seulement en $t-1$. Par conséquent, leur méthode ne fournit pas une datation en temps réel, mais avec un retard d'un trimestre[13]

### A.6.8 Hamilton (2005)

Hamilton (2005) présente un intérêt essentiellement pédagogique, en montrant l'intérêt des processus à changements de régime markovien pour l'analyse de la conjoncture. Le raisonnement procède en deux temps :

- il existe réellement des cycles économiques réels : Hamilton les constate sur l'emploi. Un modèle linéaire ne peut rendre compte de cette dynamique, et en particulier de l'asymétrie entre phases d'expansion et de récession : il est impératif d'utiliser un modèle non-linéaire, par exemple à changements de régimes markoviens.
- en outre, ces fluctuations réelles sont partiellement dues à des cycles dans la sphère monétaire.

#### A.6.8.1 Les cycles de l'économie réelle

En considérant le taux de chômage trimestriel américain**trouver quelques références biblio sur ce sujet spécifique**, Hamilton commence par montrer que celui-ci connaît des phases de hausse et de baisse, mais il note que la durée de ces cycles n'est pas fixée. Il rejette donc une approche fréquentielle, à la Baxter-King. Toutefois, il ne s'arrête pas là mais propose une définition alternative du cycle : un cycle correspond à un jeu particulier de paramètres de modèle linéaire, et on alterne entre les cycles (donc les DGP) avec un processus MS. Il montre alors qu'une telle modélisation à régimes produit de meilleurs résultats qu'une simple modélisation linéaire, sur deux plans :

11. Maximisation numérique de la log-vraisemblance
12. Me semble-t-il.
13. À comparer avec la datation du NBER qui ne parvient que plusieurs mois après les faits.

– en termes statistiques purement : l'ajustement est meilleur [14], comme en témoigne la log-vraisemblance. Il vérifie cela par en réalisant le test de Carrasco, Hu et Ploberger.
– mais, et c'est ce point qui intéresse l'économètre, en terme économiques : le modèle à régime réplique mieux les fluctuations que le modèle linéaire simple.

#### A.6.8.2 Les cycles monétaires

Ici, les régimes n'affectent pas la moyenne mais la variance des innovations. En confrontant les périodes de troubles monétaires, caractérisées par une variance plus élevée, Hamilton montre que plusieurs récessions (au sens du NBER) sont précédées de volatilité monétaire élevée. Néanmoins, ce n'est pas le cas de toutes.

#### A.6.8.3 Commentaire de Watson (2005)

Watson focalise sa réponse sur les aspects statistiques :
– il commencer par rappeler que les modèles à changements de régimes markoviens ne sont pas les seuls à introduire de la non-linéarité : Watson cite Slutsky (1937), qui avait déjà montré qu'un modèle purement linéaire pouvait induire des cycles. Toutefois, il me semble que l'argument de Watson n'a aucun rapport avec l'analyse d'Hamilton, puisque ce dernier rejette précisément une définition fréquentielle du cycle ! Non seulement Watson cite un argument qui ne répond pas exactement à ce qu'il avance, mais en plus son argument ne répond pas à Hamilton.
– Watson cherche ensuite à évaluer si les modèles non linéaires réalisent de meilleurs performances que les modéles linéaires. Il conclut que c'est probable, mais incertain. Toutefois, la littérature a depuis montré que les modèles MS posent un problème important pour la prévision : ils sont incapables de prévoir des changements de régime.

Comme Watson ne répond pas à ce qui constitue à mon sens le principal intérêt du papier de Hamilton, à savoir que les propriétés du cycle défini par les processus MS ouvre de nouvelles pistes, ou éclaire des points théoriques, il doit falloir comprendre qu'il l'accepte, auquel cas il n'a plus qu'à répondre sur la statistique. Or sa réponse présente un contenu statistique négligeable.

14. La modélisation par une loi de Student (au lieu d'une gaussienne) des erreurs accroît sensiblement la log-vraisemblance, ce qui est probablement dû à la souplesse des queues de distribution autorisée par la loi de Student. Hamilton estime en effet à 5.09 le nombre de degrés de liberté, ce qui implique un excess kurtosis supérieur à 6 (rappel : pour une loi de Student à $n$ degrés de liberté, l'excess kurtosis n'existe que si $n > 4$ et vaut $6/(n-4)$).

# Annexe B

# Rappels sur les chaînes de Markov

On commence par étudier rapidement un cas simple, puis on présente les concepts utiles à l'étude des processus à changements de régime markoviens. Il ne faut pas perdre de vue que les processus estimés auront tous de bonnes propriétés, sauf pathologie particulière et rare.

## B.1 Cas $M = 2$

La chaîne la plus simple comporte deux états, sa matrice de transition $P$ est :

$$P = \begin{pmatrix} p_1 & q_1 \\ q_2 & p_2 \end{pmatrix}$$

où l'on suppose que les deux états communiquent : $p_i \in ]0;1[$. $P$ est diagonalisable, car ses deux valeurs propres sont distinces, 1 et $\lambda = p_1 + p_2 - 1 = 1 - q_1 - q_2$. Les sous-espaces propres associés sont engendrés par $(1,1)$ et $(q_1, -q_2)$.

La loi marginale de $S_t$, notée $\pi^t$ vérifie $\pi^{t+1} = P'\pi^t$. La chaîne converge vers sa distribution stationnaire $\pi^\infty$ : $\pi^\infty = P'\pi^\infty$. Donc : $\pi^{t+1} - \pi^\infty = (P')(\pi^t - \pi^\infty)$ de sorte que :

$$\forall t \in \mathbb{N}, \pi^t = \pi^\infty + (P')^h(\pi^0 - \pi^\infty) \tag{B.1}$$

où l'on a autorisé la chaîne à prendre n'importe quelle distribution de probabilités initiale, $\pi^0$. Dans ce cas simple, on peut exprimer la loi limite $\pi^\infty = (\pi_1^\infty, \pi_2^\infty)$ :

$$\boxed{\pi_1^\infty = \frac{q_2}{q_1 + q_2} \quad \text{et donc} \quad \pi_2^\infty = 1 - \pi_1^\infty = \frac{q_1}{q_1 + q_2}}$$

Intuitivement, si l'état 1 est "fort" ($p_1$ proche de 1, donc $q_1$ petit) alors la chaîne de Markov y passe beaucoup de temps. On montre aisément que $(B.1)$ se ré-écrit :

$$\boxed{\forall t \in \mathbb{N}, \pi^t = \pi^\infty + \lambda^h(\pi^0 - \pi^\infty)} \tag{B.2}$$

Pour ce faire, on calcule $P^h$ en la diagonalisant :

$$P = \begin{pmatrix} 1 & q_1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \frac{1}{q_1 + q_2} \begin{pmatrix} q_2 & q_1 \\ 1 & -1 \end{pmatrix}$$

soit, en élevant à la puissance $h$ :

$$\begin{aligned} P^h &= \frac{1}{q_1 + q_2} \begin{pmatrix} q_2 + \lambda^h q_1 & q_1 - \lambda^h q_1 \\ q_2 - \lambda^h q_2 & q_1 + \lambda^h q_2 \end{pmatrix} \\ \Longrightarrow P^h &= \frac{1}{q_1 + q_2} \begin{pmatrix} q_2 & q_1 \\ q_2 & q_1 \end{pmatrix} + \frac{\lambda^h}{q_1 + q_2} \begin{pmatrix} q_1 & -q_1 \\ -q_2 & q_2 \end{pmatrix} \end{aligned}$$

où le premier terme est la matrice limite, $P^\infty = \mathbb{1}\pi'$. Puis [$\pi = \pi^\infty$] :

$$(P')^h(\pi^0 - \pi) = \left[\pi\mathbb{1}' + \frac{\lambda^h}{q_1 + q_2} \begin{pmatrix} q_1 & -q_2 \\ -q_1 & q_2 \end{pmatrix}\right](\pi^0 - \pi)$$

Or $\mathbb{1}'\pi^0 = \mathbb{1}'\pi = 0$ de sorte que le premier terme du développement s'annule. Puis :

$$\begin{aligned}
& \frac{1}{q_1+q_2}\begin{pmatrix} q_1 & -q_2 \\ -q_1 & q_2 \end{pmatrix}(\pi^0 - \pi) = \pi^0 - \pi \\
\iff & \begin{pmatrix} q_1 & -q_2 \\ -q_1 & q_2 \end{pmatrix}(\pi^0 - \pi) = (q_1+q_2)(\pi^0 - \pi) \\
\iff & \begin{pmatrix} -q_2 & -q_2 \\ -q_1 & -q_1 \end{pmatrix}(\pi^0 - \pi) = (\pi^0 - \pi) \\
\iff & -q\mathbb{1}'(\pi^0 - \pi) = 0
\end{aligned}$$

et le même argument que précédemment permet de conclure.

