UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESTIMATION DE VRAISEMBLANCE MAXIMALE ET VARIANCE ASYMPTOTIQUE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

AMIROUCHE MESROUR

AOÛT 2020

REMERCIEMENTS

Merci à mes chers parents pour leurs sacrifices et encouragements qui sont pour moi les piliers fondateurs de ce que je suis et de ce que je fais, ainsi que mes frères et ma sœur pour le soutien indéfectible.

Merci au professeur qui m'a enseigné mon premier cours à l'UQAM, Mr. René Ferland, qui m'a encadré tout au long de ce mémoire et qui m'a fait partager ses brillantes idées. Qu'il soit aussi remercié pour sa relecture attentive de mon texte, sa disponibilité permanente, sa gentillesse, et surtout pour tout ses conseils et discussions enrichissantes.

Merci à mes amis et collègues pour les discussions de mathématiques souvent éclairantes.

Merci aux professeurs du département de mathématiques de l'UQAM qui m'ont offert une éducation de qualité.

Merci à ma chère copine pour son soutien quotidien et son enthousiasme contagieux à l'égard de mes travaux comme de la vie en général.

Merci à mes collègues et supérieurs professionnels qui m'ont mis dans des conditions favorables pour achever ce travail.

Tanmirt – nwen s umata

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'inférence par vraisemblance maximale est utilisée dans de vastes champs applicatifs. Cette méthode ainsi que les propriétés asymptotiques de ses estimations sont un domaine de recherche toujours actif un siècle après les fondations posées par Sir. Ronald Fisher en 1912. Après avoir rappelé les fondamentaux bibliographiques des propriétés asymptotiques des estimateurs par ML (Maximum likelihood) et souligné les difficultés pratiques pour exprimer et évaluer les estimateurs de la variance asymptotique, ce mémoire présente les travaux réalisés pour deux estimateurs de remplacement de la variance découlant du score partiel empirique et de l'information observée. Les travaux engagés visent à démontrer et confirmer les propriétés asymptotiques des MLE (Maximum likelihood estimation) : convergence non biaisée et normalité de la distribution dans trois contextes d'hypothèses d'application. Avec une gradation de complexité, nous commençons par un modèle de variables indépendantes et parentes, continuons par un modèle de chaîne de Markov et finalement terminons avec un modèle de Markov cachée. Nous présentons pour chacun de ces processus aléatoires les résultats de simulation afin de confronter nos résultats théoriques à l'expérience in silico.

Mots clés : maximum de vraisemblance, asymptotique, convergence, variance, estimateur, MLE, chaîne de Markov, chaîne de Markov cachée.

INTRODUCTION

De nombreuses techniques statistiques ont été développées pour aider les scientifiques à comprendre les données qu'ils collectent. Ces techniques sont généralement classées comme descriptives ou déductives. Alors que les statistiques descriptives permettent aux scientifiques de résumer rapidement les principales caractéristiques d'un ensemble de données, les statistiques inférentielles vont plus loin en aidant les scientifiques à découvrir des modèles ou des relations dans un ensemble de données ou à porter des jugements sur les données.

Dans la plupart des situations concrètes où l'application des statistiques inférentielles est nécessaire, il est rare que les données de la population concernée par l'étude soient connues pour toute une série de raisons (exemple : il est impossible de tester la température de tous les contaminés de la covid-19 afin d'estimer la température moyenne des contaminés) et il est souvent nécessaire d'estimer un paramètre caractérisant la population à partir des données observées (échantillon).

Une des méthodes d'estimation les plus utilisées par les scientifiques est la méthode du maximum de vraisemblance. L'origine de la méthode vient de (Fisher, 1912) qui est considéré par certains comme le plus grand des successeurs de Darwin et par d'autres comme le fondateur de la statistique moderne. La méthode est connue pour sa précision dans l'estimation, car un paramètre estimé n'a aucune valeur si la précision de l'estimation réalisée n'est pas connue. Ceci peut être réalisé, soit en calculant sa variance, soit en déterminant autour de la valeur estimée, un intervalle dont on a de bonnes raisons de croire qu'il contient la vraie valeur du paramètre recherché (un intervalle de confiance).

Dans un jargon un peu plus mathématique, la précision de la méthode de vraisemblance maximale est connue sous le nom de la convergence vers la vraie valeur, à ceci s'ajoute une deuxième propriété très importante qui est la normalité de la distribution des estimations. Ces propriétés ont été prouvées dans le cas des observations qui proviennent des populations indépendantes quand la méthode a fait son apparition dans les années 1920. Plus tard, les propriétés de la méthode ont aussi été prouvées pour des processus un peu plus complexes comme les chaînes de Markov par (Billingsley, 1961) et les modèles avec chaînes de Markov cachées par (Baum et Petrie, 1966).

La variance des estimations de vraisemblance maximale est définie comme l'inverse de l'information de Fisher qui est elle-même définie comme l'espérance de la dérivée seconde par rapport aux paramètres de la vraisemblance conjointe des observations. Dans certaines situations, surtout dans des processus un peu plus complexes, cette information est parfois compliquée, voire impossible à calculer (exemple : l'espérance de la fonction conjointe n'est pas calculable pour la majorité des processus stochastiques).

Pour répondre à cette difficulté de calcul de l'information de Fisher, (Efron et Hinkley, 1978) ont proposé comme estimateur de remplacement, soit *la deuxième dérivée de la vraisemblance conjointe sur les paramètres*, qu'ils ont appelés *l'information observée*. Depuis, cette méthode est utilisée pour le calcul de la variance des estimateurs, même si sa convergence n'a pas été profondément étudiée.

Dans ce travail, nous essayons de montrer mathématiquement la convergence de l'information observée vers l'information de Fisher théorique, et de confirmer par simulation cette convergence. Pour répondre à cet objectif, nous avons besoin de détailler les démonstrations des propriétés des estimations de vraisemblance maximale. Dans un ordre chronologique de complexité, ceci sera fait pour trois

types d'observations qui seront présentés dans trois chapitres. Le premier est le cas des données provenant des populations indépendantes et parentes (de même loi), le deuxième est le cas d'une chaîne de Markov et finalement un modèle avec chaîne de Markov cachée.

Pour chaque cas, nous présentons d'abord des généralités sur le modèle statistique et l'estimateur de vraisemblance maximale. Nous étudions ensuite la convergence et la normalité asymptotique, du point de vue théorique, en suivant si possible la littérature d'origine pour le modèle considéré. Nous terminons chaque chapitre par une (petite) étude de simulation qui illustre les propriétés présentées et éclaire l'usage de l'estimateur. On en profite aussi pour expliquer comment calculer l'estimateur et surmonter les problèmes numériques afférents à ce calcul.

CHAPITRE I

VARIABLES INDÉPENDANTES ET PARENTES

1.1 Le principe de vraisemblance maximale

L'idée fondamentale de l'estimation par maximum de vraisemblance (ML) est, comme le nom l'indique, de trouver un ensemble d'estimations des paramètres, appelé $\hat{\theta}$, telles que la probabilité d'observer l'échantillon obtenu (sa vraisemblance) soit maximale. Ainsi la densité de probabilité conjointe pour le modèle que l'on estime est évaluée aux valeurs observées des variables aléatoires et traitée comme une fonction de paramètres du modèle. Le vecteur $\hat{\theta}$ des estimations ML donne alors le maximum de cette fonction. Ce principe d'estimation est très largement applicable : si nous pouvons écrire la densité conjointe de l'échantillon, nous pouvons en principe utiliser le maximum de vraisemblance, soumis bien sûr à certaines conditions de régularité que nous détaillerons dans les prochaines sections. Par ailleurs, l'estimateur ML a un nombre de propriétés commodes, dont nous discuterons plus en détail dans le reste de ce chapitre. Il possède également quelques propriétés peu pratiques, et pour cela, le praticien doit parfois être méfiant.

La manière la plus simple de saisir l'idée fondamentale de l'estimation par ML est de l'illustrer sur un exemple. Supposons que chaque observation x_t est générée par la densité

$$f(x_t;\theta) = \theta e^{-\theta x_t}, \quad x_t > 0, \quad \theta > 0,$$

et qu'elle est indépendante de toutes les autres x_t. Il s'agit de la densité de la distribution exponentielle. Il y a un seul paramètre inconnu θ que nous désirons estimer, et nous disposons de n observations avec lesquelles nous allons travailler. La densité conjointe des x_t sera appelée la *fonction de vraisemblance* et notée $L(x;\theta)$; pour toute autre valeur de θ, cette fonction nous renseigne sur la probabilité que nous aurions d'observer l'échantillon $x = (x_1, \ldots, x_n)$.

Comme les x_t sont indépendants, leur densité conjointe est simplement le produit des densités marginales. Ainsi, la fonction de vraisemblance s'écrit

$$L(x;\theta) = \prod_{t=1}^{n} \theta e^{-\theta x_t}. \tag{1.1}$$

Dans le cas d'échantillons de grande taille, (1.1) peut devenir extrêmement petite, et prendre des valeurs qui sont bien au-delà des possibilités des nombres à virgule flottante que les ordinateurs manipulent. À ceci on peut aussi ajouter la simplicité du calcul. Pour ces raisons, parmi d'autres, il est d'usage de maximiser le logarithme de la fonction vraisemblance plutôt que la fonction de vraisemblance elle-même. Bien évidemment, nous obtiendrons la même réponse en procédant ainsi, car la fonction de *log-vraisemblance* $\ell(x;\theta) = \log\left(L(x;\theta)\right)$ est une fonction monotone croissante de $L(x;\theta)$; si $\hat{\theta}$ maximise $\ell(x;\theta)$, il doit aussi maximiser $L(x;\theta)$. Dans le cas de (1.1), la fonction de log-vraisemblance est

$$\ell(x;\theta) = \sum_{t=1}^{n} \log(\theta) - \theta x_t = n\log(\theta) - \theta \sum_{t=1}^{n} x_t \,. \tag{1.2}$$

La maximisation de la fonction de log-vraisemblance, par rapport au seul paramètre inconnu θ, est une procédure directe. Dériver l'expression (1.2) par rapport à θ et poser la dérivée égale à zéro donne la condition du premier ordre

$$\frac{n}{\theta} - \sum_{t=1}^{n} x_t = 0 \tag{1.3}$$

et nous trouvons, par résolution, que l'estimateur ML $\hat{\theta}$ est

$$\hat{\theta} = \frac{n}{\sum_{t=1}^{n} x_t} \cdot$$

Dans ce cas, il n'est pas nécessaire de se soucier d'éventuelles solutions multiples de (1.3). La dérivée seconde de (1.2) est toujours négative, ce qui nous permet de conclure que $\hat{\theta}$ est l'unique estimateur ML. Notons que cela ne sera pas toujours le cas. Pour certains problèmes, les conditions du premier ordre peuvent mener à des solutions multiples.

Dès à présent, nous pourrions à juste titre poser certaines questions relatives aux propriétés de $\hat{\theta}$. Est-ce dans tous les sens du terme un bon estimateur à utiliser ? Est-il biaisé ? Est-il convergent ? Comment est-il distribué ?

Deux propriétés attrayantes majeures des estimateurs ML sont la convergence et la normalité asymptotique. Celles-ci seront longuement étudiées dans les prochaines sections.

1.2 Généralités et notations

L'estimation ML repose sur la notion de *vraisemblance* d'un ensemble donné d'observations relatives à un modèle (ou une variable aléatoire provenant d'une distribution de probabilité). Nous développons maintenant la notation à partir de laquelle nous pouvons exprimer une telle caractérisation qui est particulièrement utile pour nos objectifs. Nous supposons que chaque observation x_t pour tout échantillon de taille n est une réalisation d'une variable aléatoire X_t, $t = 1, \ldots, n$, prenant des valeurs dans $\mathbb{R}^m$. Il est plus commode de laisser la notation vectorielle x désigner l'échantillon entier

$$x = [x_1, x_2, \ldots, x_n]$$

si chaque observation est un scalaire, x est un vecteur de dimension n, tandis que si chaque observation est un vecteur de dimension m, x est une matrice de dimension $n \times m$. Le vecteur ou la matrice x peut posséder une densité de probabilité, c'est-à-dire une densité conjointe.

Nous pouvons à présent définir formellement la fonction de vraisemblance associée à un modèle donné pour un échantillon x donné. Cette fonction dépend des paramètres du modèle et de l'ensemble d'observations donné par x ; sa valeur correspond exactement à la loi associée au vecteur (ou matrice) aléatoire X caractérisée par le vecteur paramétrique $\theta \in \Theta$, évaluée au point d'échantillon x. L'ensemble Θ désigne ici l'espace paramétrique dans lequel θ prend ses valeurs ; nous supposerons que c'est un sous-ensemble de $\mathbb{R}^d$. Nous désignerons la fonction de vraisemblance par $L : \mathcal{X} \times \Theta \to \mathbb{R}$ et sa valeur pour θ et x par $L(x;\theta)$ ($\mathcal{X}$ est l'ensemble des observations des x). Dans bien des cas pratiques, tel que celui examiné à la section précédente, les observations x_t sont indépendantes et chaque x_t a une densité de probabilité $L_t(x_t;\theta)$. La fonction de vraisemblance dans ce cas est alors

$$L(x;\theta) = \prod_{t=1}^{n} L_t(x_t;\theta). \tag{1.4}$$

La fonction de vraisemblance (1.1) de la section précédente est évidemment un cas particulier de ce cas présent. Quand chacune des variables aléatoires X_t est identiquement distribuée selon une densité $f(x_t;\theta)$, nous dirons que les variables sont *parentes*.

Même quand la fonction de vraisemblance ne peut pas s'écrire sous la forme (1.4), il est souvent possible de factoriser $L(x;\theta)$ en une série de contributions, chacune provenant d'une ou plusieurs observations. Prenons à titre d'exemple des observations individuelles x_t, $t = 1, \ldots, n$ ordonnées chronologiquement comme dans les séries temporelles ou les chaînes de Markov (nous allons explorer ce dernier cas dans le prochain chapitre).

Comme nous l'indiquions dans la section précédente, on utilise dans la pratique la fonction log-vraisemblance $\ell(x;\theta)$ plutôt que la fonction de vraisemblance $L(x;\theta)$. La décomposition de $\ell(x;\theta)$ en contributions provenant d'observations résulte

de (1.4). Elle peut être écrite comme suit

$$\ell(x;\theta) = \sum_{t=1}^{n} \ell_t(x_t;\theta) \tag{1.5}$$

où $\ell_t(x_t;\theta) = \log\left(L_t(x_t;\theta)\right)$.

Nous sommes à présent en position de donner la définition de l'estimation par maximum de vraisemblance. Nous disons que $\hat{\theta}_n \in \Theta$ est une estimation par maximum de vraisemblance, une estimation ML, ou le MLE, pour les données x si

$$\ell(x;\hat{\theta}_n) \geq \ell(x;\theta), \quad \forall\, \theta \in \Theta\,.$$

Si l'inégalité est stricte, alors $\hat{\theta}$ est l'unique MLE.

Le MLE peut ne pas exister en général (exemple : le cas d'une distribution uniforme définie sur un ensemble d'observations qui dépend du paramètre), à moins que la fonction de log-vraisemblance ne soit continue par rapport aux paramètres θ et que l'ensemble Θ ne soit *compact* (c'est-à-dire fermé et borné). C'est pourquoi il est d'usage, dans les traitements formels de l'estimation par maximum de vraisemblance, de supposer que Θ est compact. Nous ne désirons pas formuler cette hypothèse, parce qu'elle s'accorde très mal avec la pratique, pour laquelle une estimation est valable partout dans $\mathbb{R}^d$. Mais cela signifie que nous devrons vivre avec la possible non-existence du MLE.

Il est aussi souvent commode d'utiliser une autre définition du MLE, qui n'est pas équivalente en général. Si la fonction de vraisemblance atteint un maximum intérieur à l'espace paramétrique, alors elle, ou de façon équivalente la fonction de log-vraisemblance, doit satisfaire les conditions du premier ordre pour un maximum. Ainsi, le MLE peut se définir comme une solution aux équations de vraisemblance, qui correspond précisément aux conditions du premier ordre suivantes :

$$S(x;\hat{\theta}_n) = 0 \tag{1.6}$$

où le vecteur score, ou vecteur gradient, $S \in \mathbb{R}^d$ est défini par

$$S(x;\theta) = \nabla_\theta \ell(x;\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ell(x;\theta) \\ \frac{\partial}{\partial\theta_2}\ell(x;\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_d}\ell(x;\theta) \end{bmatrix}. \tag{1.7}$$

Alors S est un vecteur colonne des dérivées partielles de la fonction log-vraisemblance ℓ par rapport aux paramètres θ.

Comme il peut arriver que plus d'une valeur de θ satisfasse les équations de vraisemblance (1.6), la définition nécessite par ailleurs que l'estimation $\hat{\theta}_n$ soit associée à un maximum local de ℓ et que

$$\ell(x;\hat{\theta}_n) \geq \ell(x;\theta^*)$$

pour n'importe quelle autre solution θ^* des équations de vraisemblance.

Dans le cadre de notre étude et surtout lors de l'étude des propriétés asymptotiques, nous allons considérer seulement le cas où la solution de la fonction (1.6) est unique.

Nous concluons cette section par une variété de définitions qui seront utilisées dans le reste du chapitre et plus généralement dans le reste du mémoire. En utilisant la décomposition (1.5) de la fonction de log-vraisemblance $\ell(x;\theta)$, nous pouvons définir une matrice $G(x;\theta)$ de dimension $n \times d$ dont l'élément type est

$$G_t(x;\theta) = \frac{\partial \ell_t(x;\theta)}{\partial \theta_i}$$

Nous appellerons $G(x;\theta)$ la matrice du gradient. Cette matrice est intimement reliée au vecteur gradient S.

La matrice hessienne associée à la fonction de log-vraisemblance $\ell(x;\theta)$ est la

matrice $H(x;\theta)$ de dimension $d \times d$ dont l'élément type est

$$H_{ij}(x;\theta) = \frac{\partial^2 \ell_t(x_t;\theta)}{\partial\theta_i \partial\theta_j}.$$

Nous définissons l'information contenue dans l'observation t par $I_t(\theta)$, la matrice de dimension $d \times d$ dont l'élément type est

$$\left(\mathcal{I}_t(\theta)\right)_{ij} = \mathbb{E}_\theta \left[n^{-1} G_{ti}(\theta) G_{tj}(\theta) \right]. \tag{1.8}$$

La matrice $\mathcal{I}_t(\theta)$ est symétrique, en général semi-définie positive. Celle-ci est connue sous le nom de matrice d'information de Fisher que nous détaillerons dans les prochaines sections. La matrice moyenne pour un échantillon de taille n est définie par

$$\mathcal{I}(\theta) = \frac{1}{n} \sum_{t=1}^{n} \mathcal{I}_t(\theta) = n^{-1} \mathcal{I}^n(\theta). \tag{1.9}$$

La matrice $\mathcal{I}_t(\theta)$ mesure la quantité espérée d'information contenue dans l'observation x_t et $\mathcal{I}^n(\theta) = n\mathcal{I}(\theta)$ mesure la quantité espérée d'information contenue dans l'échantillon entier.

Nous avons toutes les définitions nécessaires pour commencer l'étude des deux propriétés attrayantes de l'estimation ML. Dans les deux prochaines sections, nous détaillerons la convergence et la normalité asymptotique du MLE.

1.3 La convergence

Une des raisons pour lesquelles l'estimation par maximum de vraisemblance est largement utilisée est que les estimateurs ML sont, sous des conditions assez générales, convergents. Dans cette section, nous expliquons pourquoi c'est le cas.

On retrouve deux sortes de convergence de l'estimateur ML vers la vraie valeur, la faible et la forte. Notre curiosité s'est révélée pour la convergence forte de ce dernier en suivant le texte de (Ferguson, 1996).

Définition 1.1. *Soit une séquence d'estimateurs* $\hat{\theta}_n = \theta_n(X_1, X_2, \ldots, X_n)$ *et un paramètre* θ *dans* Θ*. On dit que* $\hat{\theta}_n$ *converge presque sûrement (*$\mathcal{P}.\mathcal{S}$*) vers* θ *si*

$$P_\theta \left\{ \lim_{n \to +\infty} \hat{\theta}_n = \theta \right\} = 1 \,.$$

La convergence asymptotique concerne une séquence d'estimateurs. Cette dernière est définie dans le contexte suivant : pour une variable aléatoire X d'une distribution $f(x, \theta)$ observée n fois, la séquence $\hat{\theta}_n = \theta(X_1, X_2, \ldots, X_n)$ est construite en utilisant la même fonction avec des tailles d'échantillons différentes.

Les données sont générées par un cas particulier du modèle qui sont des observations x provenant de la distribution paramétrique avec $\theta = \theta^0$. Ce θ^0 est la vraie valeur avec laquelle les données sont générées. Dans la pratique, cette valeur n'est pas connue et on cherche à l'estimer.

Nous commençons maintenant à présenter les résultats dont nous aurons besoin, le premier est une loi forte des grands nombres uniforme dans le cas d'une fonction paramétrique. Ce résultat est défini dans le contexte suivant : soit $X_1, X_2, X_3, \ldots$ une suite de variables aléatoires indépendantes qui proviennent d'une distribution $f(x; \theta)$ et $U(x; \theta)$ est une fonction continue en x telle que $U(X_j; \theta)$ est d'espérance finie pour tout θ. Dans ces conditions, on sait déjà que

$$\frac{1}{n} \sum_{j=1}^{n} U(X_j; \theta) \;\hookrightarrow^{\mathcal{P}.\mathcal{S}}\; \mathbb{E}\left[U(X; \theta)\right] = \mu(\theta) \tag{1.10}$$

où X est une variable générique de loi f. Le résultat (1.10) découle de la loi forte des grands nombres pour la suite $U(X_j; \theta)$, $j = 1, 2, \ldots$ car celles-ci restent des variables aléatoires pour un paramètre $\theta \in \Theta$ fixe. Pour la suite, il sera nécessaire de renforcer le résultat pour le rendre uniforme en θ :

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{j=1}^{n} U(x_j; \theta) - \mu(\theta) \right| \hookrightarrow^{\mathcal{P}.\mathcal{S}} 0 \tag{1.11}$$

quand n tend vers l'infini. Sous cette forme, le résultat est dû à (LeCam, 1960) et, bien sûr, il est obtenu sous certaines conditions pour Θ, U et f. Nous le démontrons en appendice.

Le deuxième résultat fait intervenir *l'information de Kullback-Leibler*, une mesure de divergence entre deux distributions de probabilité.

Définition 1.2. *Soient $l_{\theta_0}(x)$ et $l_{\theta_1}(x)$ deux densités paramétriques continues, l'information de Kullback-Leibler est définie par*

$$K(l_{\theta_0}, l_{\theta_1}) = \mathbb{E}_{\theta_0}\left[\log \frac{l_{\theta_0}(X)}{l_{\theta_1}(X)}\right] = \int \log \frac{l_{\theta_0}(x)}{l_{\theta_1}(x)}\, l_{\theta_0}(x)\, dx.$$

Nous pouvons montrer que l'information de Kullback-Leibler est non négative en utilisant l'inégalité de Jensen (Nous le démontrons en Appendice A). Le résultat est obtenu comme suit. On observe d'abord que

$$\mathbb{E}_{\theta_0}\left[\frac{l_{\theta_1}(X)}{l_{\theta_0}(X)}\right] = \int \frac{l_{\theta_1}(x)}{l_{\theta_0}(x)}\, l_{\theta_0}(x) = \int l_{\theta_1}(x) \leq 1\,.$$

Cependant, par la convexité de la fonction logarithme, il vient que

$$K(l_{\theta_0}, l_{\theta_1}) = \mathbb{E}_{\theta_0}\left[-\log \frac{l_{\theta_1}(X)}{l_{\theta_0}(X)}\right] \geq -\log \mathbb{E}_{\theta_0}\left[\frac{l_{\theta_1}(X)}{l_{\theta_0}(X)}\right] \geq 0 \tag{1.12}$$

avec égalité si seulement si $P_{\theta_0}\left\{l_{\theta_0}(X) = l_{\theta_1}(X)\right\} = 1$. L'inégalité (1.12) est connue sous le nom de l'inégalité de Shannon-Kolmogorov.

Une condition de régularité importante qui doit être satisfaite afin qu'un estimateur ML soit convergent : le modèle doit être identifiable. Par définition, nous dirons qu'un modèle paramétrique est identifiable si la condition ci-dessous est vérifiée :

$$\theta = \theta^0 \quad \text{si et seulement si} \quad L(x;\theta) = L(x;\theta^0)\,.$$

Alors, si le modèle est identifiable, la valeur de $\hat{\theta}_n$ est l'unique solution de (1.6).

Nous pouvons finalement énoncer le théorème suivant, que l'on doit à (Wald, 1949).