## B.2 Chaîne de Markov homogène

## B.3 Loi

## B.4 Irréductibilité

## B.5 Convergence

Sous certaines hypothèses bien choisies (chaîne apériodique), la suite de matrice $(P^n)_n$ converge vers $\pi\mathbb{1}'_M$, où le vecteur $\pi$ vérifie l'équation : $\pi = P'\pi$. Comme la solution de cette équation est une droite vectorielle au moins, il est nécessaire d'imposer une condition nécessaire. Comme on cherche une probabilité, on a nécessairement : $\mathbb{1}'_M\pi = 1$.

Le système s'écrit donc :

$$\begin{pmatrix} \mathrm{Id}_M - P \\ \mathbb{1}'_M \end{pmatrix}\pi = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

qu'on note $A\pi = e$ où $e$ est le dernier vecteur de la matrice $\mathrm{Id}_{M+1}$ : il est composé de M zéros puis un 1. On résoud ce système en $\pi = (A'A)^{-1}A'e$ où $A'e = \mathbb{1}_M$ (c'est la dernière ligne de $A$).

Notons que cet algorithme ne fonctionne que si $A'A$ est inversible, ce qui n'est pas nécessairement le cas. Il suffit de considérer $P = \begin{pmatrix} P_2 & 0 \\ 0 & 1 \end{pmatrix}$ où $P_2 \in \mathcal{S}_2$ ; la matrice $P$ ainsi définie admet 1 comme valeur propre (au moins) double, de sorte que $A'A$ n'est pas inversible. Avec des chiffres :

$$P_2 = \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix} \Longrightarrow A = \begin{pmatrix} 1-\alpha & -\beta & 0 \\ \alpha - 1 & \beta & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

or cette matrice $A$ est de rang 2.

# Annexe C

# Propriétés des MS

## C.1 Kurtosis

Soit le modèle MS à deux états, dépendant du régime uniquement dans la moyenne :

$$Y_t = \mu_{S_t} + \varepsilon_t$$

où $S_t \in \{1, 2\}$, $\mathbb{V}(\varepsilon_t) = \sigma^2$ et $\Delta\mu = \mu_2 - \mu_1 \neq 0$. On suppose la chaîne stationnaire :

$$\forall t, \quad \mathbb{P}(S_t = i) = \mathbb{P}(S_\infty = i)$$

Krolzig (1997) écrit page 23 :

$$\kappa = \frac{\mathbb{E}(y - \mathbb{E}y)^4}{\mathbb{V}(y)^2} = 3 + \frac{(\Delta\mu)^4 \pi_1 \pi_2 (1 - 3\pi_1\pi_2)}{(\sigma^2 + (\Delta\mu)^2 \pi_1 \pi_2)^2}$$

Comme $\pi_1\pi_2 \leq \frac{1}{4} < \frac{1}{3}$, ce calcul implique une kurtosis supérieure à 3, donc des queues systématiquement épaisses. La formule est fausse, la conclusion aussi.

### C.1.1 Variance

Le dénominateur du kurtosis est le carré de la variance : $\mathbb{V}(Y_t) = \mathbb{E}\mathbb{V}(Y_t|S_t) + \mathbb{V}\mathbb{E}(Y_t|S_t)$. Le premier terme est simple : $\mathbb{V}(Y_t|S_t) = \sigma^2$. Pour le second terme

$$\begin{aligned}
\mathbb{V}\mathbb{E}(Y_t|S_t) &= \mathbb{V}(\mu_{S_t}) \\
&= \mathbb{E}(\mu_{S_t}^2) - [\mathbb{E}(\mu_{S_t})]^2 \\
&= \mu_1^2\pi_1 + \mu_2^2\pi_2 - (\mu_1\pi_1 + \mu_2\pi_2)^2 \\
&= \mu_1^2\pi_1 + \mu_2^2\pi_2 - \mu_1^2\pi_1^2 - \mu_2^2\pi_2^2 - 2\mu_1\mu_2\pi_1\pi_2 \\
&= \mu_1^2\pi_1\pi_2 + \mu_2^2\pi_1\pi_2 - 2\mu_1\mu_2\pi_1\pi_2 \\
\Longrightarrow \mathbb{V}(Y_t) &= \sigma^2 + (\Delta\mu)^2\pi_1\pi_2 \qquad \text{(C.1)}
\end{aligned}$$

### C.1.2 Numérateur

Soit $\mu = \mathbb{E}(Y_t) = \pi_1\mu_1 + \pi_2\mu_2$, alors

$$\mathbb{E}(Y_t - \mu)^4 = \pi_1 \mathbb{E}\left[(Y_t - \mu)^4 | s_t = 1\right] + \pi_2 \mathbb{E}\left[(Y_t - \mu)^4 | s_t = 2\right]$$

Comme $Y_t|S_t$ est normale, il suffit de calculer le moment d'ordre 4 non centré d'une normale. Soit $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, $N = (X - \mu)/\sigma \rightsquigarrow \mathcal{N}(0, 1)$, donc :

$$\begin{aligned}
\mathbb{E}(X^4) &= \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\sigma + \mu\right)^4\right] = \mathbb{E}(\sigma N + \mu)^4 \\
&= \sigma^4\mathbb{E}(N^4) + 6\sigma^2\mu^2\mathbb{E}(N^2) + \mu^4 \\
\Longrightarrow \mathbb{E}(X^4) &= 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4
\end{aligned}$$

Par conséquent :

$$\begin{aligned}
\mathbb{E}(Y_t-\mu)^4 &= \pi_1\mathbb{E}\left[\mathcal{N}(\mu_1-(\mu_1\pi_1+\mu_2\pi_2),\sigma^2)^4\right]+\pi_2\mathbb{E}\left[\mathcal{N}(\mu_2-(\mu_1\pi_1+\mu_2\pi_2),\sigma^2)^4\right]\\
&= \pi_1\mathbb{E}\left[\mathcal{N}((\mu_1-\mu_2)\pi_2,\sigma^2)^4\right]+\pi_2\mathbb{E}\left[\mathcal{N}((\mu_2-\mu_1)\pi_1,\sigma^2)^4\right]\\
&= \pi_1\left(3\sigma^4+6\sigma^2\pi_2^2(\Delta\mu)^2+(\Delta\mu)^4\pi_2^4\right)+\pi_2\left(3\sigma^4+6\sigma^2\pi_1^2(\Delta\mu)^2+(\Delta\mu)^4\pi_1^4\right)\\
\Longrightarrow \mathbb{E}(Y_t-\mu)^4 &= 3\sigma^4+6\sigma^2(\Delta\mu)^2\pi_1\pi_2+(\Delta\mu)^4\pi_1\pi_2(\pi_1^3+\pi_2^3) \qquad \text{(C.2)}
\end{aligned}$$

### C.1.3 Kurtosis

De $(C.1)$ et $(C.2)$, on déduit l'expression du kurtosis :

$$\begin{aligned}
\kappa &= \frac{3\sigma^4+6\sigma^2(\Delta\mu)^2\pi_1\pi_2+(\Delta\mu)^4\pi_1\pi_2(\pi_1^3+\pi_2^3)}{\left(\sigma^2+(\Delta\mu)^2\pi_1\pi_2\right)^2}\\
&= \frac{3\sigma^4+6\sigma^2(\Delta\mu)^2\pi_1\pi_2+(\Delta\mu)^4\pi_1\pi_2(\pi_1^3+\pi_2^3)}{\sigma^4+(\Delta\mu)^4\pi_1^2\pi_2^2+2\sigma^2(\Delta\mu)^2\pi_1\pi_2}\\
&= 3+\frac{(\Delta\mu)^4\pi_1\pi_2(\pi_1^3+\pi_2^3)-3(\Delta\mu)^4\pi_1^2\pi_2^2}{\cdots}\\
\Longrightarrow \kappa &= 3+\frac{(\Delta\mu)^4\pi_1\pi_2}{(\sigma^2+(\Delta\mu)^2\pi_1\pi_2)^2}(\pi_1^3+\pi_2^3-3\pi_1\pi_2)
\end{aligned}$$

Soit :

$$\boxed{\kappa-3=\frac{\mathbb{E}\left[(y_t-\mu)^4\right]}{(\mathbb{E}(y_t-\mu)^2)^2}-3=\frac{(\Delta\mu)^4\pi_1\pi_2}{(\sigma^2+(\Delta\mu)^2\pi_1\pi_2)^2}\left(1-6\pi_1\pi_2\right)}$$

Et donc [1] :

$$\kappa\geq 3 \iff \pi_1\pi_2\leq\frac{1}{6} \iff \pi_1 \text{ ou } \pi_2\leq\pi^*=\frac{1}{2}-\frac{1}{\sqrt{12}}\approx 0.211$$

Finalement, les MS peuvent générer des lois à queues épaisses, mais pour certaines valeurs des paramètres uniquement : il est nécessaire que le temps passé dans la chaîne soit très asymétrique, beaucoup de temps dans un état, et donc beaucoup moins dans l'autre.

## C.2 Loi du temps de séjour

On explicite les lois des processus $T$ et $D$ définis dans la section $1.2.5$ :
- $T$ est une chaîne de Markov,
- chaque durée $D|T$ suit une loi géométrique.

Les paramètres de ces processus sont équivalents à ceux de la chaîne de Markov : ils contiennent la même information (autrement dit : on peut déduire les uns des autres, et les autres des uns).

### C.2.1 Loi jointe

Pour calculer la loi de $(T_1,\ldots,T_n,D_1,\ldots,D_n)$, on revient au processus d'état $S_t$ :

$$\begin{aligned}
\mathbb{P}(T_1,\ldots,T_n,D_1,\ldots,D_n) &= \mathbb{P}\Big(\underbrace{\bigcap_{i=1}^{n}\{S_{E_i}=\cdots=S_{F_i}=T_i\}}_{A_n},S_{F_n+1}\neq T_n\Big)\\
&= \mathbb{P}\left(S_{F_n+1}\neq T_n|A_n\right)\cdot\mathbb{P}(A_n)
\end{aligned}$$

Or :

$$\mathbb{P}\left(S_{F_n+1}\neq T_n|A_n\right)=\mathbb{P}(S_{F_n+1}\neq T_n|S_{F_n}=T_n)=q(T_n)$$

où $q(i)=1-p_{ii}$ est la probabilité de quitter l'état $i$. Soit :

$$B_k=\{S_{E_k}=\cdots=S_{F_k}=T_k\}$$

1. $x_0=\frac{1}{2}-\frac{1}{\sqrt{12}}$ est bien sûr une solution de l'équation $x(1-x)=\frac{1}{6}$, l'autre étant $1-x_0$.

alors :

$$\begin{aligned}\mathbb{P}(T_1,\ldots,T_n,D_1,\ldots,D_n) &= q(T_n)\cdot\mathbb{P}\left(B_n|B_{n-1},\ldots,B_1\right)\cdot\mathbb{P}(B_{n-1},\ldots,B_1)\\ &= q(T_n)\cdot p(T_{n-1},T_n)\cdot p(T_n)^{D_n-1}\cdot\mathbb{P}(B_{n-1},\ldots,B_1)\end{aligned}$$

Finalement :

$$\boxed{\mathbb{P}(T_1,\ldots,T_n,D_1,\ldots,D_n) = q(T_n)\cdot\prod_{i=2}^{n}p(T_{i-1},T_i)\prod_{i=1}^{n}p(T_i)^{D_i-1}\cdot\pi(T_1)} \tag{C.3}$$

### C.2.2 Loi de $T$

On obtient la loi de $T_1,\ldots,T_n$ en intégrant $(C.3)$ en $(D_1,\ldots,D_n)\in(\mathbb{N}^*)^n$ :

$$\mathbb{P}(T_1,\ldots,T_n) = \sum_{(D_1,\ldots,D_n)\in(\mathbb{N}^*)^n}\left[q(T_n)\prod_{i=2}^{n}p(T_{i-1},T_i)\prod_{i=1}^{n}p(T_i)^{D_i-1}\pi(T_1)\right]$$

Or :

$$\sum_{(D_1,\ldots,D_n)\in(\mathbb{N}^*)^n}\prod_{i=1}^{n}p(T_i)^{D_i-1} = \sum_{(D_2,\ldots,D_n)\in(\mathbb{N}^*)^n}\prod_{i=2}^{n}p(T_i)^{D_i-1}\underbrace{\sum_{D_1\in\mathbb{N}^*}p(T_1)^{D_1-1}}_{=\frac{1}{1-p(T_1)}} = \prod_{i=1}^{n}q(T_i)^{-1}$$

Donc :

$$\boxed{\mathbb{P}(T_1,\ldots,T_n) = \prod_{i=2}^{n}\frac{p(T_{i-1},T_i)}{q(T_{i-1})}.\pi(T_1)} \tag{C.4}$$

Par conséquent $(T_n)$ est une chaîne de Markov dont la matrice de transition $\tilde{P}$ a pour terme général :

$$\tilde{p}_{ij} = \begin{cases} 0 & \text{si i = j}\\ \frac{p_{ij}}{1-p_{ii}} & \text{sinon}\end{cases}$$

Par construction, le processus $T$ ne peut rester dans le même état : $\tilde{p}_{ii}=0$.

En outre, $T$ n'est pas stationnaire : $\forall n\in\mathbb{N}^*$, le vecteur $(\mathbb{P}(T_n=m))_m$ est égal à $(\tilde{P}')^{n-1}(\pi_1-\tilde{\pi}_\infty)+\tilde{\pi}_\infty$. Donc, si $\pi_1\neq\tilde{\pi}_\infty=\mathbb{P}(T_\infty)$, $T$ n'est pas stationnaire. Comme[2] : $\tilde{\pi}_{\infty,i}=\lambda(1-p_{ii})\pi_i$, si $\pi_1=\pi_\infty$, alors $\pi_1\neq\tilde{\pi}_\infty$ ssi $p_{ii}$ ne dépend pas de $i$.

### C.2.3 Loi de $D|T$

De $(C.3)$ et $(C.4)$, et avec :

$$\mathbb{P}(D_1,\ldots,D_n|T_1,\ldots,T_n) = \frac{\mathbb{P}(T_1,\ldots,T_n,D_1,\ldots,D_n)}{\mathbb{P}(T_1,\ldots,T_n)}$$

on déduit :

$$\boxed{\mathbb{P}(D_1,\ldots,D_n|T_1,\ldots,T_n) = \prod_{i=1}^{n}\left\{(1-p(T_i)).p(T_i)^{D_i-1}\right\}} \tag{C.5}$$

Autrement dit : $D|T$ suit une loi géométrique de paramètre $q(T)=1-p(T)$. L'indépendance des $(D_i|T_i)_i$ découle aussi de $(C.5)$.

2. En effet (en notant $\tilde{\pi}$ pour $\tilde{\pi}_\infty$) : $\tilde{\pi}_j=\sum_i(\tilde{P})_{ij}\tilde{\pi}_i=\sum_{inej}\frac{p_{ij}}{1-p_{ii}}\tilde{\pi}_i$. Soit alors $u_i=\tilde{\pi}_i(1-p_{ii})^{-1}$, alors $u_j=\sum_{i=1}^{M}p_{ij}u_i$, de sorte que $u$ est proportionnel à $\pi$ (la loi stationnaire de la chaîne $S$. Finalement : $\tilde{\pi}_i=\lambda(1-p_{ii})\pi_i$.

**Rappel sur la loi géométrique** La variable aléatoire $X$ suit une loi géométrique de paramètre $p \in [0;1]$, notée $\mathcal{G}(p)$, si : $\forall n \in \mathbb{N}^*,\ \mathbb{P}(X = n) = (1-p)^{n-1}p$. [ Attention à la définition : il en existe une version alternative, sur $\mathbb{N}$ au lieu de $\mathbb{N}^*$.]