Théorème 1.1. Soit x un échantillon de n réalisations de la variable X d'un modèle qui est caractérisé par la loi $f(x;\theta)$, $\theta \in \Theta$. On suppose que θ^0 est la vraie valeur du paramètre. Supposons de plus que :

1. Θ est compact ;
2. $f(x;\theta)$ est semi-continue supérieurement en θ pour tout x ;
3. il existe une fonction $K(x)$ tel que $\mathbb{E}_{\theta^0}[K(X)] < +\infty$ et

$$U(x,\theta) = \log f(x;\theta) - \log f(x;\theta^0) \leq K(x) \text{ pour tout } x \text{ et } \theta$$

4. $\forall\, \theta \in \Theta$ et $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} f(x;\theta')$ est continue en x.
5. $f(x;\theta) = f(x;\theta^0) \implies \theta = \theta^0$ (identifiabilité).

Dans ces conditions, pour toute suite $\{\hat{\theta}_n\}$ d'estimateurs du maximum de vraisemblance de θ, on a

$$\hat{\theta}_n \hookrightarrow^{\mathcal{P}.\mathcal{S}} \theta^0 .$$

Démonstration. La preuve de ce théorème se base sur l'information de Kullback-Leibler, la loi forte des grands nombres uniforme et le lemme de Fatou-Lebesgue qui sont énoncés en appendice. Nous savons que $\hat{\theta}_n$ est celui qui maximise $\ell(x;\theta)$, ce dernier maximise aussi,

$$\begin{aligned} \ell(x;\theta) - \ell(x;\theta^0) &= \sum_{t=1}^{n} \log \frac{f(x_t;\theta)}{f(x_t;\theta^0)} \\ &= \sum_{t=1}^{n} U(x_t;\theta). \end{aligned} \tag{1.13}$$

Cette première écriture (1.13) nous permet de faire un lien avec l'information de Kullback-Leibler. En outre, d'après (1.10) on peut écrire,

$$\frac{1}{n}\sum_{t=1}^{n} \log \frac{f(x_t;\theta)}{f(x_t;\theta^0)} \hookrightarrow^{\mathcal{P}.\mathcal{S}} \mathbb{E}_{\theta^0}\left[\log \frac{f(X;\theta)}{f(X;\theta^0)}\right].$$

Ainsi, en appliquant la définition de l'information de Kullback-Leibler et l'inégalité de Shannon-Kolmogorov (1.12), on obtient

$$\frac{1}{n}\sum_{t=1}^{n}\log\frac{f(x_t;\theta)}{f(x_t;\theta^0)} \;\hookrightarrow^{\mathcal{P}.\mathcal{S}}\; -k(\theta^0,\theta)=\mu(\theta)<0$$

La fonction $\mu(\theta)$ est semi-continue supérieurement. En effet, on a

$$\limsup_{\theta'\to\theta}\mu(\theta')=\limsup_{\theta'\to\theta}\mathbb{E}\left[U(X;\theta')\right]$$

et comme $U(x;\theta)$ est majorée par la fonction $K(x)$ d'après la condition 3, alors, d'après le lemme de Fatou-Lebesgue, on a

$$\limsup_{\theta'\to\theta}\mathbb{E}\left[U(X;\theta')\right]\leq\mathbb{E}\left[\limsup_{\theta'\to\theta}U(X;\theta')\right]\leq\mathbb{E}\left[U(X;\theta)\right]\leq\mu(\theta).$$

D'où

$$\forall\,\theta'\in\Theta,\;\exists\,\theta\in\Theta,\quad\limsup_{\theta'\to\theta}\mu(\theta')\leq\mu(\theta)\,. \tag{1.14}$$

Jusqu'à présent, nous avons montré que $U(x;\theta)$ converge en moyenne vers une fonction $\mu(\theta)$ définie négative. Maintenant, pour commencer le raisonnement par l'absurde, supposons $\rho>0$ tel que $S=\{\theta:|\theta-\theta_0|\geq\rho\}$ un ensemble compact. Les conditions de la loi forte des grands nombres uniforme sont vérifiées pour $U(x;\theta)$ avec $\theta\in S$, on a donc

$$P_{\theta^0}\left\{\limsup_{n\to+\infty}\sup_{\theta\in S}\frac{1}{n}\sum_{1}^{n}U(X_j;\theta)\leq\sup_{\theta\in S}\mu(\theta)\right\}=1. \tag{1.15}$$

Par (1.14), la fonction $\mu(\theta)$ est semi-continue supérieurement sur l'ensemble compact S. Alors, on peut déduire que $\mu(\theta)$ atteint son maximum sur S. Pour cela, on pose $\delta=\sup_{\theta\in\Theta}\mu(\theta)<0$, le résultat (1.15) devient

$$P_{\theta^0}\left\{\limsup_{n\to+\infty}\sup_{\theta\in S}\frac{1}{n}\sum_{1}^{n}U_\theta(X;\theta)\leq\;\delta\right\}=1.$$

Donc, avec probabilité 1,

$$\exists\;N\in\mathbb{N},\;\forall\,n>N,\quad\sup_{\theta\in S}\sum_{t=1}^{n}U(X_t;\theta')<\delta/2<0\,.$$

Mais nous savons que, pour le maximum de vraisemblance $\hat{\theta}_n$,

$$\frac{1}{n}\sum_{t=1}^{n} U(X_t;\hat{\theta}_n) = \sup_{\theta\in\Theta}\frac{1}{n}\sum_{t=1}^{n} U(X_t;\theta) = 0\,.$$

On peut conclure alors que $\hat{\theta}_n \notin S$ pour $n > N$. Donc, $\hat{\theta}_n \in S^c$ et pour N assez grand, on a $P_{\theta^0}\Big\{|\hat{\theta}_n - \theta^0| < \rho,\ \forall\, n > N\Big\} \to 1$. □

Il existe deux ensembles majeurs de circonstances dans lesquelles les estimations ML peuvent ne pas être convergentes. Le premier survient quand le nombre de paramètres n'est pas fixe, mais augmente avec n. Cette possibilité n'est pas considérée dans le théorème précédent où θ est indépendant de n. Mais il n'est pas surprenant que cela engendre des problèmes, car si le nombre de paramètres n'est pas fixe, il est loin d'être évident que la quantité d'information que l'échantillon nous donne à propos de chacun d'eux augmentera suffisamment rapidement lorsque n tend vers l'infini. Dans le deuxième, on retrouve ceux dans lesquels le modèle n'est pas identifiable. De tels problèmes sont bien au-delà des objectifs de notre étude ; on pourra consulter entre autres (Neyman et Scott, 1948), (J. Kiefer, 1956) et (Kalbfleisch et Sprott, 1970).

1.4 La distribution asymptotique

Les résultats simples concernant la distribution asymptotique des estimations ML sont obtenus dans le contexte du modèle de vraisemblance maximale. Par conséquent, nous limiterons notre attention au cas ou $\hat{\theta}_n$ est une solution du problème (1.6). Il est alors relativement simple de montrer que $\hat{\theta}_n$ possède la propriété de normalité asymptotique. Pour un modèle de vraisemblance caractérisé par θ^0, le vecteur des estimations paramétriques $\hat{\theta}_n$ tend vers la limite non stochastique θ^0. Cependant, si nous multiplions la différence $\hat{\theta}_n - \theta^0$ par $n^{1/2}$, la quantité résultante $n^{1/2}(\hat{\theta}_n - \theta^0)$ aura une limite en loi qui est une variable aléatoire avec une distri-

bution normale multivariée (ou univarié pour $d = 1$). L'écriture mathématique de ce résultat est donnée par

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \hookrightarrow^{\mathcal{L}} \mathcal{N}_d\left(0, \mathcal{I}(\theta^0)^{-1}\right) \tag{1.16}$$

où $\mathcal{I}(\theta^0)$ est l'information de Fisher introduite par (1.8) et $\mathcal{L}$: fait référence à la convergence ne lois. Elle s'écrit dans le cas d'un modèle de vraisemblance avec une distribution caractérisée par une densité comme

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2 \qquad d \times d\,. \tag{1.17}$$

Maintenant, nous énonçons un théorème qui précise les hypothèses sur le modèle d'estimation de vraisemblance maximale pour avoir le résultat (1.16). La condition 2 du théorème de normalité asymptotique énonce que l'ordre de différentiation et d'intégration sont interchangeables. Ce résultat est valide sous une variété de conditions de régularité, parmi lesquelles la plus simple est que le domaine de définition de f soit indépendant de θ ; nous allons faire appel aussi à une version continue de la loi forte des grands nombres uniforme (consulter l'appendice A).

Théorème 1.2. Soit x une séquence d'observations de n variables aléatoires X, et θ^0 la vraie valeur du paramètre. Si

1. Θ est un ouvert de $\mathbb{R}^d$;
2. $f(x;\theta)$ a les propriétés suivantes :
   - deux fois continûment dérivable sur Θ ;
   - continue sur l'ensemble des observations x ;
   - la deuxième dérivée partielle par rapport à θ peut passer sous le signe de l'intégrale $\int f(x;\theta)dx$.
3. la deuxième dérivée par rapport à θ de $\log f(x;\theta)$ est majorée au voisinage de θ^0 par une fonction $K(x)$, tel que $\mathbb{E}\left[K(X)\right] < \infty$;

4. $I(\theta^0)$ est définie positive ;
5. Identifiablitié : si $f(x;\theta) = f(x;\theta^0) \implies \theta = \theta_0$.

Alors, il existe une séquence d'estimateurs $\hat{\theta}_n$ qui converge presque sûrement vers θ^0 et telle que

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \hookrightarrow^{\mathcal{L}} \mathcal{N}\left(0, \mathcal{I}(\theta^0)^{-1}\right) .$$

Démonstration. La preuve est divisée en deux parties. Dans la première, nous allons essayer de satisfaire les conditions du théorème 1.1 pour garantir l'existence d'un MLE convergent.

1. *Existence d'un MLE convergent*

Fixons $\rho > 0$ et soit $S_\rho = \{\theta : |\theta - \theta^0| \leq \rho\}$. On a :

a. S_ρ est un ensemble compact ;
b. $f(x;\theta)$ est continue en x d'après la condition 2 ;
c. la fonction $U(x;\theta)$ définie par $\log f(x;\theta) - \log f(x;\theta^0)$ (condition 3 du théorème 1.1) est délimitée par une fonction continue et finie sur x. En effet, pour le voir, nous devons écrire $U(x;\theta)$ en utilisant le développement de Taylor au voisinage de θ^0 à l'ordre 1. Ce développement s'écrit

$$U(x;\theta) = U(x;\theta^0) + G(x;\theta^0)^T(\theta - \theta^0) + (\theta - \theta^0)^T R(x;\theta)(\theta - \theta^0). \quad (1.18)$$

Observons que pour les nouvelles composantes de $U(x;\theta)$, nous avons :

i. $U(x;\theta^0)$ est nulle (par définition de U) ;
ii. $G(x;\theta^0)$ est une fonction dérivable et finie d'après les conditions 2 et 3 ;
iii. il nous reste à montrer que $R(x;\theta)$ est finie. Une manière d'y arriver et d'écrire $R(x,\theta)$ en fonction de sa deuxième dérivée $H(x;\theta)$ avec θ dans

S_ρ vu que cette dernière est finie d'après la condition 3. Ceci nécessite de développer le reste $R(x,\theta)$ de (1.18) par l'intégrale de Laplace (voir l'appendice A) :

$$\begin{aligned} R(x;\theta) &= \int_0^1 (1-t)H(x;\theta^0 + t(\theta-\theta^0))dt \\ &= \int_0^1 \left(\int_t^1 d\lambda\right) H(x;\theta^0 + t(\theta-\theta^0))dt \\ &= \int_0^1 \left(\int_0^\lambda H(x;\theta^0 + t(\theta-\theta^0))dt\right) d\lambda\,. \end{aligned}$$

Le passage de la deuxième à la troisième égalité est fait en utilisant le théorème de Fubini pour inverser les bornes de l'intégration sous l'hypothèse $H(x;\theta)$ est intégrable.

Maintenant, en faisant le changement de variable $t = \lambda\mu$, on obtient

$$R(x;\theta) = \int_0^1 \left(\int_0^1 \lambda H(x;\theta^0 + \lambda\mu(\theta-\theta^0))d\mu\right) d\lambda\,.$$

Comme $R(x;\theta)$ est une intégrale finie d'une fonction $H(x;\theta)$ qui est majorée uniformément sur S_ρ. On conclut que celle-ci est une fonction finie. De cette manière, d'après i, ii et iii, on peut conclure que $U(x;\theta)$ est délimitée par une fonction finie.

d. la continuité de $\sup_{|\theta'-\theta|<\rho} f(x;\theta')$ se déduit directement de la continuité de $f(x;\theta)$ sur S_ρ d'après la condition 2 ;

e. $f(x;\theta) = f(x;\theta^0) \implies \theta = \theta^0$ d'après la condition 5.

Toutes les conditions du théorème 1.1 sont vérifiées, l'existence d'un MLE $\hat{\theta}_n$ convergent vers θ^0 est démontrée. Nous pouvons maintenant continuer avec la démonstration de la normalité asymptotique de $\hat{\theta}_n$.

2. *Normalité asymptotique*

Afin de simplifier la preuve, cette partie sera divisée en trois sous-parties.

1. Commençons par étudier la variation de la log-vraisemblance. Un premier réflexe est de regarder la variation de sa dérivée au voisinage du vrai paramètre. Pour commencer, écrivons le développement de Taylor de $\frac{\partial}{\partial\theta} l(x;\theta) = \ell'(x;\theta)$ au voisinage de θ^0 à l'ordre 0 :

$$\begin{aligned}\ell'(x;\theta) &= \ell'(x;\theta^0) + (\theta-\theta^0)R(x;\theta)\\ &= \ell'(x;\theta^0) + \Big(\int_0^1 \sum_{t=1}^{n} G(x_t;\theta^0+\lambda(\theta-\theta^0))d\lambda\Big)(\theta-\theta^0)\,. \qquad (1.19)\end{aligned}$$

Le passage de la première égalité à la deuxième est fait en utilisant le développement par l'intégrale de Laplace. Comme $\hat{\theta}_n$ est une racine de la fonction du score, alors (1.19) devient

$$0 = \ell'(x;\theta^0) + \Big(\int_0^1 \sum_{t=1}^{n} G(x_t;\theta^0+\lambda(\hat{\theta}_n-\theta^0))d\lambda\Big)(\hat{\theta}_n-\theta^0)$$

et, en divisant par $\sqrt{n}$, il vient

$$\frac{1}{\sqrt{n}}\ell'(x;\theta^0) = \Big(-\int_0^1 \frac{1}{\sqrt{n}}\sum_{t=1}^{n} G(x_t;\theta^0+\lambda(\hat{\theta}_n-\theta^0))d\lambda\Big)(\hat{\theta}_n-\theta^0)\,.$$

Posons

$$A_n = -\int_0^1 \frac{1}{n}\sum_{t=1}^{n} G(x_t;\theta^0+\lambda(\hat{\theta}_n-\theta^0))d\lambda\,.$$

L'équation précédente s'écrit maintenant sous la forme

$$\frac{1}{\sqrt{n}}\,\ell'(x;\theta^0) = A_n\sqrt{n}\,\big(\,\hat{\theta}_n-\theta^0\,\big)\,. \qquad (1.20)$$

Notre but est d'utiliser le théorème central limite sur le résultat (1.20) pour montrer la normalité asymptotique de $\ell'(x,\theta_0)$ et, pour y arriver, nous allons commencer par montrer la convergence presque sûre de A_n vers $\mathcal{I}(\theta^0)$.

2. Pour montrer la convergence presque sûre de A_n vers l'information de Fisher, nous aurons besoin de satisfaire les conditions de la version continue de la loi forte des grands nombres uniforme. Nous avons

(a) S_ρ est compact (fermé et borné) ;

(b) $H(x;\theta)$ est continue en x pour chaque $\theta \in S_\rho$ (condition 2) ;

(c) $H(x;\theta)$ est borné supérieurement par $K(x)$ avec $\mathbb{E}\left[K(x)\right] < \infty$ (condition 3).

Donc, d'après la version continue de la loi forte des grands nombres uniforme (voir appendice A), nous avons

$$P\left\{\lim_{n\to+\infty}\sup_{\theta\in S_\rho}\left|\frac{1}{n}\sum_{t=1}^{n}H(x_t;\theta)-\mathbb{E}_{\theta^0}\left[H(X_t;\theta)\right]\right|=0\right\}=1\,.$$

En d'autres mots, presque sûrement, quand le nombre d'observations n devient très grand, on a

$$\sup_{\theta\in S_\rho}\left|\frac{1}{n}\sum_{t=1}^{n}H(x_t;\theta)-\mathbb{E}_{\theta^0}\left[H(x_t;\theta)\right]\right|\leq\epsilon\,. \tag{1.21}$$

Notre objectif est d'en dire autant pour $|A_n - I(\theta^0)|$, ce qui suit facilement :

$$\begin{aligned}
|A_n - I(\theta^0)| &\leq \int_0^1 \left|\frac{1}{n}\sum_{t=1}^{n}H(x_t;\theta^0+\lambda(\hat{\theta}_n-\theta^0))+I(\theta^0)\right| d\lambda \\
&\leq \int_0^1 \sup_{\theta\in S_\rho}\left[\left|\frac{1}{n}\sum_{t=1}^{n}H(x_t;\theta)-\mathbb{E}_{\theta^0}\left[H(X;\theta)\right]\right| \right. \\
&\qquad\qquad \left. + \left|\mathbb{E}_{\theta^0}\left[H(X;\theta)\right]+I(\theta^0)\right|\right] d\lambda \\
&\leq \int_0^1 \left[2\epsilon\right]\, d\lambda \\
&\leq 2\epsilon\,.
\end{aligned}$$

Nous justifions le passage de la première inégalité à la deuxième par l'hypothèse de l'unicité du MLE. En effet, $\ell(x;\theta)$ est concave au voisinage de θ^0 et de plus $\ell(x;\theta^0) \geq \ell(x;\hat{\theta}_n)$. D'après le résultat d'existence $\hat{\theta}_n \in S_\rho$ et pour $\lambda \in [0,1]$, nous avons

$$\ell(x;\theta^0+\lambda(\hat{\theta}_n-\theta^0)) \geq \sup_{\theta\in S_\rho}\ell(x;\theta)$$

ceci est justifié par le faite que $S_\rho = \{\theta : |\theta - \theta_0| \leq \rho\}$ et pour N très grand, $n > N \implies |\hat{\theta} - \theta^0| < \rho$.

Comme la fonction $\ell(x;\theta)$ est concave (fonction logarithmique sur un domaine positif) sur S_ρ alors

$$\frac{\partial^2}{\partial\theta^2}\ell(x;\theta^0 + \lambda(\hat{\theta}_n - \theta^0)) \leq \frac{\partial^2}{\partial\theta^2}\sup_{\theta\in S_\rho}\ell(x;\theta),$$

ce qui est équivalent à

$$\sum_{t=1}^{n} H(x_t;\theta^0 + \lambda(\hat{\theta}_n - \theta^0)) \leq \sup_{\theta\in S_\rho}\sum_{t=1}^{n} H(x_t;\theta)\,.$$

Le résultat de concavité de la fonction $l(x;\theta)$ est obtenue par le faite que la fonction $f(x;\theta)$ est semi-continue supérieurement et que l'ensemble des paramètres Θ est *compact*. Le passage de la deuxième à la troisième inégalité découle de la proposition 1.3 de la prochaine section. Ceci achève notre démonstration de convergence presque sûre de A_n vers $\mathcal{I}(\theta^0)$.

3. Démontrons finalement la normalité asymptotique du MLE. Pour commencer, reprenons le résultat (1.20) :

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} G(X_t;\theta^0)\right) = A_n\sqrt{n}(\hat{\theta}_n - \theta^0)\,.$$

De plus, nous avons, $\mathbb{E}_{\theta^0}\left(G(X;\theta^0)\right) = 0$ et $\mathrm{Var}_{\theta^0}\left(G(x;\theta^0)\right) = \mathcal{I}(\theta^0)$ d'après (1.8). Donc, le théorème central limite (voir appendice A) nous donne le résultat suivant

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} G(X_t;\theta^0)\right) \hookrightarrow^{\mathcal{L}} N_d(0,\mathcal{I}(\theta^0)).$$

À ceci s'ajoute le résultat déjà prouvé $A_n \hookrightarrow^{\mathcal{P}.\mathcal{S}} \mathcal{I}(\theta^0)$. Sachant que $\mathcal{I}(\theta^0)$ est définie positive (condition 4), cela nous permet d'écrire

$$\sqrt{n}(\hat{\theta}_n - \theta^0) = A_n^{-1}\frac{1}{\sqrt{n}}\ell'(x;\theta^0)\,.$$

Pour finir, nous faisons appel au théorème de Slutsky qui nous donne :

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \hookrightarrow^{\mathcal{L}} N_d(0, \mathcal{I}(\theta^0)^{-1}) \,.$$

□

Par normalité asymptotique, nous signifions que la suite de variables aléatoires $n^{1/2}(\hat{\theta}_n - \theta^0)$ admet une limite en loi qui est celle d'une variable aléatoire de l'ordre de l'unité, normalement distribuée d'espérance nulle et de matrice de covariance qui est l'inverse de (1.17).

1.5 Information de Fisher

Dans cette section, nous examinerons plus en détails la matrice d'information que nous avons abordée brièvement dans les deux sections précédentes. Cette matrice est d'une importance substantielle pour le calcul de la covariance asymptotique du MLE.

L'information de Fisher est donnée par la matrice de variance-covariance de la fonction du score (1.6) sous la loi $L(x;\theta)$

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[S(X;\theta) S(X;\theta)' \right] - \mathbb{E}_\theta \left[S(X;\theta) \right] \mathbb{E}_\theta \left[S(X;\theta) \right]' .$$

Cette expression se simplifie car $E_\theta \left[S(X;\theta) \right] = 0$. En effet, on sait que

$$\int L(x;\theta)\, dx = 1. \tag{1.22}$$

La fonction (1.22) étant constante, son gradient est nul :

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int L(x;\theta)\,dx \\
&= \int \frac{\partial}{\partial\theta}\,L(x;\theta)\,dx, \quad \text{justifié par la C2 théorème 1.2} \\
&= \int \left(\frac{1}{L(x;\theta)}\frac{\partial}{\partial\theta}\,L(x;\theta)\right) L(x;\theta)\,dx \\
&= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log L(X;\theta)\right] \\
&= \mathbb{E}_\theta\left[S(X;\theta)\right]\,. \qquad (1.23)
\end{aligned}$$

Donc, nous avons

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[S(X;\theta)S(X;\theta)'\right]. \qquad (1.24)$$

On peut établir une autre écriture de l'information de Fisher.

Proposition 1.3. Sous la condition 2 du théorème 1.2, la matrice d'information est aussi égale à

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\,S(X;\theta)\right] = -\mathbb{E}_\theta\left[H(X;\theta)\right]\,.$$

Démonstration. On remarque que l'on peut écrire

$$\frac{\partial}{\partial\theta}S(x;\theta) = \frac{\partial}{\partial\theta}\frac{L'(x;\theta)}{L(x;\theta)} = \frac{L''(x;\theta)}{L(x;\theta)} - \frac{L'(x;\theta)^2}{L(x;\theta)^2} = \frac{L''(x;\theta)}{L(x;\theta)} - S(x;\theta)^2\,.$$

Ainsi, on a

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}S(X;\theta)\right] = \mathbb{E}_\theta\left[\frac{L''(x;\theta)}{L(x;\theta)}\right] - \mathcal{I}(\theta)\,.$$

En observant que

$$\mathbb{E}_\theta\left[\frac{L''(x;\theta)}{L(x;\theta)}\right] = \int L''(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int L(x;\theta)dx = 0,$$

on obtient le résultat annoncé. □

Il est souvent nécessaire de calculer la matrice de covariance des estimateurs. Celle-ci nécessite la formule de $\mathcal{I}(\theta^0)$, pas toujours facile à trouver et qui dépend d'un paramètre inconnu. Il y a deux estimateurs de remplacement.

1.5.1 Estimateur basé sur le score partiel empirique

Il s'inspire de l'écriture de $\mathcal{I}(\theta)$ en termes des scores partiels :

$$\begin{aligned}
\mathcal{I}(\theta)_{ij} &= \mathbb{E}_\theta \left[S(X;\theta)_i S(X;\theta)_j \right] \\
&= \int_x S(x;\theta)_i S(x;\theta)_j L(x;\theta)\, dx \\
&= \int_{x_1} \cdots \int_{x_n} \frac{\partial \log L}{\partial \theta_i}(x;\theta) \frac{\partial \log L}{\partial \theta_j}(x;\theta) \prod_{k=1}^{n} f(x_k;\theta)\, dx \\
&= \int_{x_1} \cdots \int_{x_n} \left(\sum_{k=1}^{n} \frac{\partial \log f}{\partial \theta_i}(x_k;\theta) \right) \left(\sum_{\ell=1}^{n} \frac{\partial \log f}{\partial \theta_j}(x_\ell;\theta) \right) \prod_{k=1}^{n} f(x_k;\theta)\, dx \\
&= \sum_{k=1}^{n} \sum_{\ell=1}^{n} \int_{x_1} \cdots \int_{x_n} \frac{\partial \log f}{\partial \theta_i}(x_k;\theta) \frac{\partial \log f}{\partial \theta_j}(x_\ell;\theta) \prod_{k=1}^{n} f(x_k;\theta)\, dx_1 \cdots dx_n
\end{aligned}$$

$$\begin{aligned}
&= \sum_{m=1}^{n} \int_{x_m} \frac{\partial \log f}{\partial \theta_i}(x_m;\theta) \frac{\partial \log f}{\partial \theta_j}(x_m;\theta) f(x_m;\theta)\, dx_m \\
&\qquad + \sum_{k \neq \ell} \int_{x_k} \int_{x_\ell} \frac{\partial \log f}{\partial \theta_i}(x_k;\theta) \frac{\partial \log f}{\partial \theta_j}(x_l;\theta) f(x_k;\theta) f(x_\ell;\theta)\, dx_k dx_\ell \\
&= \sum_{m=1}^{n} \mathbb{E}_\theta \left[s(X_m;\theta)_i s(X_m;\theta)_j \right] + \sum_{k \neq l}^{n} \mathbb{E}_\theta \left[s(X_k;\theta)_i \right] \mathbb{E}_\theta \left[s(X_l;\theta)_j \right]
\end{aligned}$$

avec $s(x;\theta)$ le score partiel :

$$s(x;\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log f(x;\theta) \\ \frac{\partial}{\partial \theta_2} \log f(x;\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d} \log f(x;\theta) \end{bmatrix}.$$

Comme pour le score, on a $\mathbb{E}_\theta\left[s(X_m;\theta)_k\right]=0$. Donc, on peut écrire

$$\begin{aligned}\mathcal{I}(\theta)_{ij} &= \sum_{m=1}^{n}\mathbb{E}_\theta\left[s(X_m;\theta)_i s(X_m;\theta)_j\right]\\ &= n\,\mathbb{E}_\theta\left[s(X_1;\theta)_i s(X_1;\theta)_j\right]\\ &= n\,\mathrm{Cov}_\theta\left(s(X_1;\theta)_i, s(X_1;\theta)_j\right).\end{aligned}$$

On obtient ensuite un estimateur de $\mathcal{I}(\theta^0)_{ij}$ en remplaçant la covariance théorique par la covariance dans l'échantillon, et en évaluant le score partiel en $\hat{\theta}_n$:

$$\begin{aligned}\hat{I}(\theta^0)_{ij} &= n\left[\left(\frac{1}{n}\sum_{m=1}^{n}s(x_m;\hat{\theta}_n)_i s(x_m;\hat{\theta}_n)_j\right)-\left(\frac{1}{n}\sum_{k=1}^{n}s(x_k;\hat{\theta}_n)_i\right)\left(\frac{1}{n}\sum_{\ell=1}^{n}s(x_\ell;\hat{\theta}_n)_j\right)\right]\\ &= \sum_{m=1}^{n}s(x_m;\hat{\theta}_n)_i s(x_m;\hat{\theta}_n)_j-\frac{1}{n}S(x;\hat{\theta}_n)_i S(x;\hat{\theta}_n)_j.\end{aligned}$$

Puisque $S(x;\hat{\theta}_n)=0$, il vient que :

$$\hat{I}(\theta^0)_{ij}=\sum_{m=1}^{n}s(x_m;\hat{\theta}_n)_i s(x_m;\hat{\theta}_n)_j.$$

Si on note

$$\frac{\partial\ell_k}{\partial\theta}=\begin{bmatrix}\frac{\partial}{\partial\theta_1}\log f(x_k;\theta)\\ \frac{\partial}{\partial\theta_2}\log f(x_k;\theta)\\ \vdots\\ \frac{\partial}{\partial\theta_d}\log f(x_k;\theta)\end{bmatrix}$$

alors l'estimateur de la matrice $\mathcal{I}(\theta^0)$ est

$$\hat{I}(\theta^0)=\sum_{k=1}^{n}\frac{\partial\ell_k}{\partial\theta}|_{\hat{\theta}_n}\frac{\partial\ell_k}{\partial\theta}^{t}|_{\hat{\theta}_n}. \tag{1.25}$$

Cet estimateur est appelé *l'information de Fisher empirique* et on le notera $\mathcal{I}_e(\hat{\theta}_n)$.

1.5.2 Estimateur basé sur l'information observée

Ce deuxième estimateur de l'information de Fisher repose sur une autre formule. En effet, nous avons montré dans la proposition 1.3 que, sous certaines hypothèses

de régularité, l'information (1.24) peut aussi s'écrire comme l'espérance de l'opposé de la matrice hessienne de la log-vraisemblance :

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log L(X; \theta) \right]. \tag{1.26}$$

L'information de Fisher observée est l'estimateur obtenu en ignorant l'espérance $\mathbb{E}_\theta$ de (1.26) et en évaluant $\mathcal{I}(\theta)$ sur $\hat{\theta}_n$ et les observations x :

$$\mathcal{I}_o(\hat{\theta}_n) = -\frac{\partial^2}{\partial \theta^2} \log L(x; \hat{\theta}_n). \tag{1.27}$$

Il n'y a aucune raison, en général, pour que les deux estimateurs, $\mathcal{I}_e(\hat{\theta}_n)$ et $\mathcal{I}_o(\hat{\theta}_n)$ soient identiques. Leurs performances seront comparées lors de notre étude de simulation.

La discussion précédente n'a peut-être pas rendu clair un point qui est de la plus haute importance quand on essaie de pratiquer les inférences concernant un ensemble d'estimation ML $\hat{\theta}_n$. Tout ce qui se rattache à la théorie de la distribution asymptotique se note en termes de $n^{1/2}(\hat{\theta}_n - \theta^0)$, mais en pratique nous voulons en fait utiliser $\hat{\theta}_n$ pour réaliser des inférences à propos de θ^0. Ceci signifie que nous devons baser nos inférences non pas sur des quantités qui estiment $\mathcal{I}(\theta^0)$ mais plutôt sur des quantités qui estiment $n\mathcal{I}(\theta^0)$. Alors c'est pour cette raison que nous n'avons pas divisé les deux estimateurs $\mathcal{I}_e(\hat{\theta}_n)$ et $\mathcal{I}_o(\hat{\theta}_n)$ par le nombre d'observations n.

1.6 Étude par simulation

Une caractéristique majeure indésirable du MLE est que ses propriétés avec des échantillons finis peuvent être très différentes des propriétés asymptotiques. Bien qu'elles soient convergentes, les estimations ML des paramètres sont probablement biaisées, et les estimations de la matrice de covariance ML peuvent être sérieusement trompeuses. Parce qu'en pratique les propriétés avec des échantillons finis

sont souvent inconnues, on doit décider comment se fier aux propriétés asymptotiques connues.

Les objectifs de cette étude de simulation sont multiples. Dans un premier temps, nous aimerions confirmer les résultats théoriques que nous avons prouvés et les illustrer par simulation dans le cas d'une variable aléatoire discrète. Par la suite, pour répondre au souci que nous avons évoqué au début de la section, nous voulons voir la grandeur et l'évolution du biais selon la taille de l'échantillon.

Pour garder la simplicité de notre étude, nous avons choisi d'étudier les paramètres des deux lois largement connues que sont les lois Gamma et de Poisson.

1.6.1 Convergence de l'estimateur

Nous avons simulé des échantillons de tailles petites, moyennes et très grandes $(50, 150, 300, 500, \ldots, 50000)$ qui proviennent des lois $\mathcal{G}(2, 5)$ et de $\mathcal{P}(4)$. Les estimations ML sont calculées avec ces échantillons et comparées aux vraies valeurs du paramètre.

La première simulation montre, dans les figures 1.1 et 1.2, que plus la taille de l'échantillon est grande plus les estimations ML sont proches des vraies valeurs. En d'autres mots, la propriété de convergence asymptotique est vérifiée pour les deux lois. Ceci confirme la généralisation que nous avons fait de cette propriété au cas d'une variable discrète.

Un deuxième phénomène à remarquer est qu'on voit bien que les estimations sont moins bonnes pour des grandeurs d'échantillon n petites. Ceci peut s'expliquer par un seul facteur qui est le manque d'information sur les paramètres dans l'échantillon.

1.6.2 Normalité et biais de l'estimateur

Dans cette deuxième simulation, pour chaque taille $n = (50, 150, 500, 1000)$, nous simulons 1000 réplications d'échantillon. Ces réplications serviront à calculer plusieurs statistiques. La première est la distribution d'échantillonage de $\hat{\theta}_n$. Les figures 1.3 à 1.8 donnent les distributions des estimateurs.

Le constat à tirer de toutes les figures est le suivant : la distribution des estimations ML des trois paramètres simulées avec $n = (500, 1000)$ semble très proche d'une cloche symétrique ce qui confirme l'hypothèse de normalité du MLE. Par contre, pour les petites tailles de l'échantillon (inférieur à 100), la deuxième hypothèse ne semble pas bien vérifiée. Ceci est sûrement dû à l'imprécision des estimateurs pour les petites tailles de l'échantillon. Cette imprécision est appelée le biais de l'estimateur qui sera l'objet de la prochaine simulation.

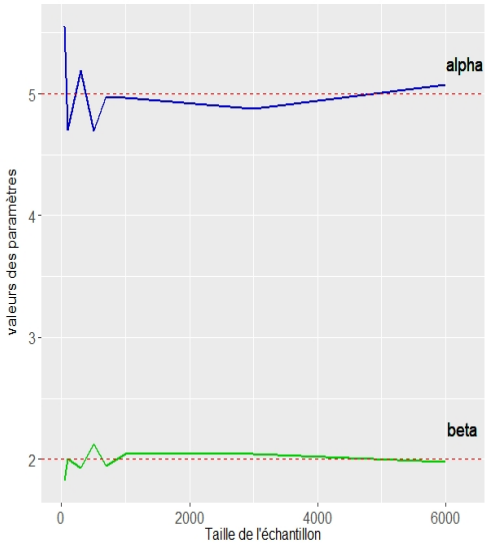

Figure (1.1) Convergence du MLE des paramètres de la loi Gamma selon la taille de l'échantillon.

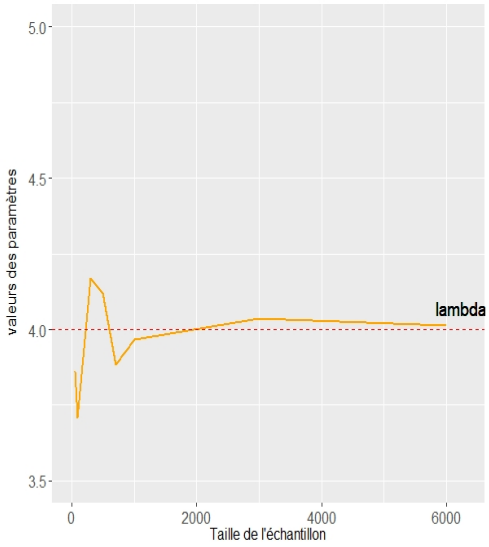

Figure (1.2) Convergence du MLE des paramètre de la loi de Poisson selon la taille de l'échantillon.

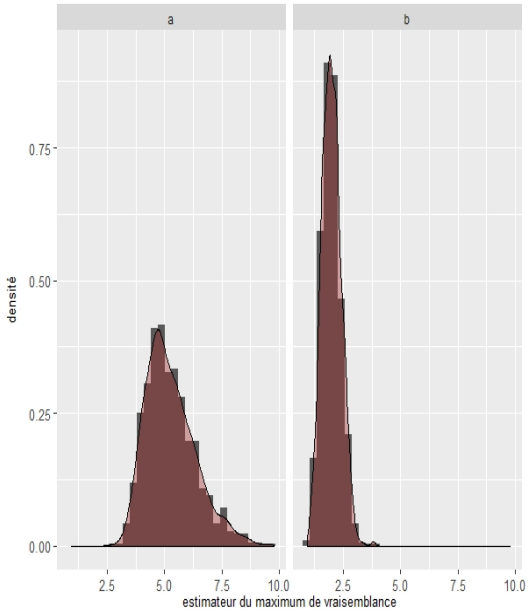

Figure (1.3) Distribution de 1000 MLE d'une loi $\mathcal{G}(5,2)$ $(n = 50)$.

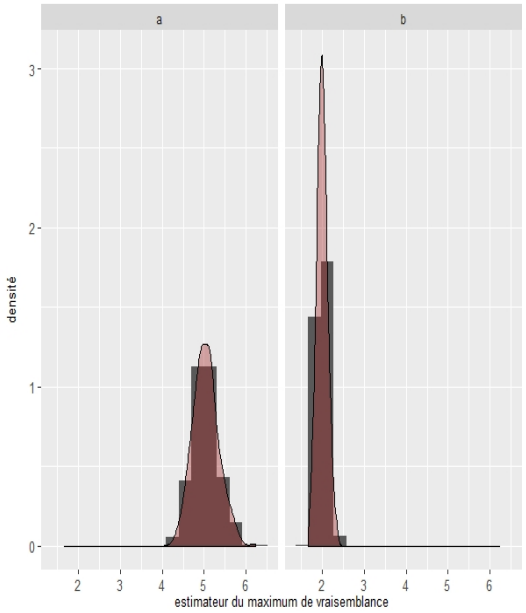

Figure (1.4) Distribution de 1000 MLE d'une loi $\mathcal{G}(5,2)$ $(n = 500)$.

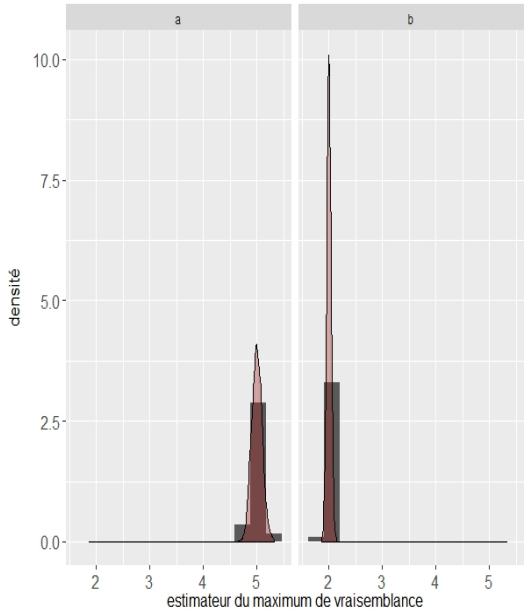

Figure (1.5) Distribution de 1000 MLE d'une loi $\mathcal{G}(5,2)$ $(n = 1000)$.

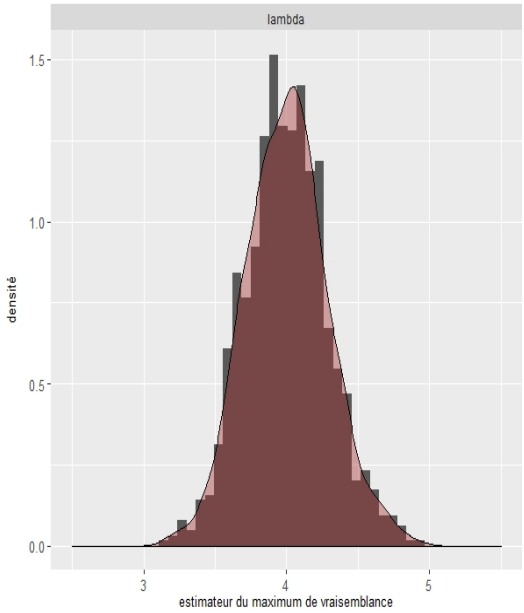

Figure (1.6) Distribution de 1000 MLE d'une loi $\mathcal{P}(4)$ ($n = 50$).

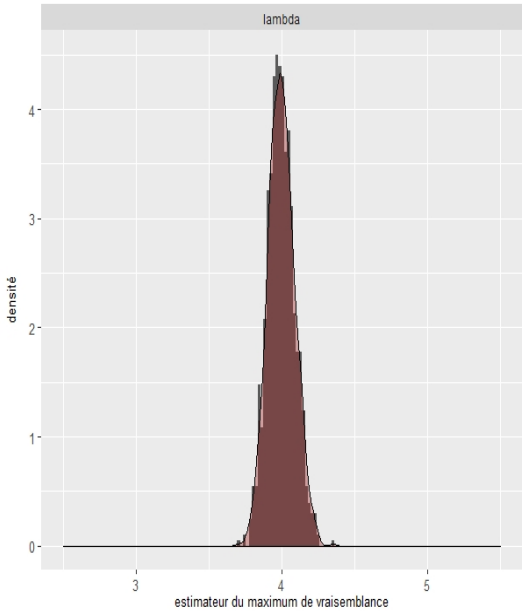

Figure (1.7) Distribution de 1000 MLE d'une loi $\mathcal{P}(4)$ ($n = 500$).

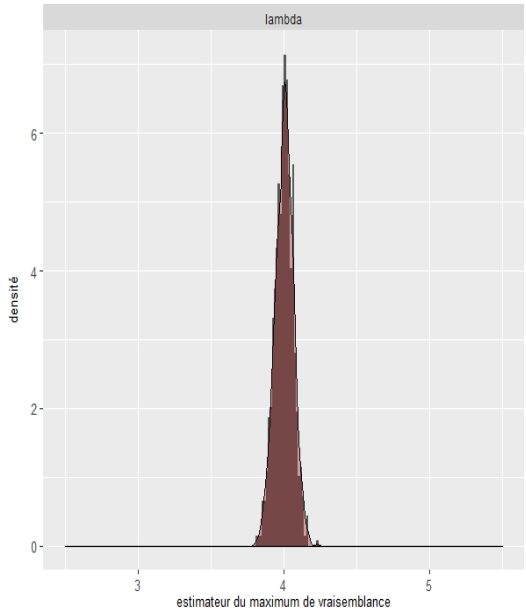

Figure (1.8) Distribution de 1000 MLE d'une loi $\mathcal{P}(4)$ ($n = 1000$).

Le biais des estimations ML est calculé en moyenne sur les 1000 réplications de chaque taille n simulées. Les résultats sont représentés dans le tableau 1.1.

N	$Biais(\hat{\alpha})$	$Biais(\hat{\beta})$	$Biais(\hat{\lambda})$
50	0.0641	0.0155	0.00609
150	0.0244	0.0074	0.00101
500	0.0083	0.0015	0.00092
1000	0.0007	0.0002	0.00016

Tableau (1.1) Biais moyen des estimateurs des lois Gamma et de Poisson simulées pour 1000 estimateurs calculés avec des tailles n différentes.

Effectivement, les chiffres du tableau 1.1 montrent que les estimations ML sont biaisées. Par contre, on remarque que le biais moyen n'est pas si élevé même pour les petites tailles d'échantillon. Il est de 0.064 pour le paramètre α estimé avec des tailles d'échantillon $n = 50$ et il tend vers 0 pour une taille n de plus en plus grande.

Un dernier point à remarquer dans les distributions des estimations ML est que la variance des MLE diminue avec la taille de l'échantillon. Cette hypothèse ainsi qu'une comparaison des deux méthodes d'estimation de la variance présentées dans la section précédente font l'objet de la dernière simulation.

1.6.3 Étude et comparaison des variances estimées

La variance asymptotique du MLE est définie comme l'inverse de l'information de Fisher (1.24) et nous avons mentionné dans la section précédente que celle-ci est une moyenne qui n'est pas toujours facile à calculer. Pour contourner ce problème, nous avons proposé deux estimateurs (1.25) et (1.27).

Pour avoir une idée sur ces deux méthodes d'estimation ainsi que sur la variance du MLE en général, nous avons calculé les variances estimées pour chacune des 1000 estimations simulées. Elles sont obtenues en inversant l'information de Fisher empirique (1.25) et l'information de Fisher observée (1.27) respectivement. Celles-ci seront comparées à la variance réelle qui est la variance du vecteur des 1000 MLE et la variance théorique (1.17) qui est calculée avec le vrai paramètre θ^0. Cette démarche est faite pour des échantillons de tailles $n = (50, 500, 1000)$ et les résultats sont résumés dans les tableaux 1.2, 1.3 et 1.4.

Paramètre	$\mathcal{I}_e(\hat{\theta}_n))^{-1}$	$\mathcal{I}_o(\hat{\theta}_n))^{-1}$	$\mathrm{Var}(\hat{\theta}_n)$	$\mathcal{I}(\theta^0))^{-1}$
α	7.4802	6.5735	4.1430	3.8712
β	0.0937	0.1361	0.1956	0.1523
λ	0.0629	0.0760	0.0844	0.0791

Tableau (1.2) Comparaison des variances théoriques, estimées et réelles pour les estimateurs des paramètres de $\mathcal{G}(5, 2)$ et $\mathcal{P}(4)$ simulés avec des échantillons de taille 50.

Paramètre	$\mathcal{I}_e(\hat{\theta}_n))^{-1}$	$\mathcal{I}_o(\hat{\theta}_n))^{-1}$	$\mathrm{Var}(\hat{\theta}_n)$	$\mathcal{I}(\theta^0))^{-1}$
α	0.31988	0.38621	0.39088	0.38712
β	0.01670	0.01647	0.01627	0.01628
λ	0.00833	0.00808	0.00843	0.00800

Tableau (1.3) Comparaison des variances théoriques, estimées et réelles pour les estimateurs des paramètres de $\mathcal{G}(5, 2)$ et $\mathcal{P}(4)$ simulés avec des échantillons de taille 500.

L'examen de ces tableaux nous permet de conclure trois points importants :

- les variances empiriques et réelles des paramètres estimés diminuent avec la taille de l'échantillon ;
- les variances estimées sont des estimateurs satisfaisants de la variance théorique. Les tableaux montrent que les valeurs de ces dernières sont proches

des variances théoriques. En d'autres mots, le biais s'approche de plus en plus de zéro avec la taille de l'échantillon qui augmente ;

- la méthode d'estimation de la variance par l'information observée est plus efficace de celle estimée par l'information empirique. On voit ça dans les tableaux par le fait que la première est plus proche des valeurs de la variance théorique dans toutes les simulations.

Paramètre	$\mathcal{I}_e(\hat{\theta}_n))^{-1}$	$\mathcal{I}_o(\hat{\theta}_n))^{-1}$	$\mathrm{Var}(\hat{\theta}_n)$	$\mathcal{I}(\theta^0))^{-1}$
α	0.20236	0.19969	0.19616	0.19356
β	0.00941	0.00826	0.00820	0.00814
λ	0.00419	0.00404	0.00415	0.00400

Tableau (1.4) Comparaison des variances théoriques, estimées et réelles pour les estimateurs des paramètres de $\mathcal{G}(5,2)$ et $\mathcal{P}(4)$ simulés avec des échantillons de taille 1000.

Cette partie de simulation sera achevée par le tableau des probabilités de couverture de chaque paramètre. Elles représentent le pourcentage d'appartenance de la vraie valeur du paramètre à l'intervalle de confiance qui est calculé avec les MLE et leurs variances théoriques. Pour chacun des 1000 estimateurs simulés, on calcule son intervalle de confiance et on vérifie si la vraie valeur du paramètre est dans cet intervalle. Les pourcentages sont représentés dans le tableau 1.5.

Taille	$P_c(\alpha^0)$	$P_c(\beta^0)$	$P_c(\lambda^0)$
50	0.931	0.939	0.947
500	0.951	0.950	0.951
1000	0.953	0.956	0.952

Tableau (1.5) Les probabilités de couverture des vrais paramètres des lois $\mathcal{G}(5,2)$ et $\mathcal{P}(4)$ calculés sur un échantillon de 1000 MLE.

Ce dernier tableau montre que les probabilités de couverture sont aux alentours de

95% pour les échantillons de taille moyenne et grande. Pour les petits échantillons, on voit que les proportions sont inférieures à 95% et ce fait peut être expliqué par le biais des estimations ML et des variances théoriques, ainsi que la normalité qui n'est pas parfaite.

Enfin, cette étude de simulation nous a permis de comprendre que le comportement global de l'estimateur du maximum de vraisemblance est très satisfaisant. Ainsi, le biais et la variance de l'estimateur sont relativement faibles et décroissent lorsque la taille de l'échantillon augmente. Nous avons aussi trouvé que les MLE sont approximativement gaussiens dès que n est suffisamment grand. Cela justifie la construction des intervalles de confiance pour les estimateurs en utilisant les quantiles de la loi $\mathcal{N}(0,1)$. Par ailleurs, les deux méthodes d'estimation de l'information de Fisher permettent généralement de fournir une bonne approximation de la variance des estimateurs quoique nous avons vu que l'information observée fournit de meilleurs résultats par rapport à l'information empirique. Suite à ce résultat, seulement l'information observée sera étudiée dans les prochains chapitres.