Quelques propriétés :
- Moyenne : $p^{-1}$, variance : $\frac{1-p}{p^2}$
- $\mathbb{P}(X = n)$ décroît quand $n$ croît,
- $\mathbb{P}(X \geq n) = (1-p)^{n-1}$
- $\mathbb{P}(X = n | X \geq n) = p$ : c'est une caractérisation de la loi géométrique. Le sens $\Longrightarrow$ est trivial ; réciproque, notons $p_n = \mathbb{P}(X = n)$, alors :

$$\mathbb{P}(X = n | X \geq n) = p \quad \Longrightarrow \quad p_n = p \sum_{i \geq n} p_i = p\left(p_n + \frac{p_{n+1}}{p}\right)$$

soit : $(1-p)p_n = p_{n+1}$. Par récurrence : $p_n = (1-p)^{n-1}p_1$. Comme $\sum_{n \in \mathbb{N}^*} p_n = 1$, $p_1 = p$. CQFD.

### C.2.4 Équivalence des paramètres

Il est équivalent de connaître les paramètres du processus $(D, T)$ (durées moyennes et matrice de transition $\tilde{P}$) et ceux de $S$ (matrice de transition $P$).

Déduire $(d, \tilde{P})$ de $P$ est déjà connu ; la réciproque se déroule en deux étapes :
- la diagonale de $P$ dérive des durées : $p_{ii} = 1 - d_i^{-1}$,
- les termes extra-diagonaux de $P$ s'écrivent : $p_{ij} = \tilde{p}_{ij}(1 - p_{ii}) = \frac{\tilde{p}_{ij}}{d_i}$.

# Annexe D

# Transition probability matrix

Cette partie vise deux objectifs :
- présenter rapidement la théorie,
- fournir des algorithmes détaillés de traitement.

## D.1 Définitions

Soit $M_n$ l'ensemble des matrices carrées d'ordre $n$ à coefficients réels. Une matrice $S$ de $M_n$ est dite stochastique si elle remplit les deux conditions suivants :

$$\forall (i,j) \in [\![1;n]\!], \qquad A_{ij} \geq 0 \tag{D.1}$$

$$\forall i \in [\![1;n]\!], \qquad \sum_j A_{ij} = 1 \tag{D.2}$$

On note $\mathcal{S}_n$ l'ensemble des matrices stochastiques d'ordre $n$, et $\mathcal{S} = \bigcup_{n\in\mathbb{N}^*} \mathcal{S}_n$.
Remarques :
- $(D.1)$+$(D.2) \implies A_{ij} \leq 1$
- Chaque ligne de $A$ est une distribution de probabilité.

Exemples, dans $M_2$ : $\begin{pmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{pmatrix}$.
La matrice $\begin{pmatrix} 1.5 & -0.5 \\ 0.3 & 0.7 \end{pmatrix}$ ne vérifie pas $(D.1)$, tandis que $\begin{pmatrix} 1.5 & 0 \\ 0.3 & 0.7 \end{pmatrix}$ ne vérifie pas $(D.2)$

## D.2 Structure de $\mathcal{S}$

### D.2.1 Algèbre linéaire

$\mathcal{S}_n$ n'est pas un sous-espace vectoriel de $M_n$, pour au moins trois raisons (quoique chacune des raisons suivantes soit individuellement suffisante) :
- la matrice nulle n'est pas dans $\mathcal{S}_n$,
- si $A, B \in \mathcal{S}_n$ alors $A + B \notin \mathcal{S}_n$,
- si $A \in \mathcal{S}_n$ et $\lambda \in \mathbb{R}$, alors $\lambda \cdot A$ n'appartient pas nécessairement à $\mathcal{S}_n$ : ce n'est le cas que si $\lambda = 1$.

### D.2.2 Structure topologique

Pour la topologie classique sur $M_n$, $\mathcal{S}_n$ est fermé dans $M_n$ (donc compact).

#### D.2.2.1 Inversibilité des matrices stochastiques

Toutes les matrices stochastiques ne sont pas inversibles : $\mathcal{S}_n \not\subset \mathrm{GL}_n$ (où $\mathrm{GL}_n$ désigne le groupe linéaire de $M_n$, *ie* l'ensemble des matrices inversibles de $M_n$). Toutefois, comme $\mathrm{GL}_n$ est dense dans $M_n$, $\mathcal{S}_n \cap \mathrm{GL}_n$ est dense dans $\mathcal{S}_n$ : on peut toujours trouver une suite de matrices inversibles qui converge vers n'importe quel élément de $\mathcal{S}_n$.

#### D.2.2.2 Diagonalisabilité des matrices stochastiques

Tous les éléments de $\mathcal{S}_n$ ne sont pas nécessairement diagonalisables ; néanmoins, comme pour $\mathrm{GL}_n$, l'ensemble des matrices diagonalisables $D_n$ est dense dans $M_n$ de sorte que $\mathcal{S}_n \cap D_n$ est dense dans $\mathcal{S}_n$ :

$$\forall M \in \mathcal{S}_n, \exists (M_n)_{n\in\mathbb{N}} \in (\mathcal{S}_n \cap D_n)^{\mathbb{N}} \quad / \quad M_n \to_{n\to\infty} M$$

On peut ainsi toujours trouver une (infinité de) suite de matrices diagonalisables qui converge vers un élément de $\mathcal{S}_n$.

Voici deux exemples de matrices de transition non diagonalisables mais irréductible et apériodique :

$$M_1 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 1 \end{pmatrix} \qquad M_2 = \frac{1}{2}\begin{pmatrix} 0 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Le polynôme caractéristique de $M_1$ est $-(\lambda-1)(\lambda-\frac{1}{4})^2$. Comme il est scindé, $M_1$ peut encore être diagonalisable. SEP$(M_1, 1)$ est une droite vectorielle (normal, cf Perron-Frobenius), mais $SEP(M_1, \frac{1}{4}) = \mathrm{Vect}(1, -2, -1)$ n'est pas de dimension 2 ! ! Donc $M_1$ est non diagonalisable. Le cas de $M_2$ est encore plus facile : son polynôme caractéristique, $(1-\lambda)(\lambda^2+\frac{1}{4})$, n'est même pas scindé. Les chaînes de Markov associées à ces deux matrices de transition sont bien sûr irréductibles et apériodiques.

### D.2.3 Métrique induite par $M_n$

$(M_n, +, \cdot)$ est un espace vectoriel euclidien [1]. Le produit scalaire classique est $< A, B >= \mathrm{trace}\,({}^tAB) = \sum_{ij} a_{ij}b_{ij}$. Ce produit scalaire induit une norme (euclidienne, par définition [2] : $||A||_2 = \sqrt{< A, A >} = (\sum_{ij} a_{ij}j^2)^{1/2}$. Mais il existe d'autres normes : $||A||_\infty = \sup_{ij} |A_{ij}|$ ou $||A||_1 = \sum_{ij} |A_{ij}|$. $(M_n, +, \cdot)$ étant un espace vectoriel de dimension finie, il est complet et donc toutes les normes sont équivalentes.

Bien entendu, toute norme induit une distance [3] par $d(A, B) = ||A - B||$.

Il existe au moins une autre semie-distance : le $q$ de Dobrushin (cf infra).

### D.2.4 Autres propriétés

$\mathcal{S}_n$ est convexe : $\forall A, B \in \mathcal{S}_n, \lambda \in [0; 1], \qquad \lambda \cdot A + (1-\lambda) \cdot B \in \mathcal{S}_n$

Propriétés de groupe :
- $(\mathcal{S}_n, +)$ n'est pas un groupe, puisque l'addition n'est pas une opération interne : elle rompt $(D.2)$.
- $(\mathcal{S}_n, \times)$ n'est pas un groupe, car $\mathcal{S}_n \not\subset \mathrm{GL}_n$. En revanche, $\mathcal{S}_n$ est bien stable par multiplication : $A, B \in \mathcal{S}_n \implies AB \in \mathcal{S}_n$, puisque $\forall i$

$$\sum_j (AB)_{ij} = \sum_{jk} A_{ik}B_{kj} = \sum_k A_{ik} \sum_j B_{kj} = \sum_k A_{ik} = 1$$