CHAPITRE II

CHAÎNES DE MARKOV

Dans le chapitre précédent, nous avons étudié le comportement de l'estimateur de vraisemblance maximale pour les variables indépendantes et de parentes. Nous voulons maintenant étudier ce même estimateur lorsque les variables cessent d'être indépendantes. En toute généralité, un tel objectif est trop ambitieux parce qu'on ne dispose pas d'une connaissance suffisante de la dépendance stochastique entre les variables.

Pour la suite, nous allons regarder deux cas, en ordre croissant de complexité : une chaîne de Markov avec un espace fini d'états et un modèle avec chaîne de Markov cachée. Le premier cas fait l'objet de ce chapitre, le second du prochain. La théorie asymptotique pour ces cas donne souvent lieu à des démonstrations techniques, avec des calculs longs et fastidieux. Pour ne pas alourdir la présentation, on cherchera principalement à explorer les résultats, plutôt que d'en faire une présentation rigoureuse.

2.1 Généralités

Nous suivons ici le texte de (Billingsley, 1961), un des premiers à traiter de la question. Soit donc $\{x_t,\, t \in \mathbb{N}\}$ une chaîne de Markov à temps discret, avec un espace d'états fini S et qui est gouvernée par une matrice de transition P_θ. Nous disposons

d'un nombre fini d'observations $(x_1, x_2, \ldots, x_{n+1})$ et nous voulons estimer θ.

Il faut d'abord spécifier la nature exacte du paramètre θ et son lien avec P_θ. On peut imaginer plusieurs situations, mais, dans ce mémoire, nous prendrons pour θ les probabilités de transition elles-mêmes. Par exemple, supposons que S n'a que deux états, c'est-à-dire $S = \{1, 2\}$, alors P_θ prendrait la forme :

$$P_\theta = \begin{bmatrix} 1 - \theta_{12} & \theta_{12} \\ \theta_{21} & 1 - \theta_{21} \end{bmatrix}$$

avec $\theta = (\theta_{12}, \theta_{21})$, un paramètre vectoriel, a priori dans $[0, 1]^2$. On devine comment généraliser à un nombre plus grand d'états, mais on doit se rappeler qu'il y a une contrainte : P_θ doit être une matrice stochastique, c'est-à-dire que chaque ligne doit être composée d'entrées non négatives qui somment à 1.

Que P_θ soit stochastique n'est pas suffisant non plus puisque, sous cette seule condition, nous pouvons obtenir une chaîne avec possiblement des états transitoires et/ou une ou plusieurs classes ergodiques. On va donc maintenant faire deux hypothèses :

H1) P_θ est stochastique ;

H2) chaque entrée $P_\theta(i, j)$ appartient à l'intervalle $]\delta; 1 - \delta[$ avec $\delta > 0$.

Du fait que les entrées sont positives, tous les états de la chaîne vont communiquer et, comme le nombre d'états est fini, il y aura une seule mesure de probabilité stationnaire π_θ. La chaîne sera ergodique.

Ces hypothèses peuvent sembler restrictives mais ne le sont pas vraiment. En effet, s'il y avait des états transitoires, alors, avec probabilité 1, ceux-ci ne seraient visités qu'un nombre fini de fois avant que la chaîne parvienne à une classe ergodique d'où elle ne sortirait plus. Puisque l'on s'intéresse ici au comportement asymptotique de l'estimateur de vraisemblance maximale, seules les observations de la classe

ergodique joueraient ultimement un rôle. Les hypothèses H1 et H2 nous placent dès le départ dans cette situation.

Par ailleurs, les résultats de (Billingsley, 1961) sont obtenus sous les conditions suivantes :

C1) l'ensemble D des couples (i, j) pour lesquels $P_\theta(i, j) > 0$ ne dépend pas de θ ;

C2) $P_\theta(i, j) > 0$ est trois fois continûment dérivable en θ ;

C3) la matrice des dérivées

$$\left[\frac{\partial P_\theta(i, j)}{\partial \theta}\right]$$

est de plein rang pour chaque θ ;

C4) il n'y a qu'une seule classe ergodique pour chaque θ.

Notre hypothèse H2 implique les conditions C1 et C4 puisque $\delta > 0$ et qu'il est le même pour tous. Quant à C2 et C3, elles sont une conséquence de la relation entre θ et P_θ. Par exemple, pour trois états, on a

$$P_\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{bmatrix}$$

La condition C2 est immédiate. La matrice en C3 contient une ligne pour chaque couple (i, j) et une colonne pour chaque θ_{ij}. Elle est donc 9×9 et se réduit à l'identité, clairement de plein rang. On aurait pu aussi choisir une autre écriture :

$$P_\theta = \begin{bmatrix} 1 - \theta_{12} - \theta_{13} & \theta_{12} & \theta_{13} \\ \theta_{21} & 1 - \theta_{21} - \theta_{23} & \theta_{23} \\ \theta_{31} & \theta_{32} & 1 - \theta_{31} - \theta_{32} \end{bmatrix}$$

et la matrice en C3 devient

$$\begin{bmatrix} -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

qui est aussi de plein rang.

La *vraisemblance* de l'échantillon $(x_1, x_2, \ldots, x_{n+1})$ s'écrit :

$$L(x_1, x_2, \ldots, x_{n+1}; \theta) = \pi_\theta(x_1) P_\theta(x_1, x_2) P_\theta(x_2, x_3) \cdots P_\theta(x_n, x_{n+1})$$

et la log-vraisemblance est

$$\log \pi_\theta(x_1) + \sum_{j=1}^{n} \log P_\theta(x_j, x_{j+1}) \,.$$

Sous les conditions ou hypothèses que nous avons énoncées, le terme initial devient négligeable face à l'autre lorsque $n \to \infty$. Puisqu'on s'intéresse ici au comportement asymptotique de l'estimateur de vraisemblance maximale, on prendra pour log-vraisemblance l'expression :

$$\ell(x_1, x_2, ..., x_{n+1}; \theta) = \sum_{j=1}^{n} \log P_\theta(x_j, x_{j+1}) \,. \tag{2.1}$$

L'estimateur de vraisemblance maximale $\hat{\theta}$ est, par définition la solution de l'équation :

$$\frac{\partial}{\partial \theta} \ell(x_1, x_2, \ldots, x_{n+1}; \theta) = 0 \tag{2.2}$$

sous la contrainte que P_θ est stochastique.

Cherchons l'estimateur dans le cas 2×2. La log-vraisemblance devient :

$$\ell(x_1, x_2, \ldots, x_{n+1}; \theta) = n_{11} \log(1-\theta_{12}) + n_{12} \log(\theta_{12}) + n_{21} \log(\theta_{21}) + n_{22} \log(1-\theta_{21})$$

où n_{ij} est le nombre de couples (x_k, x_{k+1}) égaux à (i, j). L'équation (2.2) s'écrit :

$$\begin{aligned} -\frac{n_{11}}{1-\theta_{12}} + \frac{n_{12}}{\theta_{12}} &= 0 \\ \frac{n_{21}}{\theta_{21}} - \frac{n_{22}}{1-\theta_{21}} &= 0 \end{aligned}$$

qui donne

$$\hat{\theta}_{12} = \frac{n_{12}}{n_{11} + n_{12}}, \qquad \hat{\theta}_{21} = \frac{n_{21}}{n_{21} + n_{22}}$$

L'estimateur est très naturel : la probabilité de transition θ_{ij} est estimée par sa fréquence empirique dans les observations.

2.2 Convergence de l'estimateur de vraisemblance maximale

Comme dans le chapitre précédent, nous commençons par étudier théoriquement le comportement asymptotique de l'estimateur. Dans le cas d'observations markoviennes, les démonstrations se compliquent à cause de la dépendance entre les observations. Dans le but d'alléger un peu l'écriture et de suivre de plus près (Billingsley, 1961), on va récrire la fonction de log-vraisemblance (2.1) et ses dérivées comme suit :

$$\ell_n(\theta) = \sum_{j=1}^{n} g(x_j, x_{j+1}; \theta) \tag{2.3}$$

$$\frac{\partial}{\partial \theta_u} \ell_n(\theta) = \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta), \qquad u = 1, 2, \ldots, r \tag{2.4}$$

où $g(x_j, x_{j+1}; \theta) = \log P_\theta(x_j, x_{j+1})$, $g_u(x_j, x_{j+1}; \theta)$ est la dérivée partielle de g par rapport à la u^{e} composante de θ, et r est la dimension du paramètre θ. Rappelons

que $\hat{\theta}_n$ est une racine de la fonction de score (2.4). La vraie valeur du paramètre sera notée par θ^0.

Avant de démontrer les résultats asymptotiques habituels pour l'estimateur de vraisemblance maximale, nous énonçons deux résultats préalables. Les démonstrations se trouvent à la section 9 de (Billingsley, 1961).

Théorème 2.1. Soit $\{ x_t, t \geq 1 \}$ une chaîne de Markov qui satisfait les hypothèses H1 et H2 et $\varphi : S \times S \to \mathbb{R}$ une fonction. Alors, quelle que soit la loi initiale de la chaîne,

$$\frac{1}{n} \sum_{k=1}^{n} \varphi\left(x_k, x_{k+1}\right) \to \mathbb{E}_\theta \left[\varphi\left(x_1, x_2\right) \right]$$

avec probabilité 1.

Théorème 2.2. Soit $\{ x_t, t \geq 1 \}$ une chaîne de Markov qui satisfait les hypothèses H1 et H2. Alors, pour tout $j \geq 1$, on a

$$\mathbb{E}_\theta \left[g_u\left(x_j, x_{j+1} ; \theta\right) \mid x_j \right] = 0 \,. \tag{2.5}$$

De plus,

$$n^{1/2} \sum_{j=1}^{n} g_u(x_j, x_{j+1} ; \theta) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta))$$

où

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_u} g_u(x_1, x_2 ; \theta) \frac{\partial}{\partial \theta_u} g_u(x_1, x_2 ; \theta) \right] .$$

La démonstration du théorème 2.2 s'appuie sur un théorème de la limite centrale pour les martingales que l'on doit à (Lévy, 1937). En effet, on voit facilement, comme conséquence de la propriété de Markov et de l'égalité (2.5), que les sommes partielles

$$\sum_{j=1}^{n} g_u(x_j, x_{j+1} ; \theta)$$

forment une martingale pour chaque $u = 1, 2, \ldots, r$. La normalité asymptotique découle alors du théorème de Lévy. C'est aussi une conséquence immédiate de (2.5) que le score est de moyenne nulle.

Rappelons que, dans le cas indépendant, nous avons démontré la convergence presque sûre de l'estimateur de vraisemblance maximale. Pour les chaînes de Markov, la dépendance entre les observations ne nous permet pas de faire toutes les simplifications voulues lors des calculs, comme nous avons pu le faire dans le chapitre précédent. Nous allons donc nous contenter de démontrer un théorème de convergence en probabilité énoncé par (Billingsley, 1961).

Théorème 2.3. Soit $\{ x_t, t \in \mathbb{N}\}$ une chaîne de Markov qui satisfait les hypothèses H1 et H2 (donc les conditions C1 à C4). Il existe alors un estimateur de vraisemblance maximale $\hat{\theta}_n = \hat{\theta}(x_1, x_2, \ldots, x_{n+1})$ qui converge en probabilité vers θ^0 quand $n \uparrow \infty$.

Démonstration. En prenant l'espérance dans (2.5), on a d'abord :

$$\mathbb{E}_\theta \left[g_u(x_1, x_2; \theta) \right] = 0. \tag{2.6}$$

D'après la condition C2, nous savons que pour tout (x_i, x_j), $P_\theta(x_i, x_j)$ est trois fois dérivable, et comme $\sum_{x_j \in S} P_\theta(x_i, x_j) = 1$, alors,

$$\frac{\partial}{\partial \theta_u} \sum_{x_j \in S} P_\theta(x_i, x_j) = 0 \,.$$

D'où, pour (1.24) au chapitre précédent, on peut déduire le résultat suivant

$$\mathbb{E}_\theta \left[g_{uv}(x_1, x_2; \theta) \right] = \mathbb{E}_\theta \left[g_u(x_1, x_2; \theta) g_v(x_1, x_2; \theta) \right]$$

qui se traduit en d'autres mots par :

$$\mathbb{E}_\theta \left[g_{uv}(x_1, x_2; \theta) \right] = -\mathcal{I}_{uv}(\theta) \,. \tag{2.7}$$

Comme dans le cas indépendant, choisissons un voisinage N de θ^0 et posons (où v et w sont les indices des deuxième et troisième dérivées partielles respectivement) :

$$G(x_1, x_2) = \sup_{\theta' \in N} |g_{uvw}(x_1, x_2; \theta')|$$

une quantité bien définie par la dérivabilité de P_θ (Condition C2).

Pour $\theta \in N$, un développement de Taylor à l'ordre 1 aux alentours de θ^0 donne

$$\begin{aligned} g_u(x_j, x_{j+1}; \theta) = g_u(x_j, x_{j+1}; \theta^0) + \sum_{v=1}^{r} (\theta_v - \theta_v^0) g_{uv}(x_j, x_{j+1}; \theta^0) \\ + \alpha |\theta - \theta^0| G(x_j, x_{j+1}) \,. \end{aligned}$$

où $\alpha \leq r^2/2$ et $|\theta - \theta^0|$ est la longueur dans S^r. En prenant la moyenne sur $j = 1, \ldots, n$, il vient

$$\begin{aligned} n^{-1} \frac{\partial}{\partial \theta_u} \log L(\theta) &= n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta) \\ &= n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) + \sum_{v=1}^{r} (\theta_v - \theta_v^0)\, n^{-1} \sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) \\ &\quad + \alpha |\theta - \theta^0|^2 n^{-1} \sum_{j=1}^{n} G(x_j, x_{j+1}) \,. \end{aligned} \tag{2.8}$$

Comme S est fini, $\mathbb{E}_\theta\{G(x_j, x_{j+1})\} = M$ est finie. En appliquant le théorème 2.1, (2.6) et (2.7), on obtient les limites suivantes, en probabilité :

$$\begin{aligned} \lim_{n \to +\infty} n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) &= 0, \\ \lim_{n \to +\infty} n^{-1} \sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) &= -\mathcal{I}_{uv}(\theta^0), \\ \lim_{n \to +\infty} n^{-1} \sum_{j=1}^{n} G(x_j, x_{j+1}) &= M \,. \end{aligned} \tag{2.9}$$

Fixons $\epsilon > 0$ et prenons $\delta > 0$ tel que,

$$\delta < \epsilon, \quad \{\theta : |\theta - \theta^0| \leq \delta\} \subset N \,.$$

Choisissons ensuite $n_0(\epsilon)$ assez grand pour que, grâce à (2.9), les résultats suivants soient vrais avec une probabilité $1-\epsilon$, lorsque $n \geq n_0(\epsilon)$:

$$\begin{aligned} \left| n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) \right| &< \delta^2 \\ 0 \leq n^{-1} \sum_{j=1}^{n} G(x_j, x_{j+1}) &< M+1 \\ \left| n^{-1} \sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) + \mathcal{I}_{uv}(\theta^0) \right| &< \delta \qquad u, v = 1, \ldots, r \end{aligned} \tag{2.10}$$

Pour de tels entiers, on utilise (2.8) pour écrire

$$\begin{aligned} & n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta) + \sum_{v=1}^{r} \mathcal{I}_{uv}(\theta^0)(\theta_v - \theta_v^0) \\ & \quad = n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) + \sum_{v=1}^{r} (\theta_v - \theta_v^0)\, n^{-1} \sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) \\ & \qquad + \alpha |\theta - \theta^0|^2 n^{-1} \sum_{j=1}^{n} G(x_j, x_{j+1}) + \sum_{v=1}^{r} \mathcal{I}_{uv}(\theta^0)(\theta_v - \theta_v^0) \\ & \quad = n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) + \alpha |\theta - \theta^0|^2 \; n^{-1} \sum_{j=1}^{n} G(x_j, x_{j+1}) \\ & \qquad + \sum_{v=1}^{r} \left[n^{-1} \sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) + \mathcal{I}_{uv}(\theta^0) \right] (\theta_v - \theta_v^0)\,. \end{aligned}$$

Lorsque $|\theta - \theta^0| \leq \delta$, les inégalités (2.10) mènent alors à la majoration suivante :

$$\begin{aligned} & \left| n^{-1} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta) + \sum_{v=1}^{r} \mathcal{I}_{uv}(\theta^0)(\theta_v - \theta^0) \right| \\ & \qquad \leq \delta^2 + \delta r |\theta - \theta^0| + r^2 |\theta - \theta^0|^2 (M+1)/2 \\ & \qquad \leq 3r^2 (M+1) \delta^2\,. \end{aligned}$$

Si $|\theta - \theta^0| = \delta$, alors

$$\begin{aligned}\sum_{u=1}^{r}\Big[n^{-1}\sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta)\Big](\theta_u - \theta_u^0) \\ &\leq -\sum_{u,v=1}^{r} \mathcal{I}_{uv}(\theta^0)(\theta_u - \theta_u^0)(\theta_v - \theta_v^0) + 3r^2\delta^2(M+1) \\ &\leq -\beta\delta + 3r^2\delta^2(M+1) \qquad (2.11)\end{aligned}$$

où β est une borne inférieure de la forme quadratique $x \mapsto x^t\mathcal{I}_{uv}(\theta^0)x$ lorsque $|x| = 1$; $\beta > 0$ car la matrice $\mathcal{I}_{uv}(\theta^0)$ est définie positive.

On observe alors que si $\delta < \beta/3r^2(M+1)$, le membre de droite de (2.11) est négatif. Donc, pour un δ tel que

$$\delta < \epsilon, \quad \{\theta : |\theta - \theta^0| \leq \delta\} \subset N, \quad \delta < \beta/3r^2(M+1)$$

on a

$$\sum_{u=1}^{r}\left[n^{-1}\sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0)\right](\theta_u - \theta_u^0) < 0 \qquad (2.12)$$

On peut maintenant invoquer le lemme suivant de (Aitchison et Silvey, 1958).

Lemme 2.4. Soit une fonction g satisfaisant C3 et un vecteur θ tel que $|\theta| = 1$. Si $\theta' g(\theta) < 0$ alors il existe un point $\hat{\theta}$ pour lequel $|\hat{\theta}| < 1$ et $g(\hat{\theta}) = 0$.

D'après ce lemme et (2.12), on conclut qu'il existe $\hat{\theta}$ tel que $|\hat{\theta} - \theta^0| \leq \delta \leq \epsilon$ pour lequel $n^{-1}\sum_{j=1}^{n} g_u(x_j, x_{j+1}; \hat{\theta}) = 0,\ u = 1, 2, \ldots, r$. Ceci achève la démonstration puisque, pour $n \geq n_0(\epsilon)$, on vient de montrer qu'il existe une racine $\hat{\theta}_n$ de (2.4) avec une probabilité supérieur ou égale à $1 - \epsilon$. □

Comme dans le cas indépendant, l'estimateur $\hat{\theta}_n$ est un maximum local de (2.3) puisqu'il est une racine du score. Il n'y a rien qui garantit que l'estimateur soit un maximum global de $\ell_n(\theta)$.

Si l'on veut construire des intervalles de confiance pour θ^0, il est nécessaire de connaître la loi de $\hat{\theta}_n$, laquelle est typiquement impossible à trouver pour n fixé. Généralement, on cherche à remplacer la loi de l'estimateur par une approximation plus simple, ce qui fait l'objet de la prochaine section.

2.3 Normalité asymptotique de l'estimateur de vraisemblance maximale

Dans cette section, on va démontrer un théorème de normalité asymptotique pour $\hat{\theta}_n$ donné par (Billingsley, 1961). Pour simplifier les notations, posons le vecteur aléatoire $\beta(n) = (\beta_1(n), \beta_2(n), \ldots, \beta_r(n))$ tel que,

$$\beta_u(n) = n^{-\frac{1}{2}} \sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) \qquad u = 1, 2, \ldots, r \tag{2.13}$$

où r est la dimension de θ. Pour un $\hat{\theta}_n$ convergent vers θ^0, définissons un autre vecteur aléatoire $\gamma(n)$ de longueur r, avec

$$\gamma_i(n) = n^{\frac{1}{2}}(\hat{\theta_{ni}} - \theta_i^0)\,.$$

Théorème 2.5. Supposons que $\{\, x_t,\, t \geq 1 \,\}$ satisfait les conditions C1 à C4. Si $\hat{\theta}_n$ est un estimateur convergent de θ vers θ^0, alors

$$\gamma(n) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}^{-1}(\theta^0)).$$

Démonstration. Nous avons déjà énoncé dans le théorème 2.2 que

$$\beta(n) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta^0))\,.$$

Comme $\mathcal{I}(\theta^0)$ est inversible et symétrique, on a aussi

$$\mathcal{I}(\theta^0)^{-1}\beta(n) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta^0)^{-1})\,. \tag{2.14}$$

Notre but est de trouver une suite équivalente à $\mathcal{I}(\theta^0)^{-1}\beta(n)$ en probabilité et d'utiliser un lemme de convergence qui sera présenté ci-dessous. Pour ce faire,

nous allons commencer par le développement de Taylor de (2.4) à $\hat{\theta}_n$ au voisinage de θ^0. Vu que $P_\theta(x_j, x_{j+1})$ est trois fois continûment dérivable, on va faire un développement de Taylor à l'ordre 1 :

$$n^{-1}\frac{\partial}{\partial\theta_u}L_n(\hat{\theta}_n) = n^{-1}\sum_{j=1}^{n} g_u(x_j, x_{j+1}; \theta^0) + \sum_{v=1}^{r}(\hat{\theta}_{nv} - \theta_v^0)n^{-1}\sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) + \alpha|\hat{\theta}_n - \theta^0|^2 n^{-1}\sum_{j=1}^{n} G(x_j, x_{j+1}) = 0 \tag{2.15}$$

où $|\alpha| \leq r^2/2$, $G(x_j, x_{j+1})$ est tel que défini dans la preuve précédente et l'égalité est nulle car $\hat{\theta}_n$ est une solution de (2.4).

On veut faire apparaître $\beta(n)$ dans (2.15), on multiplie par $n^{\frac{1}{2}}$ et, en se rappelant (2.13), on obtient :

$$\beta_u(n) + \sum_{v=1}^{r}\gamma_v(n)\left[n^{-1}\sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0)\right] + \alpha|\hat{\theta}_n - \theta^0|\,|\gamma(n)|\left[n^{-1}\sum_{j=1}^{n} G(x_j, x_{j+1})\right] = 0$$

où $\gamma_v(n) = (\hat{\theta}_n - \theta_v^0)$ et $u = 1, 2, \ldots, r$.