## D.3 Propriétés

### D.3.1 Convergence des itérées

C'est un résultat classique que les hypothèses d'irréductibilité et d'apériodicité garantissent la convergence hyperbolique de $P^h$ vers $\mathbb{1}\pi'$. La convergence dépend du spectre de $P$. On sait que 1 est valeur propre de $P$ et que toutes les autres sont de module $< 1$. La vitesse de convergence se ramène en fait à la question de l'ordre de multiplicité de 1. Soit $\lambda_1, \ldots, \lambda_M$ les valeurs propres de $P$, répétées selon leur ordre de multiplicité et triés dans l'ordre décroissant de leur valeur absolue : $1 = \lambda_1 \leq |\lambda_2| \leq \cdots \leq |\lambda_M|$. Soit $\lambda_* = |\lambda_2|$ :

1. *ie* muni d'un produit scalaire, qui n'est rien d'autre qu'une forme bilinéaire symétrique définie positive, notée ici : $< \cdot, \cdot >$.
2. Rappel : une norme est dite euclidienne si elle dérive d'un produit scalaire. Toutes les normes ne sont pas euclidiennes, mais on connaît une condition nécessaire et suffisante : une norme est euclidienne si et seulement si elle satisfait l'identité du parallélogramme :

$$\forall (x, y), \qquad ||x-y||^2 + ||x+y||^2 = 2(||x||^2 + ||y||^2)$$

3. Réciproque fausse, bien sûr.

– si $\lambda_* = 1$, alors
  – $P^h$ peut ne pas converger (cas $P$ périodique, par exemple)
  – $P^h$ peut converger, mais les lignes de $P^\infty$ ne sont pas nécessairement identique. C'est le cas si $P$ n'est pas irréductible, c'est-à-dire s'il existe au moins deux classes d'équivalence pour la relation de communication entre états [4]
– si $\lambda_* < 1$, alors la vitesse de convergence de $P^h$ vers $\mathbb{1}\pi'$ est liée à $\lambda_*$, dans un sens que nous précisons ci-après.

Si la chaîne prend un nombre d'état fini (c'est le cas qui nous intéresse ici ; nous considérons des chaînes avec un nombre très restreint d'états), alors $\lambda_* < 1$ si et seulement si la chaîne de Markov est irréductible et apériodique. Une condition suffisante est : $\exists n \in \mathbb{N} \ / \ \forall (i,j) \in [\![1;M]\!]^2,\ (P^n)_{ij} > 0$, ce qui est fréquemment le cas en pratique.

Dans le cas $\lambda_* < 1$, la convergence de $P^h$ vers $\mathbb{1}\pi'$ est de l'ordre de $h^{J-1}\lambda_*^h$, où $J$ est la taille du plus grand bloc de Jordan de $P$ :

– Si $P$ est diagonalisable, alors $J = 1$, de sorte que la convergence est hyperbolique, à la vitesse $\lambda_*$.
– Si $P$ n'est pas diagonalisable, alors $J > 1$, mais la convergence ralentit peu.

Dans la suite, on s'intéresse uniquement au cas diagonalisable au principe que c'est le cas en pratique. Cette preuve présente l'avantage de mettre clairement en évidence la borne. Il existe d'autres preuves (l'une figure dans `http://agreg-maths.univ-rennes1.fr/documentation/docs/agreg-Sto.pdf`), qui mettent moins clairement en évidence la borne.

#### D.3.1.1 Preuve dans le cas diagonalisable

Dans le cas de convergence $\lambda_* < 1$, on montre que la norme spectrale de $P^h - \mathbb{1}\pi'$ est $\lambda^h$ : la convergence des itérées vers la loi limite est exponentielle. Supposons que $P$ soit diagonalisable, alors ses sous-espaces propres sont en somme directe. On distingue deux cas :

– Si $0$ n'est pas une valeur propre de $P$, alors :

$$\begin{cases} \forall \lambda \in Sp(P)\backslash\{1\}, & \text{SEP}(P,\lambda) &=& \text{SEP}(P - \mathbb{1}\pi', \lambda) \\ & \text{SEP}(P,1) &=& \text{SEP}(P - \mathbb{1}\pi', 0) \end{cases}$$

– Si $0$ est valeur propre de $P$, alors

$$\begin{cases} \forall \lambda \in Sp(P)\backslash\{0,1\}, & \text{SEP}(P,\lambda) &=& \text{SEP}(P - \mathbb{1}\pi', \lambda) \\ & \text{SEP}(P,0) \oplus \text{SEP}(P,1) &=& \text{SEP}(P - \mathbb{1}\pi', 0) \end{cases}$$

**Cas 1 : 0 non valeur propre** Soit $\lambda \neq 1$ et $S \in \text{SEP}(P,\lambda)$, alors $PX = \lambda X$, soit en pré-multipliant par $\pi'$ : $\pi' X = \lambda \pi' X \Longrightarrow \pi' X = 0$ donc $X \in \text{SEP}(P - \mathbb{1}\pi', \lambda)$. La réciproque nécessite que $\lambda \neq 0$.
Pour le second morceau, $\text{SEP}(P,1) = \text{SEP}(P - \mathbb{1}\pi', 0)$, il suffit de se rappeler que $\text{SEP}(P,1)$ est la droite vectorielle engendrée par $\mathbb{1}$.

**Cas 2 : 0 non valeur propre** Pour démontrer le second morceau, il suffit de preouve que $\text{SEP}(P - \mathbb{1}\pi', 0) \subset \text{SEP}(P,0) + \text{SEP}(P,0)$. Siot $X$ tel que $PX = \mathbb{1}\pi' X$, alors $X = X - \pi' X \mathbb{1} + \pi' X \mathbb{1}$, or $X - \mathbb{1}\pi' X \in \text{SEP}(P,0)$ et $\pi' X \mathbb{1} \in \text{SEP}(P,1)$. CQFD.

**Conclusion** Les normes sur l'espace des matrices réelles étant équivalentes, choisissons en une qui nous arrange. Le rayon spectral :

$$||A|| = \max\{|\lambda| / \lambda \in Sp(A)\}$$

possède la propriété intéressante : $||AB|| = ||A|| \cdot ||B||$, puisque : $P^h - \mathbb{1}\pi' = (P - \mathbb{1}\pi')^h \Longrightarrow ||P^h - \mathbb{1}\pi'|| = ||P - \mathbb{1}\pi'||^h = \lambda_*^h$. La convergence est donc exponentielle. Voir aussi la section **??**.

#### D.3.1.2 La seconde valeur propre

Le résultat précédent met en valeur l'importance de $\lambda_* = |\lambda_2|$. Cette quantité est délicate à estimer autrement qu'en diagonalisant la matrice. De nombreux travaux cherchent à l'exprimer le plus simplement possible à partir des termes de la matrice. Signalons deux majorations :

4. Rappel : on définit une relation de communication entre états, et pour tous les états récurrents (par opposition aux états transitoires, qu'on ne peut pas revisiter une fois qu'on les a quittés), on définit des classes d'équivalence : appartiennent à une même classe d'équivalence deux états qui communiquent entre eux.

– la première est assez utile :

$$|\lambda| \leq 1 - \sum_{j=1}^{n} \min_i p_{ij}$$

– la seconde l'est moins :

$$|\lambda| \leq \sum_{j=1}^{n} \max_i p_{ij} - 1$$

puisque les matrices de transition qui nous intéressent ont fréquemment la diagonale chargée, de sorte que le majorant est supérieur à 1 !

Dans le cas $M = 2$ avec diagonale dominante $M = \begin{pmatrix} p_1 & q_1 \\ q_2 & p_2 \end{pmatrix}$ avec $p_i \geq q_i \iff p_i \geq 0.5$, ces deux majorations sont exactes : la seconde valeur propre est $p_1 + p_2 - 1$.

**Dobrushin** Dobrushin a montré un résultat intéressant. Soit $q$ l'application qui à une matrice stochastique $P$ associe :

$$q(P) = \frac{1}{2} \max_{ik} \left( \sum_{j=1}^{n} |p_{ij} - p_{ik}| \right) = 1 - \min_{ik} \sum_{j=1}^{n} \min(p_{ij}, p_{ik})$$

$q$ est une semi-norme [5] : c'est (au facteur multiplicatif $1/2$ près) le maximum des distances entre les lignes de $P$. On montre [6] que :

$$\lambda_* = \lim_{n \to \infty} q(P^n)^{1/n}$$

[ La convergence est décroissante.] En pratique, $q(P^n)$ est assez rapidement proche de $\lambda_*^n$ ($n = 10$).

## D.4 Structures particulières

Fréquemment, le modélisateur souhaite imposer une forme particulière à la matrice de transition : les contraintes traduisent des a priori sur le mécanisme étudié. En outre, d'un point de vue mathématique et informatique, il apparaît intéressant d'envisager un grand nombre de spécifications.

Deux types de spécifications sont étudiés :
– les contraintes sur une matrice de transition : typiquement, on peut souhaiter imposer restreindre la communication dans la chaîne de Markov sous-jacente, ce qui se traduit par des termes nuls dans la matrice de transition,
– les opérations sur les matrices de transition : comment créer une nouvelle matrice de transition à partir de deux (ou plus) matrices de transition.

### D.4.1 Matrices contraintes

#### D.4.1.1 Processus markoviens d'ordre supérieur à 1

Un processus markovien est d'ordre $k \in \mathbb{N}^*$ si :

$$\forall t, \qquad \mathbb{P}(S_t|(S_i)_{i<t}) = \mathbb{P}(S_t|S_{t-1}, \ldots, S_{t-k}) \neq \mathbb{P}(S_t|S_t, \ldots, S_{t-k+1})$$

Un tel processus est donc caractérisé par des transitions plus complexes : la distribution de $S_t$ ne dépend pas uniquement de $S_{t-1}$ mais aussi de $S_{t-2}, \ldots, S_{t-k}$. Toutefois, $\tilde{S}_t = (S_{t-k+1}, \ldots, S_t)$ est un processus markovien d'ordre 1 dont la matrice de transition est largement trouée. En effet : $\forall (s_0, s_2) \in [\![1; M]\!] \times [\![1; M]\!]$, $\forall (s_1, s_1') \in [\![1; M]\!]^{k-1} \times [\![1; M]\!]^{k-1}$,

$$\mathbb{P}\left(\tilde{S}_t = (s_1', s_2) | \tilde{S}_{t-1} = (s_0, s_1)\right) = \begin{cases} 0 & \text{si } s_1 \neq s_1' \\ p_{s_1 s_2} & \text{sinon} \end{cases}$$

5. Elle ne vérifie bien sûr pas $||x|| = 0 \implies x = 0$, puisqu'ici $q(P) = 0 \implies P = P.1\!\!1'$. Néanmoins, $q$ vérifie l'inégalité triangulaire $q(P+Q) \leq q(P) + q(Q)$ et l'homogénéité $q(\lambda P) = |\lambda| q(P)$. De plus, $q(PQ) \leq q(P)q(Q)$. On peut dériver une distance de $q$ : $d(P, Q) = q(P, Q)$.

6. Voir http ://arxiv.org/PS_cache/math/pdf/0307/0307056.pdf par exemple.

de sorte que sur les $M^k$ transitions à partir de l'état $(s, s_0)$, seulement $M$ sont non nulles. Donc $M^{k+1}$ de l'ensemble des $M^{2k}$ transitions sont non nulles. Un exemple dans le cas $M = 2, k = 2$ :

| | 11 | 12 | 21 | 22 |
|---|---|---|---|---|
| 11 | $p_{11\to1}$ | $p_{11\to2}$ | 0 | 0 |
| 12 | 0 | 0 | $p_{12\to1}$ | $p_{12\to2}$ |
| 21 | $p_{21\to1}$ | $p_{21\to2}$ | 0 | 0 |
| 22 | 0 | 0 | $p_{22\to1}$ | $p_{22\to2}$ |

#### D.4.1.2 Modélisation de Hamilton (1989)

Le modèle de Hamilton (1989) (cf section A.6.1, page 44) nécessite la mémoire des $p$ états précédents de la chaîne : il faut donc remplacer la chaîne $S_t$ par $\tilde{S}_t = (S_{t-p}, \dots, S_t) \in \mathbb{R}^{p+1}$, et comme précédemment de nombreux termes de la nouvelle matrice de transition sont nuls. Précisément , sur chacue des $M^{p+1}$ lignes, seuls $M$ termes sont non nuls, donc chaque terme est répété $M^{2(p+1)}/(M^{p+1} \cdot M) = M^p$ fois. Exemple[7].

#### D.4.1.3 Processus markoviens pour la détection de rupture

Frühwirth-Schnatter (2006) présente une modélisation de Chibb (1998) adaptée à la détection des ruptures : $\Delta S_t \in \{0, 1\}$. La chaîne part de l'état 1, puis y reste avec probabilité $\alpha_1$ et va vers l'état 2 avec probabilité $1 - \alpha_1$. Ensuite, la chaîne reste dans l'état 2 avec proabilité $\alpha_2$ et va à l'état 3 avec probabilité $1 - \alpha_2$. En généralisation, la matrice de transition s'écrit donc :

$$P = \begin{pmatrix} \alpha_1 & 1-\alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & 1-\alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \alpha_{M-1} & 1-\alpha_{M-1} \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

où $\forall i \in [\![1; M-1]\!], \alpha_i \in [0; 1]$. $P$ est donc triangulaire supérieure. La distribution stationnaire est $(0, \cdots, 0, 1)$.

### D.4.2 Combinaisons de matrices de transition

#### D.4.2.1 Matrix product

Comme nous l'avons déjà noté : $\forall m \in \mathbb{N}^*, \ \forall (A_1, \dots, A_m) \in \mathcal{S}_n^m$,

$$A_1 \times \cdots \times A_m \in \mathcal{S}_n$$

#### D.4.2.2 Union of TP

[ Trouvé dans la programmation par Warne de l'estimation des msvar, et utilisé dans Grégoir-Lenglart. Othman Bouabdallah m'a indiqué que cette modélisation avait été utilisée pour des estimations sur données monétaires.] Fondamentalement, il s'agit de composer des matrices de transition par produit de Kronecker, en ayant constaté que : $A, B \in \mathcal{S} \implies A \otimes B \in \mathcal{S}$.

7. Soit $P = \begin{pmatrix} p_1 & q_1 \\ q_2 & p_2 \end{pmatrix}$, les matrices $\tilde{P}$ associées avec $p = 1$ et $p = 2$ sont :

| | 11 | 12 | 21 | 22 |
|---|---|---|---|---|
| 11 | $p_1$ | $q_1$ | 0 | 0 |
| 12 | 0 | 0 | $q_2$ | $p_2$ |
| 21 | $p_1$ | $q_1$ | 0 | 0 |
| 22 | 0 | 0 | $q_2$ | $p_2$ |

et

| | 111 | 112 | 121 | 122 | 211 | 212 | 221 | 222 |
|---|---|---|---|---|---|---|---|---|
| 111 | $p_1$ | $q_1$ | 0 | 0 | 0 | 0 | 0 | 0 |
| 112 | 0 | 0 | $q_2$ | $p_2$ | 0 | 0 | 0 | 0 |
| 121 | 0 | 0 | 0 | 0 | $p_1$ | $q_1$ | 0 | 0 |
| 122 | 0 | 0 | 0 | 0 | 0 | 0 | $q_2$ | $p_2$ |
| 211 | $p_1$ | $q_1$ | 0 | 0 | 0 | 0 | 0 | 0 |
| 212 | 0 | 0 | $q_2$ | $p_2$ | 0 | 0 | 0 | 0 |
| 221 | 0 | 0 | 0 | 0 | $p_1$ | $q_1$ | 0 | 0 |
| 222 | 0 | 0 | 0 | 0 | 0 | 0 | $q_2$ | $p_2$ |

Si l'on écrit les états comme dans cet exemple, alors quelques lignes suffisent à déterminer l'expression de $\tilde{P}$ en fonction de $P$ :

$$\tilde{P} = \mathbb{1}_M \otimes \mathrm{Id}_{M^{p-1}} \otimes \bar{P}$$

où $\bar{P}$ est la matrice $M \times M^2$ qui est nulle sauf la diagonale (par blocs) formée des lignes de la matrice $P$.