Pour faire apparaître $\mathcal{I}(\theta^0)^{-1}$, nous allons faire appel aux résultats (2.9) et (2.10) de la preuve précédente et les remplacer dans l'égalité précédente :

$$\left|\,\mathcal{I}(\theta^0)^{-1}\beta_u(n) - \gamma(n)\,\right| \leq \frac{r^2}{2}|\hat{\theta}_n - \theta^0|(M+1)|\gamma(n)|\,.$$

Comme $\hat{\theta}_n$ est faiblement convergent, nous savons que $|\hat{\theta}_n - \theta^0| \to 0$ avec une probabilité proche de 1. Il s'ensuit que

$$\left|\,\mathcal{I}(\theta^0)^{-1}\beta_u(n) - \gamma(n)\,\right| \leq \epsilon_n\,|\gamma(n)|$$

où ϵ_n convergence en probabilité vers 0. Nous faisons ensuite appel à un résultat de convergence (démontré en appendice A) qui s'énonce comme suit : *Si* $u_n \hookrightarrow^{\mathcal{L}} \mu$

et si $|u_n - v_n| \leq \epsilon_n |v_n|$ *où* ϵ_n *converge en probabilité vers zéro, alors* $v_n \hookrightarrow^{\mathcal{L}} \mu$ *;* u_n *et* v_n *sont des vecteurs aléatoires dans* $\mathbb{R}^r$. De ce résultat et (2.14), on peut conclure immédiatement que $\gamma(n) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta^0)^{-1})$. □

La matrice de variance-covariance asymptotique de l'estimateur de vraisemblance maximale s'obtient donc en inversant la matrice d'information de Fisher qui, rappelons-le, se définit comme suit :

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_{ij}} \log P_\theta(x_1, x_2) \frac{\partial}{\partial \theta_{i'j'}} \log P_\theta(x_1, x_2) \right] \tag{2.16}$$

Compte tenu de la forme que nous avons donné à P_θ, il est possible d'expliciter $\mathcal{I}(\theta)$ ce qui sera commode pour l'étude de simulation à venir. Examinons ici le cas 2×2, juste pour illustrer un peu l'allure de la matrice. Commençons par le calcul des dérivées de chaque entrée de la matrice de transition P_θ par rapport à θ_{12} sont,

$$\begin{aligned}
\frac{\partial}{\partial \theta_{12}} \log P_\theta(1,1) &= \frac{\partial}{\partial \theta_{12}} \log(1 - \theta_{12}) = -\frac{1}{1 - \theta_{12}} \\
\frac{\partial}{\partial \theta_{12}} \log P_\theta(1,2) &= \frac{\partial}{\partial \theta_{12}} \log \theta_{12} = \frac{1}{\theta_{12}} \\
\frac{\partial}{\partial \theta_{12}} \log P_\theta(2,1) &= \frac{\partial}{\partial \theta_{12}} \log P_\theta(2,2) = 0
\end{aligned}$$

et par rapport à θ_{21}

$$\begin{aligned}
\frac{\partial}{\partial \theta_{21}} \log P_\theta(1,1) &= \frac{\partial}{\partial \theta_{21}} \log P_\theta(1,2) = 0 \\
\frac{\partial}{\partial \theta_{21}} \log P_\theta(2,1) &= \frac{1}{\theta_{21}} \\
\frac{\partial}{\partial \theta_{21}} \log P_\theta(2,2) &= \frac{-1}{1 - \theta_{21}}
\end{aligned}$$

Ceci nous permet de calculer les entrées de la matrice (2.16), en prenant l'espérance

pour P_θ et la mesure de probabilité stationnaire π correspondante :

$$\begin{aligned}
\mathcal{I}_{11}(\theta) &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_{12}}\log P_\theta(X_0,X_1)^2\right] \\
&= \sum_{x_0,x_1} P\{X_0=x_0,X_1=x_1\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(x_0,x_1)^2 \\
&= P\{x_0=1,x_1=1\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(1,1)^2+\cdots \\
&\qquad \cdots + P\{x_0=2,x_1=2\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(2,2)^2 \\
&= \pi(1)(1-\theta_{12})\left(\frac{-1}{1-\theta_{11}}\right)+\pi(1)\theta_{12}\left(\frac{1}{\theta_{12}}\right)^2 \\
&\qquad + \pi(2)\theta_{21}\times 0+\pi(2)(1-\theta_{21})\times 0 \\
&= \frac{\pi(1)}{1-\theta_{21}}+\frac{\pi(1)}{\theta_{21}}
\end{aligned}$$

De même, il vient

$$\begin{aligned}
\mathcal{I}_{12}(\theta) &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_{12}}\log P_\theta(x_0,x_1)\frac{\partial}{\partial\theta_{21}}\log P_\theta(x_0,x_1)\right] \\
&= \sum_{x_0,x_1} P\{X_0=x_0,X_1=x_1\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(x_0,x_1)\frac{\partial}{\partial\theta_{21}}\log P_\theta(x_0,x_1) \\
&= P\{x_0=1,x_1=1\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(1,1)\frac{\partial}{\partial\theta_{21}}\log P_\theta(1,1)+\cdots \\
&\qquad \cdots + P\{x_0=2,x_1=2\}\frac{\partial}{\partial\theta_{12}}\log P_\theta(2,2)\frac{\partial}{\partial\theta_{21}}\log P_\theta(2,2) \\
&= \pi(1)(1-\theta_{12})\left(\frac{-1}{1-\theta_{11}}\right)\times 0+\pi(1)\theta_{12}\left(\frac{1}{\theta_{12}}\right)\times 0 \\
&\qquad + \pi(2)\theta_{21}\times 0\times\frac{1}{\theta_{21}}+\pi(2)(1-\theta_{21})\times 0\times\left(\frac{-1}{1-\theta}\right) \\
&= 0
\end{aligned}$$

Le calcul de $\mathcal{I}_{21}(\theta)$ et $\mathcal{I}_{22}(\theta)$ se fait de la même manière et la matrice résultante est donnée par

$$\mathcal{I}(\theta)=\begin{bmatrix} \frac{\pi(1)}{1-\theta_{21}}+\frac{\pi(1)}{\theta_{21}} & 0 \\ 0 & \frac{\pi(2)}{1-\theta_{21}}+\frac{\pi(2)}{\theta_{21}} \end{bmatrix}$$

et puisque la loi stationnaire est

$$\pi(1) = \frac{\theta_{21}}{\theta_{12} + \theta_{21}}$$
$$\pi(2) = \frac{\theta_{12}}{\theta_{21} + \theta_{21}}$$

on obtient

$$\mathcal{I}(\theta) = \begin{bmatrix} \left(\frac{\theta_{21}}{\theta_{12}+\theta_{21}}\right)\left(\frac{1}{\theta_{11}(1-\theta_{11})}\right) & 0 \\ 0 & \left(\frac{\theta_{12}}{\theta_{12}+\theta_{21}}\right)\left(\frac{1}{\theta_{21}(1-\theta_{21})}\right) \end{bmatrix} \tag{2.17}$$

Rappelons que, dans la réalité, la vraie valeur θ^0 du paramètre est inconnue, donc $\mathcal{I}(\theta^0)$ aussi. C'est pour cela qu'on cherche un moyen de l'estimer. Dans ce contexte, la matrice d'information de Fisher sera souvent remplacée par la *matrice d'information observée* qui fait l'objet de la prochaine section.

2.4 Convergence de la matrice d'information observée

L'information de Fisher telle que définie dans le chapitre précédent est la variance du score qui se calcule comme étant l'espérance de l'information observée. Nous avons mentionné que le calcul de l'espérance n'est toujours pas évident dans la réalité. Surtout quand on ne connaît pas la distribution de provenance des observations $x_1, x_2, \ldots, x_{n+1}$. L'information observée a été introduite pour répondre à cette difficulté, elle se définit comme suit :

$$\mathcal{I}_o(\hat{\theta}_n) = -\frac{\partial^2 \ell_n}{\partial\theta\partial\theta^t}\bigg|_{\theta=\hat{\theta}_n} \tag{2.18}$$

La question, que l'on peut se poser est : est-ce que $\mathcal{I}_o(\hat{\theta}_n)$ converge vers $\mathcal{I}(\theta^0)$ Pour répondre à cette question nous allons essayer de montrer cette convergence en espérant qu'asymptotiquement (2.18) n'est pas très loin de $\mathcal{I}(\theta^0)$.

Vu que le développement de Taylor occupe une place considérable dans nos démonstrations, nous allons continuer sur cette lancée et faire un développement à

l'ordre 1 de la deuxième dérivée de la fonction de la log-vraisemblance à $\hat{\theta}_n$ au voisinage de la vraie valeur. Ceci est donné par :

$$n^{-1}\frac{\partial^2 \ell_n}{\partial\theta_u\partial\theta_v}\bigg|_{\theta=\hat{\theta}_n} = n^{-1}\sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) + r|\hat{\theta}_n - \theta^0|\Big(n^{-1}\sum_{j=1}^{n} G(x_j; x_{j+1})\Big)\,.$$

Pour faire apparaître $\mathcal{I}(\theta^0)$, nous allons la soustraire des deux côtés de l'égalité précédente :

$$\begin{aligned} n^{-1}\frac{\partial^2 \ell_n}{\partial\theta_u\partial\theta_v}\bigg|_{\theta=\hat{\theta}_n} - \mathcal{I}_{uv}(\theta_0) = n^{-1}\sum_{j=1}^{n} g_{uv}(x_j, x_{j+1}; \theta^0) - \mathcal{I}_{uv}(\theta_0) \\ + r|\hat{\theta}_n - \theta^0|\Big(n^{-1}\sum_{j=1}^{n} G(x_j; x_{j+1})\Big) \end{aligned}$$

Maintenant, en choisissant un δ assez petit pour que $\{\hat{\theta_n} : |\hat{\theta}_n - \theta| \leq \delta \leq \epsilon\} \subset N$. Alors, les inégalités (2.10) sont vérifiées pour un tel $\hat{\theta}_n$. Donc, on a

$$\Big|\; n^{-1}\frac{\partial^2 \ell_n}{\partial\theta_u\partial\theta_v}\bigg|_{\theta=\hat{\theta}_n} - \mathcal{I}_{uv}(\theta^0)\;\Big| \leq \delta + \delta r(M+1) < \epsilon$$

si $\delta < \epsilon/(1 + r(M+1))$. Ce qui conclut notre résultat.

Un peu comme dans la section précédente, nous allons terminer en effectuant le calcul de l'information observée pour le cas 2×2. Nous écrivons la log-vraisemblance comme suit :

$$\ell(\theta) = n_{11}\log(\theta_{11}) + n_{12}\log(1-\theta_{11}) + n_{21}\log(1-\theta_{22}) + n_{22}\log(\theta_{22})\,.$$

La première ligne de son hessien est donnée par

$$\begin{aligned} -\frac{\partial^2 \ell}{\partial\theta_{11}\partial\theta_{11}} &= -\frac{\partial}{\partial\theta_{11}}\left(\frac{n_{11}}{\theta_{11}} - \frac{n_{12}}{(1-\theta_{11})}\right) \\ &= -\frac{\partial}{\partial\theta_{11}}\left(\frac{n_{11} - \theta_{11}(n_{11}+n_{12})}{\theta_{11}(1-\theta_{11})}\right) \\ &= \frac{n_{11} - 2n_{11}\theta_{11} + \theta_{11}^2(n_{11}+n_{12})}{[\theta_{11}(1-\theta_{11})]^2} \end{aligned}$$

et

$$-\frac{\partial^2 \ell}{\partial \theta_{11} \partial \theta_{22}} = 0\,.$$

Un calcul analogue pour la seconde ligne mène à la forme finale de l'information observée :

$$\mathcal{I}_o(\hat{\theta}_n) = \begin{bmatrix} \dfrac{n_{11} - 2n_{11}\hat{\theta}_{11} + \hat{\theta}_{11}^2(n_{11}+n_{12})}{\left[\hat{\theta}_{11}(1-\hat{\theta}_{11})\right]^2} & 0 \\ 0 & \dfrac{n_{22} - 2n_{22}\hat{\theta}_{22} + \hat{\theta}_{22}^2(n_{22}+n_{21})}{\left[\hat{\theta}_{22}(1-\hat{\theta}_{22})\right]^2} \end{bmatrix}.$$

La matrice d'information observée joue un rôle très important en pratique puisque, habituellement, elle remplace (dans l'inférence) la matrice de variance-covariance théorique, inutilisable directement du fait qu'elle dépend de paramètres inconnus.

2.5 Étude par simulation

Les résultats des sections précédentes permettent de vérifier que les estimateurs du maximum de vraisemblance ont de bonnes propriétés asymptotiques. Cependant, en pratique, il est aussi intéressant d'avoir une idée sur la qualité des estimateurs lorsque la longueur de la séquence observée, n, est fixée. Cela permet en effet d'avoir une idée de la quantité de données nécessaires pour obtenir des estimateurs de qualité raisonnable. Pour cela, nous avons réalisé une étude de simulation, basée sur une chaine de Markov à deux états, avec la matrice de transition :

$$P_\alpha = \begin{bmatrix} 0.75 & 0.25 \\ 0.45 & 0.55 \end{bmatrix}$$

2.5.1 Convergence de l'estimateur

Nous avons simulé plusieurs fois la chaîne pour des longueurs de séquences différentes, n, correspondant à des tailles petites, moyennes et très grandes, à savoir $50, 150, 300, 500, \ldots, 50000$. Pour chacune des longueurs, nous avons calculé l'estimateur de vraisemblance maximale.

Ceci est fait pour les deux paramètres inconnus de la chaîne dans le but de regarder la variation des estimateurs autour des vraies valeurs selon la longueur de la séquence. En fait, on veut surtout voir à quelle vitesse l'estimateur de vraisemblance maximale converge.

On peut voir dans la figure 2.1 que la propriété de convergence asymptotique est vérifiée et que plus la taille de l'échantillon est grande plus l'estimateur est proche de la véritable valeur du paramètre. Pour les petits échantillons (inférieur à 100), on voit que les estimés oscillent beaucoup ce qui suggère que les estimateurs ne

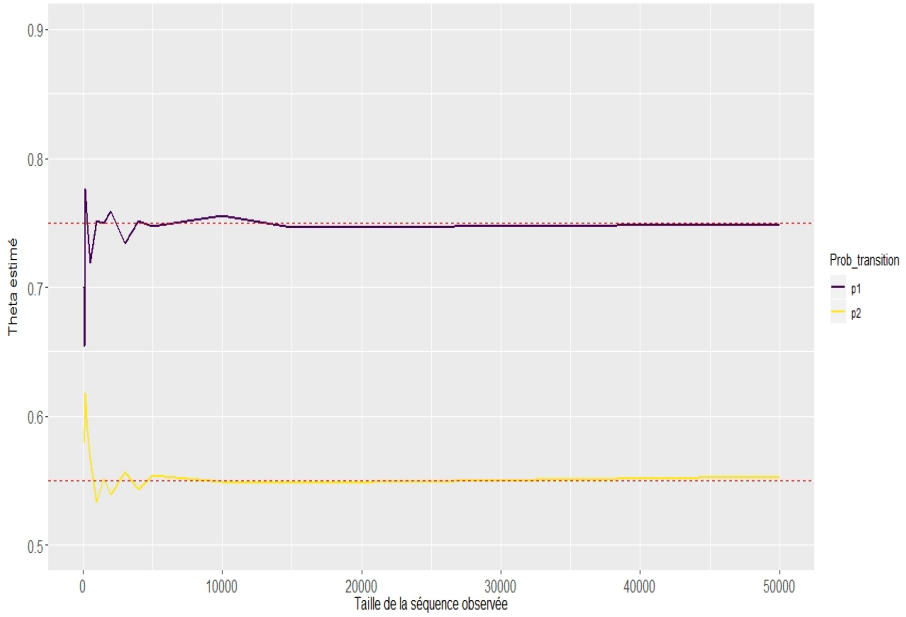

Figure (2.1) Convergence de l'estimateur de vraisemblance maximale dans le cas d'une chaîne de Markov

sont pas très précis.

2.5.2 Normalité et biais de l'estimateur

Pour confirmer la normalité du maximum de vraisemblance, à chaque fois, 1000 estimateurs ont été calculés pour des séquences simulées de la chaîne, pour des tailles 50, 200 et 6000. Les figures 2.2, 2.3 et 2.4 donnent des histogrammes de la distribution des estimateurs.

Grâce à cette simulation, on voit dans les figures 2.2 et 2.3 que les distributions sont asymétriques à gauche. Par contre, dans la figure 2.4, on voit que distribution est presque symétrique. Enfin, comme dans la simulation précédente, la conclusion est que plus la taille de l'échantillon est grande plus la normalité de l'estimateur de vraisemblance maximale est vérifiée.

Deux autres éléments peuvent être remarqués d'après les mêmes figures. Dans un premier temps, on peut voir que l'estimateur est biaisé surtout pour les tailles d'échantillon petite et moyenne. Ceci est confirmé dans le tableau 2.1 dans lequel on voit que plus la taille de l'échantillon est grande plus le biais de l'estimateur s'approche vers 0. Dans un deuxième temps, on peut voir, toujours d'après les figures, que la variance de l'estimateur diminue avec la taille d'échantillon. Cette hypothèse sera confirmée dans la prochaine simulation.

2.5.3 Variance estimée par l'information de Fisher observée

Pour chacun des estimateurs calculés dans la simulation précédente, nous avons calculé l'information de Fisher observée présentée par (2.18). Ceci nous donne 1000 estimations de la variance du maximum de vraisemblance obtenues en inver-

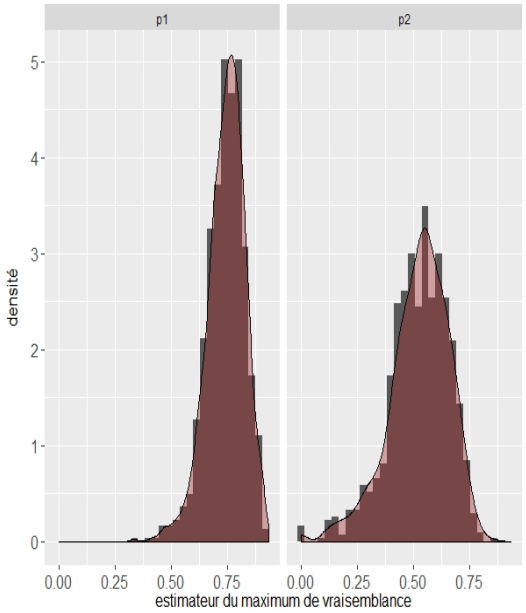

Figure (2.2) Distribution des estimateurs ($n = 50$).

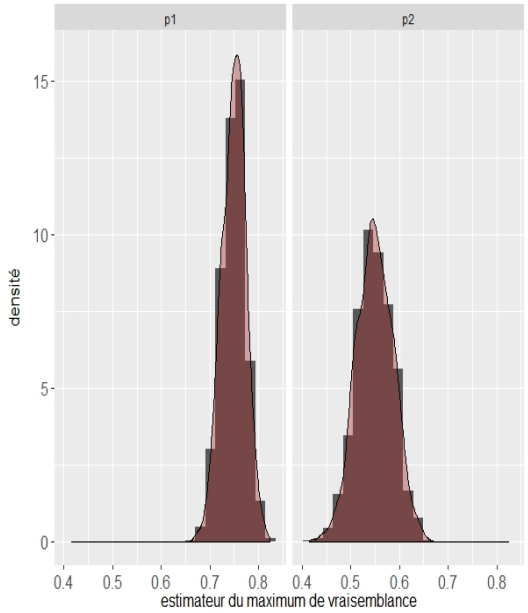

Figure (2.3) Distribution des estimateurs ($n = 500$).

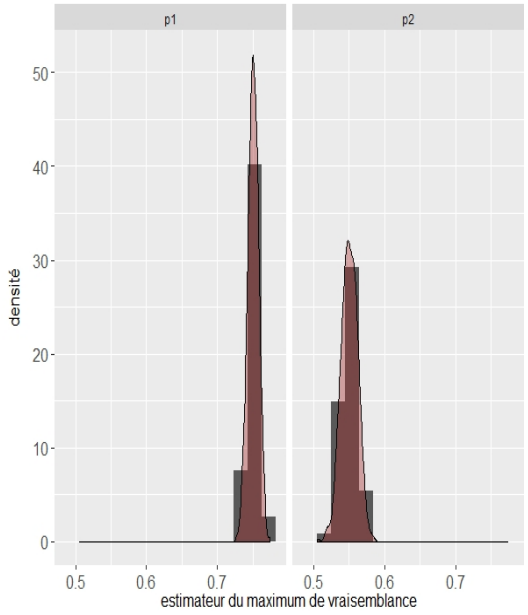

Figure (2.4) Distribution des estimateurs ($n = 5000$).

N	Biais($\hat{p}_1$)	Biais($\hat{p}_2$)
50	0.01568	0.03519
500	0.00101	0.01142
5000	0.00033	0.00063

Tableau (2.1) Biais des estimateurs selon la taille de l'échantillon simulé.

sant l'information de Fisher observée pour chaque taille d'échantillon. De plus, la variance théorique est calculée en inversant (2.17). Cette simulation est présentée par les figures 2.5, 2.6 et 2.7.

Pour les différentes tailles d'échantillon, on remarque dans un premier temps que les médianes des distributions sont proches de la valeur de la variance théorique (la variance théorique est représentée par une ligne pointillée). Par contre, la dispersion des variances estimées est considérable pour les échantillons de taille 50 et celle-ci diminue avec la taille de l'échantillon. Ce qui est confirmé dans la figure 2.7 dans laquelle on voit que la variation est presque négligeable.

Vu que nous avons la variance estimée pour chaque échantillon simulé, nous avons calculé des intervalles de confiance asymptotiques pour le paramètre et estimer la probabilité de couverture qui en découle sachant que la probabilité de couverture théorique est fixée à 95%. Les résultats sont résumés dans le tableau ci-dessous.

n	$P_c(p_1)$	$P_c(p_2)$
50	0.922	0.919
500	0.94	0.934
5000	0.957	0.955

Tableau (2.2) Probabilités de couverture selon la taille de l'échantillon.

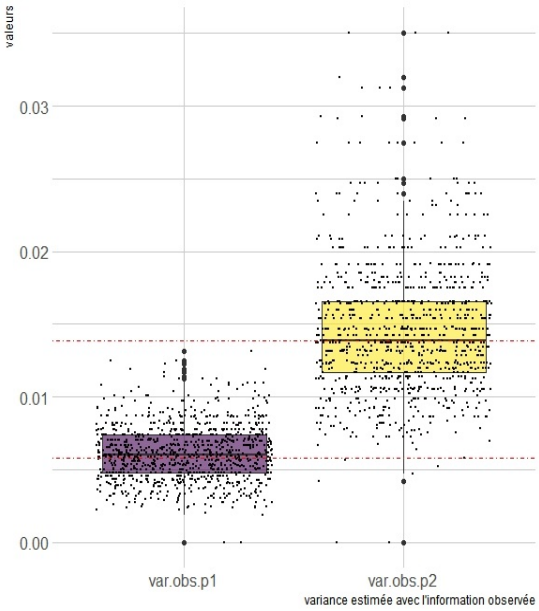

Figure (2.5) Distribution de la variance estimée ($n = 50$).

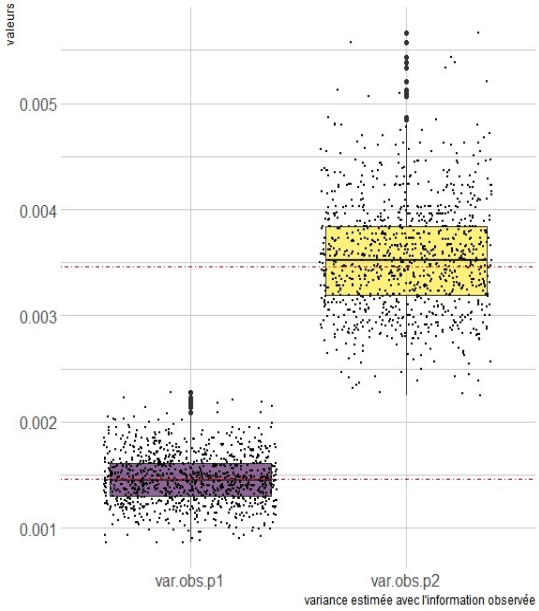

Figure (2.6) Distribution de la variance estimée ($n = 500$).

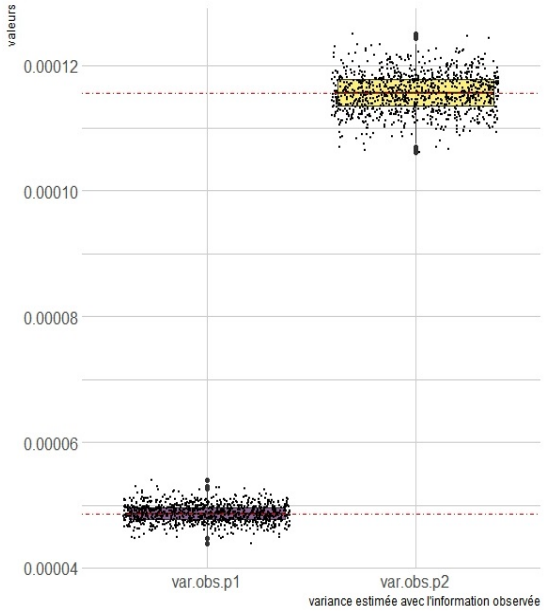

Figure (2.7) Distribution de la variance estimée ($n = 5000$).

D'après cette étude de simulation, on constate que la théorie asymptotique est vérifiée pour l'estimateur de vraisemblance maximale dans le cas d'une dépendance markovienne. En outre, comme nous avons obtenu des probabilités de couverture supérieur à 92% pour les échantillons de petite ou moyenne taille, on peut aussi dire que les estimateurs de vraisemblance maximale sont bons dans ces cas. Finalement, pour le point qui nous préoccupait le plus, à savoir l'estimation de la variance par l'information observée, les résultats que nous avons obtenus montrent que l'information observée est un estimateur raisonnable de la variance théorique.

CHAPITRE III

CHAÎNE DE MARKOV CACHÉE

Un modèle avec chaîne de Markov cachée est en quelque sorte un mélange des modèles examinés dans les deux précédents chapitres. Conditionnellement à la trajectoire de la chaîne, les données sont indépendantes mais les paramètres de leurs lois changent selon l'état de la chaîne, lequel n'est pas observé par ailleurs (d'où l'adjectif cachée). La situation est plus générale que celle des modèles avec point de rupture, puisque le changement de paramètres peut survenir à plusieurs reprises. Pour le développement du chapitre, nous allons suivre deux textes : le premier est celui de (Baum et Petrie, 1966), les premiers à étudier l'estimation de vraisemblance maximale pour ce genre de modèle et le deuxième est celui de (Hamilton, 1987) qui a proposé un algorithme itératif pour le calcul de la fonction de vraisemblance. S'en suivront des évaluations empiriques de la convergence dont la méthodologie sera précisée au préalable.

3.1 Généralités

Un modèle avec chaîne de Markov cachée est la donnée de deux processus stochastiques :

i) un premier processus X non observable, constitué d'un ensemble d'états connectés entre eux par des probabilités de transition ;

ii) un deuxième processus Y observable, constitué d'un ensemble de sorties ou observations. Chaque sortie peut être émise par n'importe quel état caché, conformément à une fonction de densité de probabilité de sortie qui dépend de cet état.

Dans bien des applications utilisant ce type de modèle, on cherche d'abord à identifier les états de la chaîne cachée qui ont le plus probablement donné lieu aux observations. Dans ce mémoire, on se préoccupera plutôt d'estimer (par vraisemblance maximale) les probabilités de transition de la chaîne ainsi que les paramètres de la loi de sortie.

Pour écrire le modèle plus formellement, on a $\{X_t, t \in \mathbb{N}\}$ une chaîne de Markov, gouvernée par une matrice de transition P_α $(s \times s)$. Pour chaque instant t, on a de plus Y_t, une variable discrète observée, dont la loi est déterminée par l'état de la chaîne X_t à cet instant :

$$P\{\, Y_t = k \mid X_t = j \,\} = b_{jk} \tag{3.1}$$

où $j \in \{1, 2, \ldots, s\}$ et $k \in \{1, 2, \ldots, r\}$. Donc, pour chaque état de la chaîne de Markov $\{X_t\}$, nous avons r probabilités pour les observations de Y_t. L'ensemble de ces probabilités forme une matrice $s \times r$ qui sera notée P_β. De plus, puisque ces probabilités ne dépendent que de l'état courant de la chaîne, les variables Y_t sont indépendantes conditionnellement à la trajectoire de celle-ci.

Nous observons $(y_1, \ldots, y_n)$ et nous voulons estimer les paramètres α et β. Les matrices P_α et P_β sont stochastiques et, comme dans le chapitre précédent, nous les écrirons directement en termes de α et β. Par exemple, pour un modèle à deux états cachés et trois états observables, l'espace d'états de la chaîne est $S = \{1, 2\}$ et P_α est donnée par

$$\begin{bmatrix} 1 - \alpha_{12} & \alpha_{12} \\ \alpha_{21} & 1 - \alpha_{21} \end{bmatrix}$$

avec $\alpha = \{\alpha_{12}, \alpha_{21}\}$. L'ensemble des états observables est défini par $R = \{1, 2, 3\}$ et la matrice d'émission (3.1) s'écrit

$$\begin{bmatrix} 1-\beta_{12}-\beta_{13} & \beta_{12} & \beta_{13} \\ \beta_{21} & 1-\beta_{21}-\beta_{23} & \beta_{23} \end{bmatrix}$$

où $\beta = \{\beta_{12}, \beta_{13}, \beta_{21}, \beta_{23}\}$.

Les deux paramètres vectoriels α et β sont dans $[0,1]^2$ et $[0,1]^4$ respectivement. Nous supposons P_α et P_β satisfont les hypothèses H1 et H2 du chapitre précédent. Suite à ceci, la chaîne $\{X_t\}$ est irréductible, récurrente positive avec une seule classe ergodique et admet une seule mesure de probabilité stationnaire π_α.

Ces hypothèses sur X_t et Y_t font en sorte que le couple $Z_t = (Y_t, X_t)$ est une chaîne de Markov homogène sur l'ensemble d'états $S \times R$ avec une matrice de transition P_θ donnée par

$$P_\theta\left((x,y),(x',y')\right) = P_\alpha(x,x')P_\beta(x',y')$$

Il s'agit d'une matrice $rs \times rs$. Le paramètre $\theta = (\alpha, \beta)$ appartient à un ensemble Θ_δ qui est compact. Nous ajoutons l'indice δ pour rappeler le fait que les entrées des matrices P_α et P_β sont toutes supérieures à $\delta > 0$.

La chaîne $\{Z_t\}$ est également irréductible et récurrente positive sous les hypothèses H1 et H2. L'irréductibilité vient du fait que, par hypothèse, les coefficients de P_θ sont strictement positifs. De plus, l'espace des états est fini donc l'irréductibilité entraîne que la chaîne est récurrente positive, ce qui entraîne à son tour que la chaîne admet une unique loi stationnaire.

On tient à noter que, même si $\{X_t\}$ et $\{Z_t\}$ sont des chaînes de Markov, ce n'est pas le cas de $\{Y_t\}$. Ces variables ne sont pas non plus indépendantes.

3.2 Vraisemblance et entropie de Shannon

Dans le cas des chaînes de Markov, la fonction de vraisemblance n'est pas trop difficile à calculer car X_j dépend seulement de X_{j-1}. Le cas dans lequel on est en ce moment est un peu particulier car non seulement on n'observe pas les états de la chaîne $\{X_t\}$ mais les observations $(y_1, y_2, \ldots, y_n)$ sont dépendantes.

La chaîne $\{Z_t\}$ est ergodique et n'a qu'une seule mesure de probabilité stationnaire. Il est donc possible de travailler avec une version stationnaire de cette chaîne, définie pour des $t \in \mathbb{Z}$. Dans ce contexte, les observations $(y_1, y_2, \ldots, y_n)$ sont une sorte de fenêtre finie sur un processus issu de $t = -\infty$ et allant vers $t = \infty$. C'est le point de vue que nous adopterons pour la suite.

La fonction de vraisemblance est définie par :

$$L(\theta; y_1, y_2, \ldots, y_n) = \sum_{x_1} \cdots \sum_{x_n} \pi_\alpha(x_1) P_\beta(x_1, y_1) \cdots P_\alpha(x_{n-1}, x_n) P_\beta(x_n, y_n)\,.$$

Cette fonction est difficile à étudier analytiquement et son calcul numérique, sous cette forme, peut devenir très lourd. De plus, la définition de l'information de Fisher n'est pas claire dans ce contexte. La première contribution de (Baum et Petrie, 1966) est de prouver l'existence d'une entropie limite, dont sera déduite la variance asymptotique de l'estimateur de vraisemblance maximale.

Expliquons de quoi il s'agit. La chaîne $\{Z_t\}$ est stationnaire ce qui implique que $\{Y_t\}$ l'est aussi. De plus, les probabilités conditionnelles

$$\Pr{}_\theta \{Y_0 = y_0 \mid Y_{-1} = y_{-1}, \ldots, Y_{-k+1} = y_{-k+1}\}$$

sont bien définies puisque toutes les entrées de P_θ sont positives. On peut donc définir une variable aléatoire $f_k(\theta, Y)$ en remplaçant dans cette probabilité, les valeurs $y_0, y_{-1}, \ldots, y_{-k+1}$ par les variables $Y_0, Y_{-1}, \ldots, Y_{-k+1}$. Le point de départ

de (Baum et Petrie, 1966) est le fait que la limite

$$f(\theta, Y) = \lim_{k \to \infty} f_k(\theta, Y) \tag{3.2}$$

existe pour toute trajectoire du processus $\{Y_t\}$. Ceci découle d'une série d'inégalités techniques qui font l'objet de la section 2 de l'article (et qu'on ne détaillera pas ici). Si on pose $g(\theta, Y) = \log f(\theta, Y)$, l'entropie H est alors donnée par

$$H(\theta) = \mathbb{E}_{\theta^0} [\, g(\theta, Y) \,] \tag{3.3}$$

avec θ^0 la vraie valeur de θ (celle qu'on cherche à estimer) et où l'espérance est prise sous la loi qui lui correspond.

Comme il est démontré dans (Baum et Petrie, 1966), cette entropie atteint son maximum en θ_0. L'égalité $H(\theta) = H(\theta_0)$ implique $\{Y_t\}$ a la même loi sous θ ou θ_0. En d'autre mots, le modèle n'est alors pas identifiable. De fait, étant donné la forme de P_θ comme produit de matrices stochastiques, il se peut que deux valeurs distinctes de θ donne lieu à la même loi pour $\{Y_t\}$ (voir à ce sujet, (Gilbert, 1959)). Cependant la paramétrisation que nous donnons à P_α et P_β fait en sorte que ce ne sera pas le cas, comme on peut le déduire de la discussion présente dans la section 1 de (Baum et Petrie, 1966). Dès lors, on a toujours $H(\theta) < H(\theta^0)$ aussitôt que $\theta \neq \theta^0$.

L'entropie joue un rôle important parce qu'on peut démontrer que

$$n^{-1} \log L(\theta; Y_1, \ldots, Y_n) \to H(\theta)$$

presque sûrement pour la loi de $\{Y_t\}$ sous θ^0 ((Baum et Petrie, 1966), théorème 3.2). De cette convergence, on conçoit intuitivement que les maximums locaux de la vraisemblance convergent vers ceux de H, lesquels se réduisent à θ^0. En d'autres mots, l'estimateur de vraisemblance maximale est convergent.

3.3 Propriétés asymptotiques de l'estimateur de vraisemblance maximale

Cherchons d'abord une expression différente pour la vraisemblance. On peut l'écrire avec un produit de probabilités conditionnelles :

$$L(\theta; y_1, \ldots, y_n) = \Pr{}_\theta (Y_1 = y_1) \prod_{k=2}^{n} \Pr{}_\theta (Y_k = y_k \mid Y_{k-1} = y_{k-1}, \ldots, Y_1 = y_1) \ .$$

Si on remplace dans ces probabilités conditionnelles, les valeurs observées par les variables, on voit que L peut s'écrire en termes des fonctions f_k de la section précédente :

$$L(\theta; Y_1, \ldots, Y_n) = \prod_{k=1}^{n} f_k(\theta, T^k Y)$$

où T est l'opérateur de décalage sur une trajectoire de $\{Y_t\}$: $(TY)_i = Y_{i+1}$.

La log-vraisemblance prend alors une forme familière, semblable à (2.3) :

$$h_n(\theta, Y) = n^{-1} \log L(\theta; Y_1, \ldots, Y_n) = n^{-1} \sum_{k=1}^{n} g_k \left(\theta, T^k Y\right)$$

avec $g_k(\theta, Y) = \log f_k(\theta, Y)$. Nous calculons l'estimateur de vraisemblance maximale en cherchant une racine de la dérivée de la fonction de log-vraisemblance par rapport aux paramètres. Ceci est un peu plus compliqué dans le cas d'un modèle avec chaîne de Markov cachée car, avant tout, nous devons vérifier la dérivabilité de $h_n(\theta, Y)$ ainsi que la convergence de la dérivée vers celle de l'entropie $H(\theta)$.

Pour ce faire, (Baum et Petrie, 1966) ont démontré un résultat qui prend une place importante dans l'étude des propriétés asymptotiques de l'estimateur.

Théorème 3.1. Soit $(d\,)$ (en indice supérieur) une dérivée d'ordre d par rapport aux paramètres. Alors, uniformément en θ et pour toute trajectoire de $\{Y_t\}$, on a

a) $\lim\limits_{k\to\infty} g_k^{(d)}\,(\theta, Y) = g^{(d)}\,(\theta, Y)\,;$

b) $H^{(d)}(\theta)$ existe et $\lim\limits_{n\to\infty} h_n^{(d)}(\theta, Y) = H^{(d)}(\theta)$.

La fonction $f_k(\theta, Y)$ est dérivable en θ comme conséquence de la paramétrisation de P_α et P_β que nous avons choisie. Donc, $g_k(\theta, Y)$ l'est aussi et le théorème précédent assure que cette dérivabilité se transporte à la limite.

Théorème 3.2. Soit $\{(X_t, Y_t)\}$ un modèle avec chaîne de Markov cachée. Pour chaque n, il existe une solution de $\frac{\partial}{\partial\theta} h_n(\theta, Y) = 0$, notée $\hat{\theta}_n$, telle que $\hat{\theta}_n \to \theta^0$ en probabilité.

Démonstration. La preuve de ce théorème ressemble à celle du théorème de convergence dans le cas des chaînes de Markov. Il suffit d'utiliser le théorème 3.1, la définition (3.3) et de suivre la preuve du chapitre précédent, en remplaçant les termes $g_u(x_j, x_{j+1}, \theta^0)$, $g_{uv}(x_j, x_{j+1}, \theta^0)$ et $g_{uvw}(x_j, x_{j+1}, \theta^0)$ respectivement par $g_k^{(u)}(\theta^0, T^k Y)$, $g_k^{(uv)}(\theta^0, T^k Y)$ et $g_k^{(uvw)}(\theta^0, T^k Y)$. □

En suivant l'ordre de présentation des résultats dans le chapitre II, la normalité asymptotique de l'estimateur de vraisemblance maximale vient après le résultat de sa convergence. Dans ce cas d'un modèle avec chaîne de Markov cachée, la preuve du théorème central limite suit exactement les mêmes étapes que celles que nous avons détaillées pour le cas d'une chaîne de Markov simple. À cause de cela, nous allons seulement illustrer les résultats de (Baum et Petrie, 1966).

Définissons déjà la matrice de variance-covariance asymptotique de la fonction du score qui s'écrit comme :

$$\mathcal{I}_{uv}(\theta^0) = \left.\frac{\partial^2 H}{\partial\theta_u \partial\theta_v}\right|_{\theta=\theta^0} \tag{3.4}$$

qui est supposée être non singulière.

La normalité du score a joué un rôle important dans la preuve du chapitre précédent. Le théorème 2.2 s'applique aussi au cas présent. En utilisant la définition (3.3) et en remplaçant $g_u(x_j, x_{j+1}; \theta)$ par $g_k^{(u)}(\theta, T^k Y)$ dans la preuve du

théorème, le résultat s'écrit comme suit :

$$n^{1/2} h_n(\theta, Y) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta^0)) \,.$$

Maintenant, nous pouvons utiliser ce résultat en appliquant le développement de Taylor à l'ordre 1 au voisinage de la vraie valeur θ^0 à la fonction $n^{1/2} h_n(\theta, Y)$ pour montrer que

$$n^{-\frac{1}{2}} (\hat{\theta}_n - \theta^0) \hookrightarrow^{\mathcal{L}} \mathcal{N}(0, \mathcal{I}^{-1}(\theta^0)) \,.$$

Nous pouvons dégager un fil conducteur dans ces théorèmes. Lorsque les variables sont indépendantes, la probabilité conditionnelle qui définit la fonction $f_k(\theta, y)$ se réduit à la densité en cause et (3.2) est triviale. Dans le cas markovien, elle s'arrête à une variable du passé, ce qui explique pourquoi on a une fonction $g_u(x_j, x_{j+1}, \theta^0)$ qui dépend seulement de deux variables consécutives. Pour un modèle avec chaîne de Markov cachée, la dépendance entre les variables Y_t existe peu importe la suite d'instants considérés, mais f_k admet une limite et c'est cette dernière qui mène à l'équivalent approprié de l'information de Fisher.

3.4 Information de Fisher observée

Un des objectifs de notre étude est de trouver un estimateur de l'information de Fisher pour pouvoir faire de l'inférence sur nos estimateurs. Dans le chapitre précédent, nous avons constaté dans la section 2.4 que l'information observée converge vers l'information théorique. Ceci fut illustré et confirmé empiriquement par les simulations que nous avons résumées dans les figures 2.5, 2.6 et 2.7. Nous aimerions en faire de même ici, c'est-à-dire montrer que

$$-\frac{\partial^2}{\partial\theta_u \partial\theta_v} n^{-1} \log L(\theta; Y_1, \ldots, Y_n)|_{\theta=\hat{\theta}_n} \to \frac{\partial^2}{\partial\theta_u \partial\theta_v} H(\theta)|_{\theta=\theta^0} = \mathcal{I}_{uv}(\theta^0) \,.$$

Un peu comme dans le théorème 3.2, il suffit de suivre les calculs de la section 2.4 en remplaçant $g_{uv}(x_j, x_{j+1}, \theta^0)$ et $g_{uvw}(x_j, x_{j+1}, \theta^0)$ respectivement par

$g_k^{(uv)}(\theta^0, T^k Y)$ et $g_k^{(uvw)}(\theta^0, T^k Y)$.

Pour un modèle avec chaîne de Markov cachée, la matrice d'information de Fisher est définie comme une entropie limite, laquelle n'a pas une expression calculable comme dans les chapitres précédents. Pour cette raison, l'étude de simulation devient un peu plus compliquée, car dans les sections précédentes, nous avons comparé les valeurs de l'information observée aux vraies valeurs qui sont données par la formule calculable de l'information théorique. Dans l'étude à venir et faute de mieux, nous allons comparer les entrées de la matrice observée aux variances empiriques qui seront calculées pour un échantillon considérable d'estimateurs de vraisemblance maximale.

3.5 Calcul effectif de la fonction de vraisemblance et de l'estimateur

Pour trouver l'estimateur de vraisemblance maximale, il faut identifier les points critiques de la vraisemblance, ce qui est impossible à faire analytiquement. On doit donc se rabattre sur un calcul numérique. Mais ce dernier n'est pas aussi simple qu'on pourrait le croire *a priori*.

On observe seulement les valeurs de $\{Y_t\}$ et ce processus est une marginale de $\{Z_t\}$. La vraisemblance s'obtient donc en sommant sur toutes les valeurs possibles de la chaîne de Markov cachée. Pour simplifier l'écriture, notons $y^n = (y_1, \ldots, y_n)$ avec une convention similaire pour x^n, X^n et Y^n. Alors on a

$$\begin{aligned} L(\theta; y_1, \ldots, y_n) &= \Pr{}_\theta \{Y^n = y^n\} \\ &= \sum_{x^n \in S^n} \Pr{}_\theta \{X^n = x^n, Y^n = y^n\} \\ &= \sum_{x^n \in S^n} \Pr{}_\theta \{X^n = x^n\} \Pr{}_\theta \{Y^n = y^n \mid X^n = x^n\} \ . \end{aligned}$$

Mais

$$\Pr{}_\theta \{X^n = x^n\} = \pi_\alpha(x_1) \prod_{i=2}^{n} P_\alpha(x_{i-1}, x_i)$$

$$\Pr{}_\theta \{Y^n = y^n \mid X^n = x^n\} = \prod_{i=1}^{n} P_\beta(x_j, y_j)$$

puisque $\{X_t\}$ est une chaîne de Markov et que $P_\beta(x_i, y_i)$ est la probabilité d'observer y_i lorsque l'état caché de la chaîne est x_i. Il vient donc

$$L(\theta; y_1, \ldots, y_n) = \sum_{x_1,\ldots,x_n \in S} \pi_\alpha(x_1) P_\beta(x_1, y_1) \prod_{i=2}^{n} P_\alpha(x_{i-1}, x_i) P_\beta(x_i, y_i)\,.$$

Il s'agit d'un calcul direct de la fonction de vraisemblance. Mais, dans la pratique, cette façon de procéder nous mène à des calculs d'ordre exponentiel r^n qui sont rapidement très lourds en temps de calcul. On peut surmonter cette difficulté de deux manières : utiliser une méthode itérative pour calculer L ou faire appel à l'algorithme EM pour calculer l'estimateur (car cet algorithme s'applique typiquement aux cas des données partiellement observées).

3.5.1 Méthode itérative

Nous décrivons un algorithme récursif et rapide qui permet de réduire le temps de calcul de la vraisemblance. Il s'agit de l'algorithme de (Hamilton, 1987) qui fut l'un des premiers à proposer une méthode de calcul de la vraisemblance dans le contexte d'une étude économétrique. Dans la suite, nous allons montrer les différentes étapes de cet algorithme pour une séquence observée y^n. La fonction de log-vraisemblance s'écrit

$$\log L(y_1, \ldots, y_n; \theta) = \sum_{k=1}^{n} \log \Pr{}_\theta \{Y_k = y_k \mid Y^{k-1} = y^{k-1}\}$$

avec la convention que $y^k = (y_1, \ldots, y_k)$ (de même pour x^k, X^k et Y^k) et que le premier terme de la somme se réduit à la marginale de Y_1. Tout d'abord, on a

$$\begin{aligned}\Pr{}_\theta\{Y_k = y_k \mid Y^{k-1} = y^{k-1}\} \\ &= \sum_{x \in S} \Pr{}_\theta\{X_{k-1} = x \mid Y^{k-1} = y^{k-1}\} \\ &\qquad \times \Pr{}_\theta\{Y_k = y_k \mid X_{k-1} = x, Y^{k-1} = y^{k-1}\}\,.\end{aligned}$$

Il vient ensuite

$$\begin{aligned}\Pr{}_\theta\{Y_k = y_k \mid X_{k-1} = x, Y^{k-1} = y^{k-1}\} \\ &= \sum_{x' \in S} \Pr{}_\theta\{X_k = x', Y_k = y_k \mid X_{k-1} = x, Y^{k-1} = y^{k-1}\} \\ &= \sum_{x' \in S} \Pr{}_\theta\{X_k = x', Y_k = y_k \mid X_{k-1} = x, Y_{k-1} = y_{k-1}, Y^{k-2} = y^{k-2}\} \\ &= \sum_{x' \in S} \Pr{}_\theta\{Z_k = (x', y_k) \mid Z_{k-1} = (x, y_{k-1}), Y^{k-2} = y^{k-2}\} \\ &= \sum_{x' \in S} \Pr{}_\theta\{Z_k = (x', y_k) \mid Z_{k-1} = (x, y_{k-1})\} \\ &= \sum_{x' \in S} P_\alpha(x, x') P_\beta(x', y_k)\,.\end{aligned}$$

Que l'on puisse éliminer l'évènement $\{\, Y^{k-2} = y^{k-2} \,\}$ en cours de route est justifié par le fait que $\{Z_t\}$ est une chaîne de Markov. Pour faciliter le calcul de $\Pr{}_\theta\{X_{k-1} = x \mid Y^{k-1} = y^{k-1}\}$, on introduit deux notations :

$$\begin{aligned} v_k(x) &= \Pr{}_\theta\{\, X_k = x \mid Y^k = y^k \,\} \\ w_k(x') &= \Pr{}_\theta\{\, Y_{k+1} = y_{k+1}, X_{k+1} = x' \mid Y^k = y^k \,\} \\ &= \sum_{x \in S} v_k(x) \Pr{}_\theta\{\, Y_{k+1} = y_{k+1}, X_{k+1} = x' \mid Y^k = y^k, X_k = x \,\} \\ &= \sum_{x \in S} v_k(x) P_\alpha(x, x') P_\beta(x', y_{k+1})\,. \end{aligned}$$

On peut calculer de manière récursive $\Pr_\theta\{X_{k-1} = x_{k-1} \mid Y^{k-1} = y^{k-1}\}$ comme suit. On observe d'abord que

$$\begin{aligned} v_k(x) &= \Pr{}_\theta\{\, X_k = x \mid Y^k = y^k \,\} \\ &= \Pr{}_\theta\{\, X_k = x \mid Y_k = y_k, Y^{k-1} = y^{k-1} \,\} \\ &= \frac{\Pr_\theta\{\, X_k = x, Y_k = y_k \mid Y^{k-1} = y^{k-1} \,\}}{\Pr_\theta\{\, Y_k = y_k \mid Y_{k-1} = y_{k-1} \,\}} \\ &= \frac{w_{k-1}(x)}{\sum_{x' \in S} w_{k-1}(x')} \end{aligned}$$

tandis que

$$\begin{aligned} &\Pr{}_\theta\{Y_k = y_k \mid Y^{k-1} = y^{k-1}\} \\ &\qquad = \sum_{x \in S} \Pr{}_\theta\{X_{k-1} = x \mid Y^{k-1} = y^{k-1}\} \\ &\qquad\qquad\qquad \times \Pr{}_\theta\{Y_k = y_k \mid X_{k-1} = x, Y^{k-1} = y^{k-1}\} \\ &\qquad = \sum_{x \in S} v_{k-1}(x) \left(\sum_{x' \in S} P_\alpha(x, x') P_\beta(x', y_k) \right) \\ &\qquad = \sum_{x' \in S} \left(\sum_{x \in S} v_{k-1}(x) P_\alpha(x, x') P_\beta(x', y_k) \right) \\ &\qquad = \sum_{x' \in S} w_{k-1}(x') \end{aligned}$$

On déduit l'algorithme récursif de calcul ci-dessous,

$v_0 \leftarrow \pi_\alpha$ (la loi stationnaire de P_α)

pour $j = 1, 2, \ldots, n$

$w_{k-1}(x') \leftarrow \sum_{x \in S} v_{k-1}(x) P_\alpha(x, x') P_\beta(x', y_k), \quad x' \in R$

$\Pr_\theta\{\, Y_k = y_k \mid Y^{k-1} = y^{k-1} \,\} \leftarrow \sum_{x' \in S} w_{k-1}(x')\,;$

$v_k(x) \leftarrow w_{k-1}(x) / \sum_{x' \in S} w_{k-1}(x'), \quad x \in S\,;$

fin

$$\log L \leftarrow \sum_{k=1}^{n} \log \Pr{}_\theta\{\, Y_k = y_k \mid Y^{k-1} = y^{k-1} \,\}$$

Cet algorithme calcule la valeur de la fonction de log-vraisemblance pour une matrice de transition P_α donnée, une matrice d'émission P_β donnée et une série d'observations $(y_1, \ldots, y_n)$ de $\{Y_t\}$.