Pour montrer la stabilité de $\mathcal{S}$ par le produit de Kronecker, il est élégant d'utiliser les chaînes de Markov. Soient en effet $Z_t$ et $W_t$ deux chaînes de Markov indépendantes, $P^Z$ et $P^W$ leurs matrices de transition ; le processus $S_t = (Z_t, W_t)$ est clairement une nouvelle chaîne de Markov, dont on calculer la matrice de transition :

$$\begin{aligned} & \mathbb{P}(Z_t = j_z, W_t = j_w | Z_{t-1} = i_z, W_{t-1} = i_w) \\ = \; & \mathbb{P}(Z_t = j_z | Z_{t-1} = i_Z, W_t = j_w, W_{t-1} = i_w) \mathbb{P}(W_t = j_w | Z_{t-1} = i_z, W_{t-1} = j_w) \\ = \; & P^Z_{i_z j_z} P^W_{i_w j_w} \end{aligned}$$

de sorte que $P^S = P^Z \otimes P^W$, si les états sont nommés ainsi : $S = (Z-1) \cdots M_W + W$, c'est-à-dire :

$$\begin{array}{ccc} S = 1 & \iff & Z = 1 \text{ et } W = 1 \\ S = 2 & \iff & Z = 1 \text{ et } W = 2 \\ \vdots & & \vdots \\ S = M_W & \iff & Z = 1 \text{ et } W = M_W \\ S = M_W + 1 & \iff & Z = 2 \text{ et } W = 1 \\ \vdots & & \vdots \end{array}$$

Cette modélisation présente l'avantage certain de diminuer drastiquement le nombre de paramètres libres de la matrice de transition. Si $S_t^i$ prend $M_i$ états, alors $S_t = (S_t^1, \ldots, S_t^n)$ prend $M = \prod_{i=1}^n M_i$ états, mais ne compte que $\sum_{i=1}^n M_i(M_i - 1)$ paramètres libres au lieu de $M(M-1)$. À titre d'exemple, avec $M = 2$, on réduit le nombre de paramètres de $2^n(2^n - 1)$ à $2n$.

Notation : $\bigotimes_{i=1}^{n} \mathcal{S}_{m_i} = \mathcal{S}_{M_1} \otimes \cdots \otimes \cdots \mathcal{S}_{M_n} = \{P_1 \otimes \cdots \otimes P_n / \forall i \in [\![1; n]\!], P_i \in \mathcal{S}_{M_i}\}$.

Grégoir et Lenglart (2000) utilisent cette modélisation : le second processus traduit l'importance de la nouvelle information.

**Factorisation sous cette forme** Soit $(M_1, \ldots, M_n)$, $M = \prod_i M_i$ et $P \in \mathcal{S}_M$, comment projeter $P$ sur $\mathcal{S}(M) = \bigotimes_{i=1}^n \mathcal{S}_{M_i}$ ? On cherche $\tilde{P} \in \mathcal{S}(M)$ qui minimise la quantité $d(P, \tilde{P})$ où $d$ est une distance, par exemple la distance associée à la semi-norme de Dobrushin.

Pour la norme dérivée du produit scalaire trace, je conjecture la solution suivante. Notons $Q = Q_1 \otimes \cdots \otimes Q_n$ la projection orthogonale recherchée, alors :

$$\forall k \in [\![1; n]\!], \forall (i, j) \in [\![1; M_k]\!]^2, \qquad q_{ij}^k = \frac{M_k}{M} \sum (P \odot A_{ij}^k) \tag{D.3}$$

où :

$$A_{ij}^k = J_{M_1} \otimes \cdots \otimes J_{M_{k-1}} \otimes E_{ij,M_k} \otimes J_{M_{k+1}} \cdots \otimes J_{M_n}$$

où $J_M$ est la matrice carrée d'ordre $M$ composée uniquement de 1, et $E_{ij,M_k}$ est la matrice carrée d'ordre $M_k$ avec des zéros partout sauf un 1 en position $(i, j)$. Seuls $M/M_k$ termes de la somme $\sum(P \odot A_{ij}^k)$ sont non nuls de sorte que $q_{ij}^k$ n'est autre que la moyenne de toutes les bonnes transitions. $(D.3)$ provient d'un examen du cas $M = (2, 2)$ : $P = P_1 \otimes P_2$

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} p_{11}^1 & p_{12}^1 \\ p_{21}^1 & p_{22}^1 \end{pmatrix} \otimes \begin{pmatrix} p_{11}^2 & p_{12}^2 \\ p_{21}^2 & p_{22}^2 \end{pmatrix} = \begin{pmatrix} p_{11}^1 p_{11}^2 & p_{11}^1 p_{12}^2 & p_{12}^1 p_{11}^2 & p_{12}^1 p_{12}^2 \\ p_{11}^1 p_{21}^2 & p_{11}^1 p_{22}^2 & p_{12}^1 p_{21}^2 & p_{12}^1 p_{22}^2 \\ p_{21}^1 p_{11}^2 & p_{21}^1 p_{12}^2 & p_{22}^1 p_{11}^2 & p_{22}^1 p_{12}^2 \\ p_{21}^1 p_{21}^2 & p_{21}^1 p_{22}^2 & p_{22}^1 p_{21}^2 & p_{22}^1 p_{22}^2 \end{pmatrix}$$

de sorte que :

$$\left\{ \begin{array}{rcccl} p_{11} + p_{12} & = & p_{11}^1(p_{11}^2 + p_{12}^2) & = & p_{11}^1 \\ p_{21} + p_{22} & = & p_{11}^1(p_{21}^2 + p_{22}^2) & = & p_{11}^1 \end{array} \right.$$

D'où l'idée de retenir la moyenne de ces deux lignes :

$$p_{11}^1 = \frac{p_{11} + p_{12} + p_{21} + p_{22}}{2}$$

$(D.3)$ ne fait généraliser cela.

L'application qui à $P \in \mathcal{S}_M$ associe $Q \in \bigotimes_{i=1}^{n} \mathcal{S}_{M_i}$ est linéaire et c'est un projecteur [8].

#### D.4.2.3 Convex combination

$\forall m \in \mathbb{N}^*,\ \forall (A_1, \ldots, A_m) \in \mathcal{S}_n^m \forall (\lambda_1, \ldots, \lambda_m) \in [0;1]^m$ :

$$\sum_{i=1}^{m} \lambda_i = 1 \quad \Longrightarrow \quad \sum_{i=1}^{m} \lambda_i A_i \in \mathcal{S}_n$$

8. Pour la preuve, il suffit d'écrire :

$$Q \odot A_{ij}^k = Q_1 \otimes \cdots \otimes Q_{k-1} \otimes (E_{ij,M_k} \cdot Q_k) \otimes Q_{k-1} \otimes \cdots \otimes Q_n$$

Puis $\sum(A \otimes B) = (\sum A) \cdot (\sum B)$, de sorte que $\sum(Q \odot A_{ij}^k) = \frac{M}{M_k} q_{ij}^k$. $Q$ est donc invariant.

# Annexe E

# Modélisation avancée

Dans le cas d'une variable dépendante quantitative, plusieurs améliorations peuvent être incorporées :
- prise en contrainte de contraintes sur les paramètres VAR : nous avons étudié ce cas en section **??**,
- hypothèses supplémentaires sur la matrice de transition : pour enrichir l'analyse des transitions, par exemple en multipliant le nombre d'états sans multiplier pour autant le nombre de paramètres à estimer,
- hypothèses supplémentaires sur la variance : les cas homoscédastiques ou hétéroscédastiques peuvent intéresser le modélisateur.

## E.1 Modélisation avancée de la matrice de transition

Nous envisageons ici des alternatives à l'absence de modélisation de la matrice de transition, qui correspond au cas général de $M(M-1)$ paramètres libres sur les $M^2$. Ces alternatives reposent sur des paramétrisations différentes, et il convient d'adapter la phase d'estimation. Ces adaptations sont présentées en section E.1.2.

### E.1.1 Les formes envisagées

L'essentiel a été décrit dans la section D.4 ; il existe une dernière modélisation, qualitativement différente : les Time-varying transition probabilities. La matrice de transition dépend alors du temps, de la façon suivante : $\mathbb{P}(S_{t+1}=j|S_t=i)=p_{ij}^t$. **Références : qui introduit ça ?** Dans le cas $M=2$, il est usuel de modéliser la matrice de transition ainsi :

$$\begin{pmatrix} \frac{F(X_t\beta_1)}{1+F(X_t\beta_1)} & \frac{1}{1+F(X_t\beta_1)} \\ \frac{F(X_t\beta_2)}{1+F(X_t\beta_2)} & \frac{F(X_t\beta_2)}{1+F(X_t\beta_2)} \end{pmatrix}$$

### E.1.2 Adaptation de l'estimation de la matrice de transition

#### E.1.2.1 Rappel dans le cas non-contraint

Si aucune contrainte n'existe sur $p_{ij}$, la dérivée partielle de la log-vraisemblance $\mathscr{L}$ s'exprime simplement :

$$\frac{\partial \mathscr{L}}{\partial p_{ij}} = \frac{S_{ij}^{(2)}}{p_{ij}}$$

où $S^{(2)}$ est la matrice de somme des probabilités lissées consécutives d'ordre deux (cf eq. (**??**)). La log-vraisemblance introduit une contrainte (car une matrice de transition est stochastique, et donc ses colonnes somment à l'unité [1]), finalement la dérivée du lagrangien produit :

$$\frac{S_{ij}^{(2)}}{p_{ij}} = \lambda$$

de sorte que la conclusion suit : $p_{ij} \propto S_{ij}^{(2)}$ et donc $P = S^{(2)} \oslash \left(S^{(2)} J_M\right)$. Les $p_{ij}$ ainsi obtenus sont bien $\geq 0$.