3.5.2 Algorithme EM

Les modèles avec chaîne de Markov cachée (HMM) peuvent servir à de multiples applications et pour cela trois types d'estimation ont été développés dans la littérature. Pour un modèle dont on connaît les paramètres ainsi qu'une séquence observée, nous pouvons nous demander : quelle est la probabilité d'apparition de cette séquence ? L'algorithme *backward-forward* introduit par (Rabiner et Juang, 1986) permet de faire un calcul progressif et rétrograde pour déterminer la probabilité d'apparition de la séquence. Un exemple de cet usage : dans un cas d'un HMM entraîné à reconnaître quelque chose (langue d'un texte par exemple), c'est la question qu'on se pose souvent.

Un deuxième cas d'application est le suivant : pour une séquence d'observations provenant d'un HMM dont on connaît les paramètres, on cherche la séquence d'état cachés génératrice de la séquence observée. C'est la question qu'on se pose souvent dans les logiciels et systèmes de reconnaissance vocale. L'algorithme de Viterbi introduit par (Viterbi, 1967) et (Omura, 1969) utilise la programmation dynamique pour maximiser la probabilité *a posteriori* de la séquence cachée sachant la séquence observée. Cette méthode permet d'obtenir à la fois la probabilité maximale et la séquence d'états cachés associés.

Le dernier cas d'usage que nous allons illustrer est le suivant : nous avons un HMM dont nous connaissons le nombre s d'états cachés, le nombre r d'états observables et une séquence d'observations de longueur n. Nous nous demandons quelles sont les matrices P_α et P_β qui maximisent la fonction de vraisemblance. L'algorithme

Espérance-Maximisation (EM) appliqué à un HMM est une méthode itérative développée par (Baum, 1970) qui répond à ce besoin. Celui-ci permet de calculer l'estimateur de vraisemblance maximale sans effectuer l'optimisation numérique de la vraisemblance (calculée par exemple avec la méthode itérative de la sous-section 3.5.1).

L'algorithme est très connu dans la littérature, il est utilisé lorsque certaines données sont manquantes pour calculer le MLE. Dans le cas des HMM, les données manquantes sont les états de la chaîne cachée $\{X_t\}$. L'algorithme EM comporte deux étapes. Premièrement, on calcule l'espérance conditionnelle de la log-vraisemblance des données manquantes en utilisant les valeurs courantes des paramètres. Cette étape est dénotée par E et s'écrit :

$$\mathcal{Q}(\theta, \theta^{(i-1)}) = \mathbb{E}\left[\log \Pr(Y, X \mid \theta) \mid Y, \theta^{(i-1)}\right]$$

où $\theta^{(i-1)}$ sont les paramètres estimés courants $(\hat{\alpha}, \hat{\beta})$ utilisés pour évaluer l'espérance. Deuxièmement, c'est l'étape M, on maximise $\mathcal{Q}$ par rapport à θ. Les paramètres obtenus, à la fin de l'étape M, sont donc les paramètres maximisant l'espérance de l'étape E :

$$\theta^{(i)} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{(i-1)})$$

Ces deux étapes sont répétées jusqu'à ce qu'il y ait convergence des estimateurs.

L'algorithme EM est souvent facile à programmer et la convergence est assurée sous des conditions légères. Par contre, l'algorithme ne permet pas *a priori* de produire une estimation de la matrice de variance-covariance des paramètres estimés (asymptotique ou non). De plus, dans certains cas, l'étape E est impossible à effectuer analytiquement.

Dans ce mémoire, nous avons opté pour la méthode récursive de (Hamilton, 1987) pour estimer les paramètres du modèle HMM. Nous avons fait ce choix pour deux

raisons. Premièrement, la méthode est facile à comprendre et à coder. Deuxièmement, puisque la méthode calcule la vraisemblance, cette dernière peut être dérivée numériquement pour obtenir un estimateur de la matrice de variance-covariance. Cependant, l'algorithme EM sera utilisé pour trouver des valeurs initiales des paramètres pour notre routine d'optimisation de la fonction de vraisemblance.

3.6 Étude par simulation

Cette partie de simulation est un peu différente de celle des chapitres précédents. La différence réside dans le fait que nous allons devoir passer par des calculs numériques pour obtenir le MLE ainsi que la matrice de variance-covariance estimée. Pour ne pas alourdir le travail de simulation, nous allons tirer des conclusions générales des distributions des estimations plutôt que d'en faire des tests de comparaison systématiques.

3.6.1 Calcul numérique de l'estimateur

Pour commencer, nous avons mentionné dans la section 3.5.2 que nous allons utiliser l'algorithme EM pour trouver les valeurs initiales de notre routine d'optimisation numérique de la fonction de log-vraisemblance. L'algorithme EM est connu par sa convergence vers un maximum local qui dépend en lui-même des valeurs initiales choisies. Pour maximiser les chances que l'algorithme EM nous fournisse des bonnes probabilités de départ, nous avons établi le processus suivant :

1. nous simulons aléatoirement N matrices de probabilités initiales P_α et P_β sous les contraintes

$$P_\alpha(i,1) + \cdots + P_\alpha(i,s) = 1 \text{ et } P_\alpha(i,j) > 0, \qquad i,\, j \in S$$
$$P_\beta(\ell,1) + \cdots + P_\beta(\ell,r) = 1 \text{ et } P_\beta(\ell,m) > 0, \qquad \ell \in S,\ m \in R;$$

2. nous simulons une séquence observée de la chaîne de Markov cachée avec les vraies matrices de probabilités P_{α^0} et P_{β^0}. Le vecteur est de taille petite ou moyenne (entre 100 à 200) ;
3. nous calculons la valeur de la fonction de vraisemblance par l'algorithme récursif (section 3.5.1) pour chaque matrice initiale simulée avec le vecteur d'observation de l'étape 2 ;
4. pour chacune des N matrices initiales correspondant à 10 % des valeurs les plus élevées de la fonction de vraisemblance, nous lançons l'algorithme EM avec le même vecteur d'observation simulé qu'à l'étape 2 ;
5. nous choisissons l'estimation EM finale la plus fréquente [1] comme valeur de départ de notre routine d'optimisation.

Ces premières étapes de recherche de la valeur initiale sont lancées une seule fois et les valeurs de départ résultantes sont utilisées dans toutes les simulations qui suivent.

Une fois la valeur initiale obtenue, nous lançons dans un deuxième temps la routine d'optimisation qui se résume en deux étapes :

1. nous utilisons sur le langage R la fonction **nlminb** pour optimiser la fonction de vraisemblance. Cette fonction d'optimisation numérique prend en entrée la fonction qui calcule la vraisemblance, les probabilités (matrices) de départ et le vecteur des séquences observées ;
2. nous rajoutons ensuite une condition : aux résultats de l'étape 1, la dérivée numérique de la fonction de log-vraisemblance est quasi-nulle.

Les algorithmes de ces étapes sont disponibles en appendice B.

---

1. Nous avons remarqué que généralement les estimations EM convergent en moyenne de 10% à 20% des cas à la même solution.

Comme nous soupçonnons que le comportement de l'estimateur de vraisemblance maximale sera différent pour différentes matrices d'émission, deux cas de HMM seront simulés. Dans le premier cas, nous simulons des vecteurs d'observations qui proviennent d'une matrice d'émission P_{β^0} pour laquelle, dans chaque état caché, il existe un état plus probable que les autres :

$$P_{\alpha_1^0} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix} \quad \text{et} \quad P_{\beta_1^0} = \begin{pmatrix} 0.95 & 0.03 & 0.02 \\ 0.07 & 0.90 & 0.03 \end{pmatrix} .$$

Dans le deuxième cas, nous considérons un peu l'inverse du premier, à savoir, les probabilités d'émission sont plus aux moins proches entre elles, dans chaque état caché :

$$P_{\alpha_2^0} = \begin{pmatrix} 0.77 & 0.23 \\ 0.26 & 0.74 \end{pmatrix} \quad \text{et} \quad P_{\beta_2^0} = \begin{pmatrix} 0.59 & 0.27 & 0.14 \\ 0.25 & 0.51 & 0.24 \end{pmatrix}$$

Intuitivement, nous nous attendons à ce que, dans le premier cas, les estimateurs soient plus faciles à calculer par notre routine d'optimisation que dans le deuxième. Dans les deux cas, les estimations ML sont des vecteurs de longueur 6, où p_1, p_2 sont les probabilités de la matrice de transition et $p_3, \ldots, p_6$ sont les probabilités de la matrice d'émission.

$$P_{trans} = \begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix} \quad \text{et} \quad P_{emiss} = \begin{pmatrix} p_3 & p_4 & 1-p_3-p_4 \\ p_5 & p_6 & 1-p_5-p_6 \end{pmatrix}$$

Pour vérifier les propriétés asymptotiques des estimations ML, nous allons procéder comme dans les chapitres précédents. À chaque fois, 1000 estimateurs sont calculés pour des séquences simulées de la chaîne de Markov cachée, pour des tailles petite, moyenne et grande. Ceci est fait pour les deux exemples que nous avons mentionnés. Les figures 3.1, 3.2 et 3.3 résument les résultats pour le premier cas ; les figures 3.4, 3.5 et 3.6 les présentent pour le deuxième.

Pour la convergence des estimateurs, on remarque le même phénomène que dans les chapitres I et II : plus la taille de l'échantillon est grande plus la distance

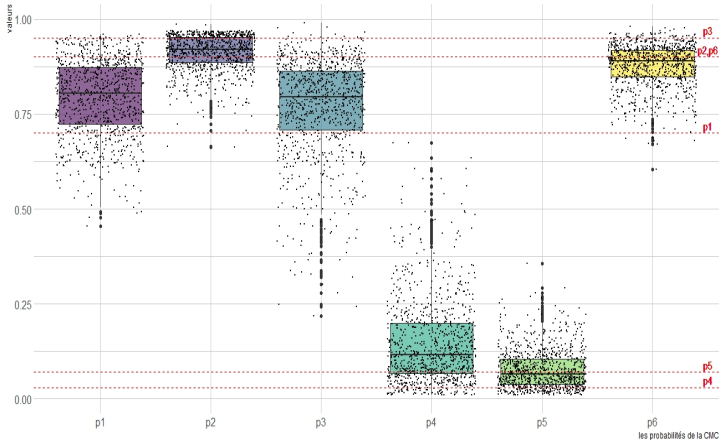

Figure (3.1) Distributions des 1000 estimateurs simulés avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ $(n = 50)$.

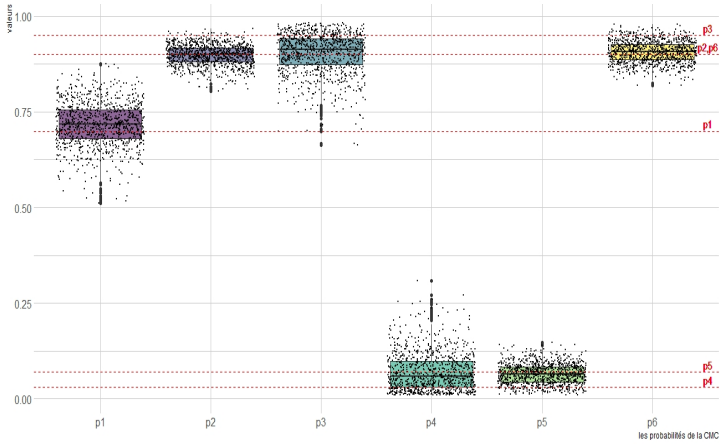

Figure (3.2) Distributions des 1000 estimateurs simulés avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ $(n = 500)$.

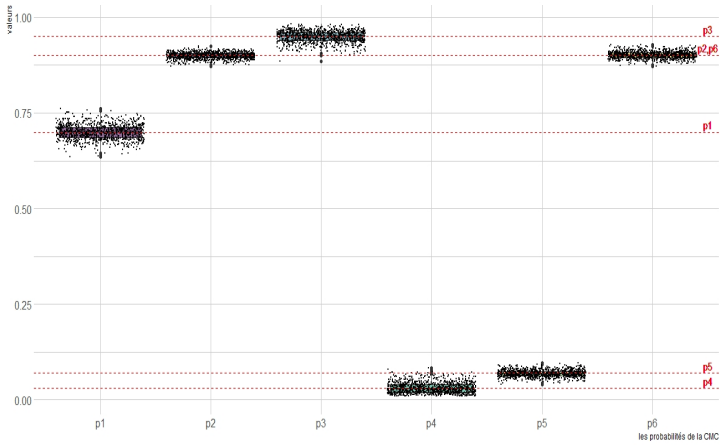

Figure (3.3) Distributions des 1000 estimateurs simulés avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ $(n = 5000)$.

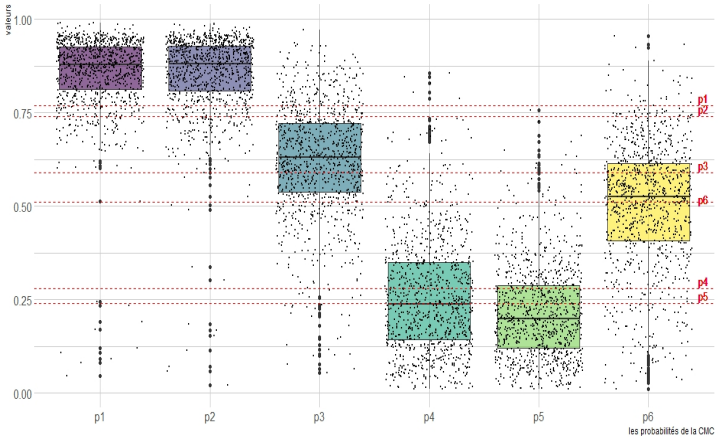

Figure (3.4) Distributions des 1000 estimateurs simulés avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ $(n = 50)$.

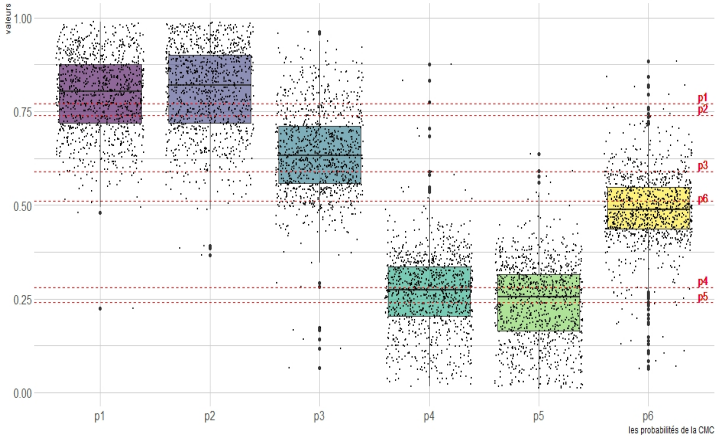

Figure (3.5) Distributions des 1000 estimateurs simulés avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ $(n = 500)$.

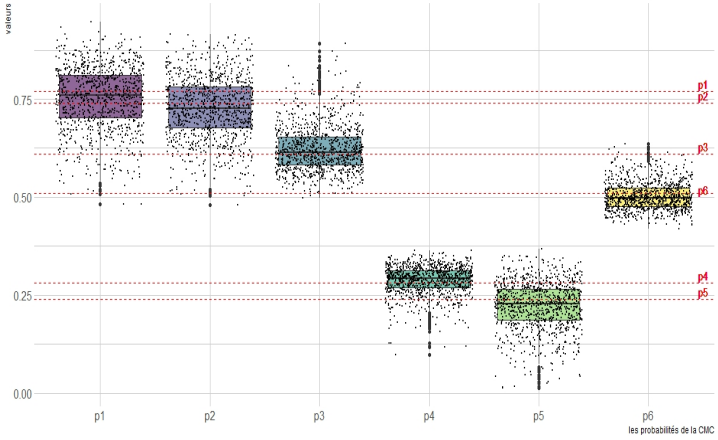

Figure (3.6) Distributions des 1000 estimateurs simulés avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ $(n = 10000)$.

entre la médiane des distributions des estimations est proche des vraies valeurs des paramètres. La même remarque peut être faite pour la normalité asymptotique, on voit que plus la taille de l'échantillon augmente plus les boîtes à moustaches semblent symétriques. Par contre, on remarque que les propriétés asymptotiques sont plus rapidement vérifiées pour le premier cas que pour le deuxième. Ce phénomène peut être expliqué par le facteur suivant : dans le premier cas, la probabilité que la chaîne soit à l'état caché 1 et l'état observable 1 (ou l'état caché 2 et l'observation 2) est plus élevée, au point que les autres états observables ont des probabilités presque nulles. Alors, si on observe 1, on sait qu'on sera dans la plupart du temps dans l'état caché 1 et c'est la même chose aussi pour l'état caché 2. Ceci permet à la méthode de vraisemblance maximale de converger plus facilement, même avec des échantillons de tailles petite ou moyenne. Par contre, le phénomène n'est pas observé dans le deuxième cas. Il s'avère que la méthode ML a besoin d'une séquence d'observations assez longue pour fournir un bon estimateur.

Les autres aspects que nous avons examinés dans les chapitres précédents sont le biais et la variance des estimateurs. Il semble que le biais est considérable pour les estimateurs calculés avec des échantillons de tailles petite, moyenne et même grande pour le deuxième cas. Ceci est confirmé dans les tableaux 3.1 et 3.2 dans lesquels on remarque que le biais tend vers 0 rapidement pour le premier cas et lentement pour le deuxième. En ce qui concerne la variance, celle-ci semble aussi diminuer quand la taille de l'échantillon est grande. On analysera plus en détail cette question dans la prochaine simulation.

3.6.2 Analyse de la variance estimée par l'information observée

L'estimateur de la variance est calculé, comme dans les chapitres précédents, en inversant l'information de Fisher observée. Rappelons la particularité du cas des

n	Biais($\hat{p}_1$)	Biais($\hat{p}_2$)	Biais($\hat{p}_3$)	Biais($\hat{p}_4$)	Biais($\hat{p}_5$)	Biais($\hat{p}_6$)
50	0.13213	0.01300	0.18945	1.3411	0.10092	0.02569
500	0.02440	0.00288	0.04992	0.78879	0.06897	0.00691
1000	0.01168	0.00184	0.02578	0.03922	0.06479	0.00534
5000	0.00039	0.000992	0.00116	0.030412	0.01788	0.00130
7000	0.00016	0.000352	0.00099	0.009563	0.00638	0.00041

Tableau (3.1) Biais moyen de 1000 estimateurs de $P_{\alpha_1^0}$ et $P_{\beta_1^0}$.

n	Biais($\hat{p}_1$)	Biais($\hat{p}_2$)	Biais($\hat{p}_3$)	Biais($\hat{p}_4$)	Biais($\hat{p}_5$)	Biais($\hat{p}_6$)
50	0.11247	0.1528	0.02171	0.07327	0.04631	0.10281
500	0.03231	0.08647	0.08223	0.03518	0.06491	0.03885
5000	0.03178	0.00267	0.07542	0.01253	0.05275	0.01700
10000	0.02325	0.001673	0.05581	0.0084	0.001160	0.01543

Tableau (3.2) Biais moyen de 1000 estimateurs de $P_{\alpha_2^0}$ et $P_{\beta_2^0}$.

chaînes de Markov cachées qui est l'absence d'une formule explicite pour les variances théoriques comme dans les chapitres précédents. Alors, nous allons comparer les variances estimées avec les variances empiriques qui sont calculées sur les 1000 MLE simulés dans chacun des cas. Les résultats sont rassemblés dans les figures, 3.7 jusqu'à 3.12.

Pour les deux cas ainsi que les différentes tailles de l'échantillon, on remarque, d'après les figures le même phénomène que dans les chapitres précédents : la variance des estimations ML diminue lorsque la taille de l'échantillon augmente. Par contre, la variation des MLE du premier cas est inférieure à celle du deuxième, ce qui est sûrement dû à la précision des estimations dans le premier cas par rapport au deuxième. Un deuxième élément à remarquer est que les médianes

des boîtes à moustaches (pour les deux cas) ne semblent pas très éloignées des variances empiriques qui sont présentées dans les tableaux 3.3 et 3.4.

Un autre élément visible dans ces figures est le biais des estimateurs de la variance. Celui-ci paraît très important pour les échantillons de petite taille (inférieure ou

n	$V(\hat{p}_1)$	$V(\hat{p}_2)$	$V(\hat{p}_3)$	$V(\hat{p}_4)$	$V(\hat{p}_5)$	$V(\hat{p}_6)$
50	0.00983	0.00225	0.01624	0.0126	0.0026	0.00317
500	0.00336	0.00073	0.00280	0.00240	0.00070	0.00081
5000	0.00038	0.000070	0.00023	0.00019	0.00008	0.00008

Tableau (3.3) Variance empirique des estimateurs $\hat{p}_i$ pour le premier cas.

n	$V(\hat{p}_1)$	$V(\hat{p}_2)$	$V(\hat{p}_3)$	$V(\hat{p}_4)$	$V(\hat{p}_5)$	$V(\hat{p}_6)$
50	0.01087	0.01386	0.0217	0.0241	0.0155	0.0292
500	0.01035	0.01182	0.0131	0.0013	0.0111	0.0092
10000	0.00630	0.00930	0.0034	0.00240	0.0036	0.0012

Tableau (3.4) Variance empirique des estimateurs $\hat{p}_i$ pour le deuxième cas.

égale à 50) dans le cas 1 et petite et moyenne pour le cas 2. On remarque, dans les figures, l'existence de certains points aberrants qui sont très loin de la médiane des distributions. À titre d'exemple, dans la figure 3.10, on voit une estimation de la variance de $\hat{p}_1$ qui est proche de 4 sachant que la médiane est au alentour de 0.01.

Le biais moyen des estimations de la variance pour chaque taille d'échantillon du cas 2 est représenté dans le tableau 3.5. On voit que le biais est très considérable pour les petites tailles d'échantillon et que celui-ci diminue quand la taille de l'échantillon devient grande. Maintenant nous aimerions connaître l'effet du

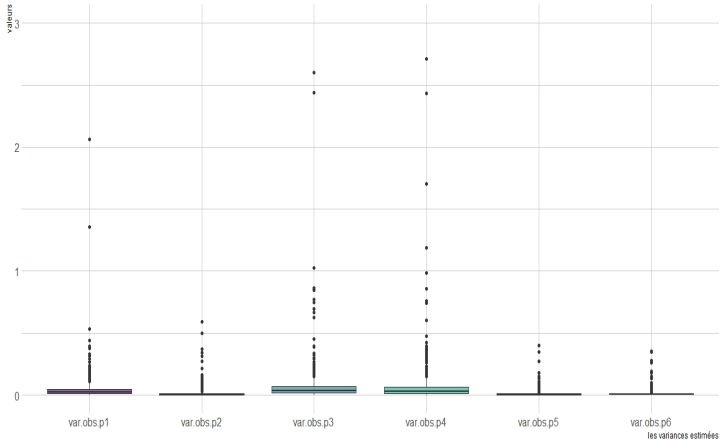

Figure (3.7) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ et $n = 50$).

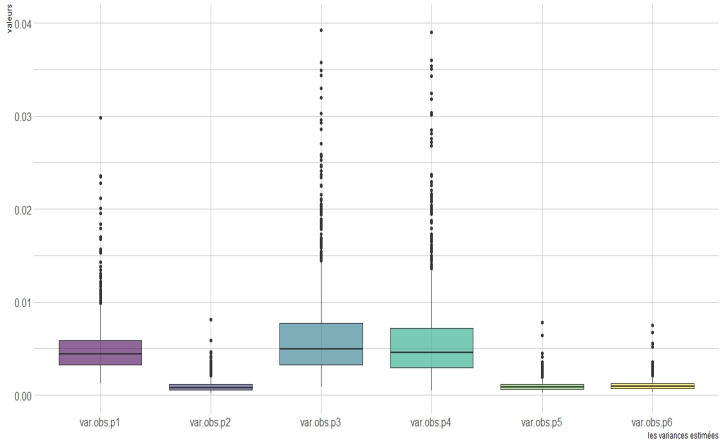

Figure (3.8) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ et $n = 500$).