1. On n'introduit pas de contraintes sur la positivité des coefficients, car c'est inutile : après la résolution, la matrice a bien tous ses coefficients positifs.

#### E.1.2.2 Cas d'une union de processus markoviens indépendants

On traite ici de l'adapatation de la procédure d'estimation de la matrice de transition pour le cas où celle-ci est modélisée conformément à la section D.4.2.2. L'adaptation est rapide. Etudions la dérivée de la log-vraisemblance par rapport au terme $(i,j)$ de la matrice $P^k$ (il s'agit de la $k^{\text{ième}}$ matrice de transition et non de la puissance $k$ de la matrice globale, bien sûr). On commence pour cela par étudier $\partial P/\partial p_{ij}^k$ : c'est une matrice composée de 0 sauf pour les lignes $L_i$ et les colonnes $C_j$ :

$$\frac{\partial p_{i'j'}}{\partial p_{ij}^k} = \frac{p_{i'j'}}{p_{ij}^k} \mathbb{1}_{L_i \times C_k}(i',j')$$

Par conséquent :

$$\frac{\partial \mathscr{L}}{\partial p_{ij}^k} = \sum_{(i',j') \in L_i \times C_k} \frac{\partial LV}{\partial p_{i'j'}} \frac{\partial p_{i'j'}}{\partial p_{ij}^k} = \sum_{(i',j') \in L_i \times C_k} \frac{S_{i'j'}^{(2)}}{p_{i'j'}} \frac{p_{i'j'}}{p_{ij}^k} = \frac{1}{p_{ij}^k} \sum_{(i',j') \in L_i \times C_k} S_{i'j'}^{(2)}$$

Soit $\tilde{S}_k^{(2)}$ la matrice (carrée d'ordre $M_k$) de terme général : $(\tilde{S}_k^{(2)})_{ij} = \sum_{(i',j') \in L_i \times C_k} S_{i'j'}^{(2)}$, alors : $\frac{\partial \mathscr{L}}{\partial P^k} = \frac{\tilde{S}_k^{(2)}}{P^k}$.
Procédant comme précédemment en introduisant la contrainte $P^k \mathbb{1} = \mathbb{1}$, on en déduit que : $P^k = \tilde{S}_k^{(2)} \oslash \left( \tilde{S}_k^{(2)} J_{M_k} \right)$.

#### E.1.2.3 Cas d'un processus markovien de rupture

On traite ici de l'adapatation de la procédure d'estimation de la matrice de transition pour le cas où celle-ci est modélisée conformément à la section D.4.1.3. Le raisonnement est direct :

$$\forall i \in [\![1; M-1]\!], \quad \frac{\partial \mathscr{L}}{\partial \alpha_i} = \sum_{j=i}^{i+1} \frac{\partial LV}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial \alpha_i} = \frac{S_{ii}^{(2)}}{\alpha_i} - \frac{S_{i,i+1}^{(2)}}{1-\alpha_i}$$

de sorte que l'annulation de cette dérivée partielle conduit à :

$$\alpha_i = \frac{S_{ii}^{(2)}}{S_{ii}^{(2)} + S_{i,i+1}^{(2)}} \in [0;1]$$

ce qui se comprend aisément.

## E.2 Modélisation avancée de la variance

### E.2.1 Modélisations envisagées

On étudie ici différentes spécifications de la matrice de variance. Les cas principaux sont systématiquement distingués en deux, suivant la dépendance au régime :

- Cas 1 : forme générale
  - Cas 1.1 : régime dépendant $\forall m, \qquad \Sigma_m = \left(\sigma_{mij}^2\right)$
  - Cas 1.2 : non régime dépendant $\forall m, \qquad \Sigma_m = \left(\sigma_{ij}^2\right)$
- Cas 2 : hétéroscédastique non corrélé, régime dépendant ou non
  - Cas 2.1 : hétéroscédastique non corrélé régime dépendant :

$$\Sigma_m = \begin{pmatrix} \sigma_{1m}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{nm}^2 \end{pmatrix} = \text{diag}(\sigma_{im}^2)_i$$

  - Cas 2.2 : hétéroscédastique non corrélé non régime dépendant : $\Sigma_m \text{diag}(\sigma_i^2)_i$
- Cas 3 : homoscédastique non corrélé, régime dépendant ou non
  - Cas 3.1 : homoscédastique non corrélé régime dépendant : $\Sigma_m = \sigma_m^2 \text{Id}_n$
  - Cas 3.2 : homoscédastique non corrélé non régime dépendant : $\Sigma_m = \sigma^2 \text{Id}_n$

### E.2.2 Adaptation de la phase M de l'estimation EM

Pour chacun de ces cas, on étudie la minimisation de :

$$\sum_m T_m \ln|\Sigma_m| + \sum_{m,t} S_{mt} X'_{mt} \Sigma_m^{-1} X_{mt} \tag{E.1}$$

où $X \in \mathbb{R}^n$, tandis que $T_m, S_{mt} \in \mathbb{R}$.

**Préliminaire** La formule univariée suivante :

$$\forall a, b, \in \mathbb{R}^+, \quad \operatorname*{argmin}_x \; a \ln x + \frac{b}{x} = \frac{b}{a}$$

se généralise aisément au cas multivarié [2].

**Cas 1.1**

$(E.1)$ devient $\sum_m T_m \ln|\Sigma_m| + \sum_{m,t} S_{mt} X'_{mt} \Sigma_m^{-1} X_{mt}$

$$\implies \Sigma_m = \frac{1}{T_m} \sum_t S_{mt} X_{mt} X'_{mt}$$

**Cas 1.2**

$(E.1)$ s'écrit $\ln|\Sigma| \sum_m T_m + \sum_{m,t} S_{mt} X'_{mt} \Sigma^{-1} X_{mt}$

$$\implies \Sigma = \frac{\sum_{m,t} S_{mt} X_{mt} X'_{mt}}{\sum_m T_m}$$

**Cas 2.1**

$(E.1)$ s'écrit $\sum_m T_m \left( \sum_{i=1}^n \ln \sigma_{im}^2 \right) + \sum_{m,t} S_{mt} \sum_i \frac{X_{mti}^2}{\sigma_{im}^2}$

$$\implies \hat{\sigma}_{im}^2 = \frac{\sum_t S_{mt} X_{mti}^2}{T_m}$$

**Cas 2.2**

$(E.1)$ s'écrit $\sum_m T_m \left( \sum_i \ln \sigma_i^2 \right) + \sum_{m,t} S_{mt} \sum_i \frac{X_{mti}^2}{\sigma_i^2}$

$$\implies \hat{\sigma}_i^2 = \frac{\sum_{m,t} S_{mt} X_{mti}^2}{\sum_m T_m}$$

**Cas 3.1**

$(E.1)$ s'écrit $\sum_m T_m n \ln \sigma_m^2 + \sum_{m,t} \frac{S_{mt}}{\sigma_m^2} ||X_{mt}||^2$ (où $||X||^2 = \sum_{i=1}^n X_i^2$)

$$\implies \hat{\sigma}_m^2 = \frac{\sum_t S_{mt} ||X_{mt}||^2}{n T_m}$$

2. Il suffit d'écrire la forme quadratique avec la trace.

**Cas 3.2**

$(E.1)$ s'écrit $\sum_m T_m n \ln \sigma^2 + \frac{1}{\sigma^2} \sum_{m,t} S_{mt} ||X_{mt}||^2$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{n} \frac{\sum_{m,t} S_{mt} ||X_{mt}||^2}{\sum_m T_m}$$