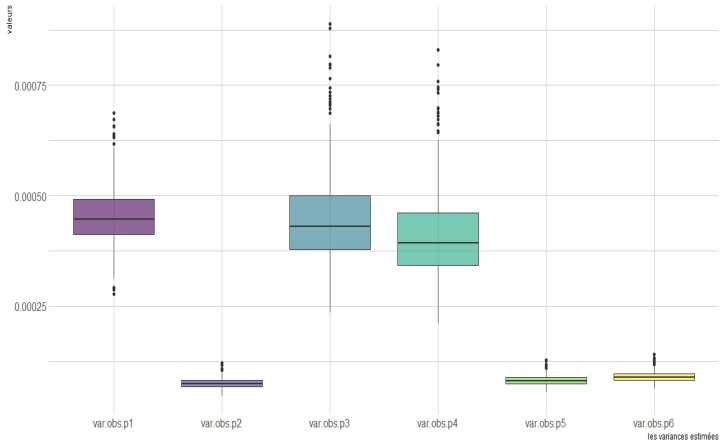

Figure (3.9) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_1^0}$ et $P_{\beta_1^0}$ et $n = 500$).

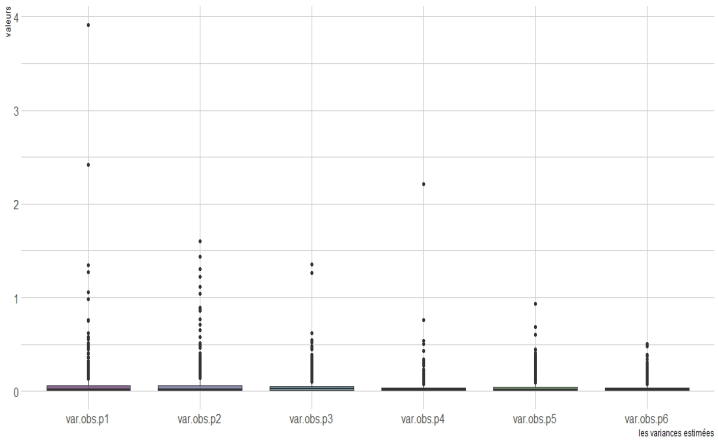

Figure (3.10) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ et $n = 50$).

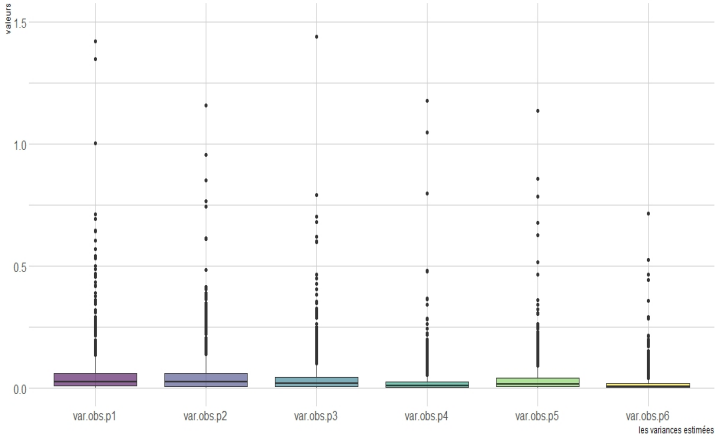

Figure (3.11) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ et $n = 500$).

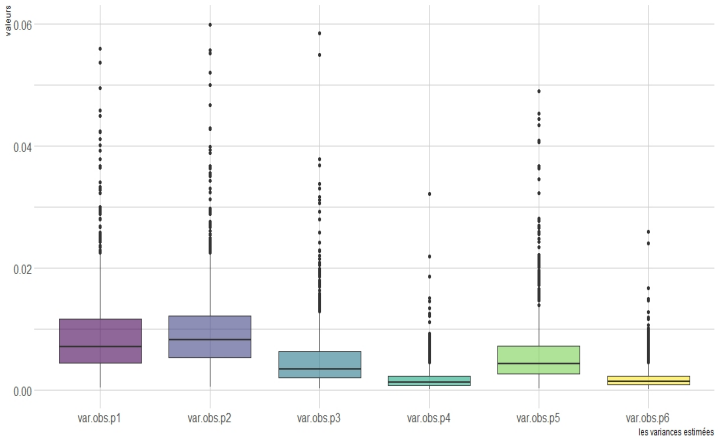

Figure (3.12) Distributions des 1000 variances estimées (simulé avec $P_{\alpha_2^0}$ et $P_{\beta_2^0}$ et $n = 10000$).

n	$B(\hat{V}_{\hat{p}_1})$	$B(\hat{V}_{\hat{p}_2})$	$B(\hat{V}_{\hat{p}_3})$	$B(\hat{V}_{\hat{p}_4})$	$B(\hat{V}_{\hat{p}_5})$	$B(\hat{V}_{\hat{p}_6})$
50	3.610	2.273	2.9454	2.700	1.695	1.122
500	1.440	1.678	1.2909	0.9544	0.4045	0.756
10000	0.483	0.652	0.5323	0.2102	0.174	0.382

Tableau (3.5) Biais moyen des variances estimées de $\hat{p}_i$ pour le deuxième cas.

biais des variances estimées sur les intervalles de confiance et les probabilités de couverture. Les résultats des deux cas sont résumés dans les tableaux ci-dessous.

n	$P_c(\hat{p}_1)$	$P_c(\hat{p}_2)$	$P_c(\hat{p}_3)$	$P_c(\hat{p}_4)$	$P_c(\hat{p}_5)$	$P_c(\hat{p}_6)$
50	0.81	0.946	0.948	0.956	0.944	0.946
500	0.92	0.949	0.952	0.963	0.960	0.955
5000	0.96	0.953	0.956	0.973	0.974	0.964

Tableau (3.6) Probabilités de couverture selon la taille de l'échantillon (cas 1).

n	$P_c(\hat{p}_1)$	$P_c(\hat{p}_2)$	$P_c(\hat{p}_3)$	$P_c(\hat{p}_4)$	$P_c(\hat{p}_5)$	$P_c(\hat{p}_6)$
50	0.799	0.792	0.948	0.883	0.890	0.91
500	0.876	0.930	0.970	0.870	0.945	0.929
10000	0.950	0.958	0.972	0.947	0.953	0.955

Tableau (3.7) Probabilités de couverture selon la taille de l'échantillon (cas 2).

Enfin, d'après cette simulation, on peut tirer une conclusion générale commune avec les chapitres précédents qui est : la théorie asymptotique est vérifiée aussi pour les estimations ML dans le cas d'une dépendance de Markov cachée. À travers les deux cas que nous avons étudiés, nous avons remarqué que ces propriétés sont plus rapidement vérifiées pour le cas d'une chaîne de Markov cachée avec

des probabilités d'états plus au moins distinguables. Par ailleurs, plus ces probabilités sont proches entre elles, plus l'estimation ML nécessite des séquences d'observations de grande taille pour pouvoir bien capturer l'information.

De plus, comme nous avons obtenu des probabilités de couverture inférieures à 90% pour la majorité des probabilités d'états du cas 2, on peut dire que les estimations sont moins bonnes dans ces cas. Quant à la variance observée, on peut aussi dire que cet estimateur est parfois biaisé pour les petites tailles d'échantillon et que ce biais s'approche de zéro, lorsque la taille de la séquence d'observations devient grande.

CONCLUSION

Dans un article pionnier de la méthodologie statistique, publié en 1922, Ronald A. Fisher souligne qu'il est important de trouver, si possible, des estimateurs convergents et exhaustifs, dont on peut connaître (ou du moins approcher) la distribution. L'objectif de ce mémoire était d'étudier ces questions pour l'estimateur de vraisemblance maximale.

Pour des modèles de complexité croissante, nous avons évalué la convergence et la normalité asymptotique de l'estimateur, d'abord du point de vue théorique, pour ensuite illustrer ces propriétés par simulation de Monte Carlo. Nous avons aussi examiné le problème de l'estimation de la variance de l'estimateur de vraisemblance maximale et nous avons découvert le rôle fondamental de l'information de Fisher dans ce contexte.

On a aussi vu les difficultés techniques que soulève la démonstration des propriétés asymptotiques, même dans le cas classique des variables indépendantes et parentes. Pour exposer ces questions avec rigueur, il faut savoir manier avec doigté la convergence stochastique et l'analyse en général. Ainsi, plutôt que de tomber dans un formalisme pesant, nous avons tenté de rester pédagogique et d'exposer les idées plus que les détails, en particulier lorsque ces derniers devenaient trop lourds.

En ce qui concerne les résultats directs que nous pouvons dégager de cette étude, on a vu que les deux propriétés asymptotiques (convergence et normalité des estimations) sont vérifiées mathématiquement pour les trois modèles, et que l'information observée converge vers la matrice d'information de Fisher asymptoti-

quement. Quant aux simulations, elles nous ont montré que la vitesse de convergence vers ces propriétés asymptotiques diminue quand les modèles sont de plus en plus complexes. À ceci, s'ajoute la difficulté empirique de trouver des estimateurs convergents dans les modèles de Markov cachée pour lesquels nous avons pu construire une méthode de simulation qui nous permet à la fois d'avoir des estimateurs convergents et de calculer l'information observée.

Les modèles que nous avons étudiés sont parmi les premiers considérés dans la littérature statistique. Il va de soi que l'estimation de vraisemblance maximale a depuis été employée pour bien d'autres modèles, tous plus complexes les uns que les autres. Mais pour ces modèles, plus que jamais, les distributions exactes sont inconnues et la statistique asymptotique est alors incontournable.

S'il est une leçon que ce mémoire nous a appris, c'est que la statistique asymptotique, bien qu'essentielle, pose des difficultés de rédaction non négligeables et que le calcul effectif de l'estimateur de vraisemblance maximale et de sa variance asymptotique estimée n'est pas toujours évident. Nous croyons que ces sujets mériteraient d'être exposés en profondeur et avec soin dans le cas d'autres modèles stochastiques (comme les modèles à changement de régime et les modèles de Markov à temps continu), en ayant au premier chef, une volonté de clarté et de simplicité. La statistique, ou du moins ceux qui veulent l'apprendre, y gagnerait beaucoup.

APPENDICE A

L'objectif de cet appendice est de présenter certains théorèmes et résultats que nous avons utilisés au chapitre I et le chapitre II.

A.1 Théorèmes et résultats du chapitre I

Pour montrer la loi forte des grands nombres uniformes, nous avons besoin d'introduire quelques théorèmes largement connus dans la littérature statistique.

Théorème A.1. Convergence dominée

Soit f_n, $n \in \mathbb{N}$, des fonctions mesurables sur un ensemble $\mathcal{E}$ de $\mathbb{R}$, telles que :

- $\lim_{n\to\infty} f_n(x)$ existe pour tout $x \in \mathcal{E}$;
- il existe une fonction intégrable g telle que pour tout entier naturel n :

$$|f_n(x)| \leq g(x), \quad x \in \mathcal{E}\,.$$

Alors, il existe une fonction f telle que f_n converge vers f presque partout, et

$$\lim_{n\to\infty} \int_{\mathcal{E}} f_n(x)dx = \int_{\mathcal{E}} f dx\,.$$

Théorème A.2. Lois forte des grands nombres

Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires indépendantes et parentes. Posons $S_n = X_1 + X_2 + \cdots + X_n$ et supposons que $\mathbb{E}\,[\,|X_1|\,] < \infty$. Alors S_n/n converge persque sûrement vers $\mathbb{E}\,[\,X_1\,]$.

Nous avons besoin de rappeler la notion de convergence uniforme.

Définition A.1. Convergence uniforme

Soit I un intervalle de $\mathbb{R}$, f_n une suite de fonctions définies sur I, et f une fonction définie sur I. On dit que f_n *converge uniformément* vers f sur I si

$$\forall\, \epsilon > 0,\, \exists\, n_0 \in \mathbb{N},\, \forall\, x \in I,\, \forall\, n \geq n_0,\, |f_n(x) - f(x)| < \epsilon\,.$$

En d'autres mots, si on note

$$||f_n - f||_\infty = \sup\{\, |f_n(x) - f(x)| :\, x \in I\,\},$$

alors f_n converge uniformément vers f si l'on a $||f_n - f||_\infty \to 0$.

Ces premiers résultats nous permettent de démontrer une loi forte des grands nombres uniformes qui s'énonce comme suit :

Théorème A.3. Loi forte des grands nombres uniforme

Soit X_1, $X_2, X_3, \ldots$, des variables aléatoires indépendantes, chacune de même loi $F(x;\theta)$ et soit $U(x;\theta)$ une fonction continue en x pour tout $\theta \in \Theta$. Supposons que :

1. l'espace des paramètres Θ est compact ;
2. $U(x;\theta)$ est semi-continue supérieurement en θ pour tout x ;
3. il existe une fonction $K(x)$ tel que, $\mathbb{E}\,[K(X)] < \infty$ et $U(x;\theta) \leq K(x)$ pour tout x et θ ;
4. pour tout $\theta \in \Theta$ et pour tout $\rho > 0$ suffisamment petit, $\sup_{|\theta' - \theta| < \rho} U(x;\theta')$ est mesurable en x.

Dans ces conditions, on a

$$\Pr\left\{ \limsup_n \sup_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} U(X_j;\theta) \leq \sup_{\theta \in \Theta} \mu(\theta) \right\} = 1\,.$$

Démonstration. Commençons par poser

$$\varphi(x;\theta;\rho) = \sup_{|\theta'-\theta|<\rho} U(x;\theta') \tag{A.1}$$

avec ρ assez petit ($\rho \to 0$) pour que θ' soit accès proche de θ. Nous nous intéressons dans un premier temps au comportement de la fonction φ quand $\rho \to 0$. Remarquons que :

(a) la fonction $\varphi(x;\theta;\rho)$ est mesurable d'après la condition 4 ;

(b) la fonction $\varphi(x;\theta;\rho)$ est bornée supérieurement par une fonction intégrable d'après la condition 3 ;

(c) $\varphi(x;\theta;\rho) \longrightarrow U(x;\theta)$ quand $\rho \to 0$ par (A.1).

Les propriétés (a), (b) et (c) sont les conditions du théorème A.1. De ceci, on peut déduire le résultat de convergence limite suivant :

$$\int \varphi(x;\theta;\rho)\, dF(x) \longrightarrow \int U(x;\theta)\, dF(x) = \mu(\theta) \qquad \text{quand} \quad \rho \to 0. \tag{A.2}$$

Ce premier résultat nous montre que l'espérance de la fonction φ converge vers l'espérance de U quand $\rho \to 0$. Fixons $\varepsilon > 0$ et pour chaque θ, choisissons $\rho_\theta > 0$ tel que

$$\int \varphi(x;\theta;\rho_\theta) dF(x) < \mu(\theta) + \varepsilon\,. \tag{A.3}$$

Les boules $S(\theta;\rho_\theta) = \{\theta' : |\theta - \theta'| < \rho_\theta\}$ forment un recouvrement de Θ. Puisque Θ est compact, il existe un sous-recouvrement fini :

$$\Theta \subset \bigcup_{k=1}^{m} S(\theta_k;\rho_{\theta_k})\,.$$

Cette inclusion signifie que, pour chaque θ, il existe un indice $k \in \{1,\ldots,m\}$ tel que $\theta \in S(\theta_k;\rho_{\theta_k})$. Par définition de φ, on a alors $U(x,\theta) \le \varphi(x,\theta_k,\rho_{\theta_k})$ pour ce θ et pour tout x. Donc on a

$$\frac{1}{n}\sum_{j=1}^{n} U(X_j;\theta) \le \frac{1}{n}\sum_{j=1}^{n} \varphi(X_j;\theta_k;\rho_{\theta_k})$$

de sorte que

$$\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}U(X_j;\theta)\leq\sup_{1\leq k\leq m}\frac{1}{n}\sum_{j=1}^{n}\varphi(X_j;\theta_k;\rho_{\theta_k})\,. \tag{A.4}$$

On applique la loi forte des grands nombres à la suite $\{\varphi(X_j;\theta_k;\rho_{\theta_k}),\ j\geq 1\}$ pour chaque k. Comme il n'y a qu'un nombre fini d'entiers k, on peut écrire

$$\Pr\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\varphi(X_j;\theta_k;\rho_{\theta_k})\leq\mu(\theta_k)+\varepsilon,\ k=1,\ldots,m\right\}=1$$

à cause de (A.2) et (A.3). On en déduit que

$$\Pr\left\{\limsup_{n}\sup_{1\leq k\leq m}\frac{1}{n}\sum_{j=1}^{n}\varphi(X_j;\theta_k;\rho_{\theta_k})\leq\sup_{1\leq k\leq m}\mu(\theta_k)+\varepsilon\right\}=1\,. \tag{A.5}$$

En combinant (A.4) et (A.5), on trouve

$$\Pr\left\{\limsup_{n}\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}U(X_j;\theta)\leq\sup_{\theta\in\Theta}\mu(\theta)+\varepsilon\right\}=1$$

ce qui donne le résultat puisque ε est arbitraire. □

On peut renforcer la conclusion du théorème A.3 en supposant que $U(x,\theta)$ est continue en θ pour tout x. Dans ce cas, les conditions 2 et 4 sont satisfaites et $\mu(\theta)$ devient continue en θ. On applique alors le théorème à $U(x,\theta)-\mu(\theta)$ et $-U(x,\theta)+\mu(\theta)$ et on peut en déduire que

$$\Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}U(X_j;\theta)-\mu(\theta)\right|=0\right\}=1\,.$$

Toujours dans le cadre de la démonstration de la convergence du MLE, nous avons fait appel à *l'inégalité de Jensen* et au *lemme de Fatou*.

Théorème A.4. Inégalité de Jensen

Si g est une fonction convexe, alors $\mathbb{E}\left[g(X)\right]\geq g(\mathbb{E}[X])$ pour toute variable aléatoire X, aussitôt que les espérances existent.

Démonstration. Posons $\mu = \mathbb{E}[X]$. Par la définition de la convexité, il existe une fonction linéaire $h(t) = a\,t + b$ qui est une équation de la tangente à g en $t = \mu$. Alors, pour chaque t, $g(t) \geq a\,t + b$. Donc

$$\mathbb{E}\left[g(X)\right] \geq \mathbb{E}[a\,X + b] = a\,\mathbb{E}[X] + b = h(\mu) = g(\mu) = g(\mathbb{E}[X])\,.$$

□

Lemme A.5. Fatou-Lebesgue

Soit f_n une suite de fonction décroissante et continue sur un ensemble $\mathcal{E}$ de $\mathbb{R}$. Alors,

$$\int_{\mathcal{E}} \liminf f_n(x)dx \leq \liminf \int_{\mathcal{E}} f_n(x)dx$$

et

$$\limsup \int_{\mathcal{E}} f_n(x)dx \leq \int_{\mathcal{E}} \limsup f_n(x)dx$$

(en multipliant la première inégalité par -1).

Dans le cadre de la démonstration de la normalité asymptotique, nous avons utilisé la version renforcée du théorème A.3. De plus, nous avons fait appel aux deux résultats suivants.

Théorème A.6. Dérivation sous le signe de l'intégrale

Soit $f : (t, x) \to f(t, x)$ un fonction de $I \times E$ dans $\mathcal{C}$. On suppose que :

- pour tout $t \in I$, $x \to f(t, x)$ est intégrable ;
- pour tout $x \in E$, $t \to f(t, x)$ admet une dérivée sur I, notée $\frac{\partial f}{\partial t}$;
- il existe une fonction $\phi : E \to \mathbb{R}^+$ mesurable telle que $\int \phi(x)\,dx < \infty$ et pour tout $t \in I$, pour tout $x \in E$, $\left\|\frac{\partial f}{\partial t}(t, x)\right\| \leq \phi(x)$.

Alors la fonction

$$F : t \to F(t) = \int f(t, x)\ dx$$

est dérivable sur I et pour tout $t \in I$,

$$F'(t) = \int \frac{\partial f}{\partial t}(t,x)\,dx\,.$$

Formule de Taylor avec reste intégral de Laplace

Soient I un intervalle de $\mathbb{R}$, a un élément de I et f une fonction de I dans E dérivable en a et de classe $n \geq 1$. Pour tout nombre réel $x \in I$, on a la formule de Taylor-Young suivante :

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \mathcal{R}_n(x)$$

où $\mathcal{R}_n(x)$ est une fonction négligeable par rapport à $(x-a)^n$ au voisinage de a. Si la fonction f est de classe $n+1$ sur I et a valeur dans l'espace réel, alors, pour tout $x \in I$:

$$\mathcal{R}_n(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!}\,(x-t)^n\,dt\,.$$

Théorème A.7. Théorème central limite

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles définies sur le même espace probabilisé, indépendantes et de mêmes lois, et $S_n = X_1 + X_2 + \cdots + X_n$ et $\overline{X}_n = S/n$. Alors

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}\left[X_1\right]\right) \hookrightarrow^{\mathcal{L}} N(0,\sigma^2)$$

où $\sigma^2 = \mathrm{Var}(X_1)$.

Extension au cas vectoriel

Le théorème précédent se prolonge au cas des vecteurs aléatoires X_n à valeurs dans $\mathbb{R}^d$: $X_n = (X_{1,n}, X_{2,n}, \ldots, X_{d,n})$. Pour l'écrire, posons

$$S_n = \begin{bmatrix} X_{1,1} \\ X_{2,1} \\ \vdots \\ X_{d,1} \end{bmatrix} + \cdots + \begin{bmatrix} X_{1,n} \\ X_{2,n} \\ \vdots \\ X_{d,n} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mathbb{E}\left[X_{1,1}\right] \\ \mathbb{E}\left[X_{2,1}\right] \\ \vdots \\ \mathbb{E}\left[X_{d,1}\right] \end{bmatrix} \quad \text{et} \quad \overline{X}_n = \frac{S_n}{n}\,.$$

Alors

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \hookrightarrow^{\mathcal{L}} N(0, \Sigma)$$

où Σ est la matrice de variance-covariance de X_1.

A.2 Théorèmes et résultats du chapitre II

Théorème A.8.

Soit u_n un vecteur aléatoire dans E^r qui satisfait,

$$u_n \hookrightarrow^{\mathcal{L}} u,$$

où u est une mesure de probabilité constante dans E^r. Supposons aussi un autre vecteur aléatoire dans E^r tel que,

$$|u_n - v_n| \leq \epsilon_n |u_n|, \qquad \epsilon_n \hookrightarrow 0 \tag{A.6}$$

ou

$$|u_n - v_n| \leq \epsilon'_n |v_n|, \qquad \epsilon'_n \hookrightarrow 0 \tag{A.7}$$

Dans les deux cas, $u_n \sim v_n$ et $v_n \hookrightarrow^{\mathcal{L}} u$.

Démonstration. Si $|\epsilon'_n| < \frac{1}{2}$, alors

$$|v_n| \leq |u_n| + |v_n - u_n| \leq |u_n| + \epsilon'_n |v_n|$$

et ceci nous permet d'écrire,

$$|u_n - v_n| \leq \epsilon'_n |v_n| / (1 - \epsilon'_n) \leq 2\epsilon'_n |u_n|$$

et comme $|\epsilon'_n| < \frac{1}{2}$ car $\epsilon'_n \hookrightarrow 0$, on peut conclure alors que $A.7$ implique $A.6$.

Il suffit maintenant de montrer que $\epsilon_n |v_n| \hookrightarrow 0$, pour se faire, supposons F la distribution asymptotique de $|u_n|$. Étant donné δ_0 accès petit, considérons δ tel que $\frac{\delta_0}{\delta}$ est un point de continuité de F.

Alors,

$$P(\epsilon_n|u_n| \geq \delta_0) \leq P(|\epsilon_n > \delta|) + P(|u_n| \geq \frac{\delta_0}{\delta})$$

En supposant que $\delta \to 0$, tel que, $\delta < \epsilon_n$, on obtient

$$\begin{aligned} P(\epsilon_n|u_n| \geq \delta_0) &\leq 0 + P(|u_n| \geq \frac{\delta_0}{\delta}) \\ &\leq 1 - F(\frac{\delta_0}{\delta}) \\ &\leq 0 \end{aligned}$$

D'où comme, $\epsilon_n|v_n| \hookrightarrow 0$ alors on peut conclure que $u_n \sim v_n$.

□

APPENDICE B

Nous rassemblons ici les codes R que nous avons utilisés pour obtenir la majorité des graphiques et tableaux du chapitre III.

B.1 Obtention des valeurs initiales

B.1.1 Simulation des probabilités initiales et de la chaîne de Markov cachée

```
#-----------------------------------------
#  Debut declaration librairie
#-----------------------------------------
library(HMM)
library(seqHMM)
library(plyr)
library(numDeriv)
#-----------------------------------------
#  Fin declaration librairie
#-----------------------------------------

#-----------------------------------------
# Debut simulation des vecteurs initiales
#-----------------------------------------

init_input = function(nb_vect_sim, length_seq_cmc, p_trans,
             p_emiss, int_cm_probs){
# input;
# @nb_vect_sim : nombre de vecteur initiale a simuler;
# @length_seq_cmc: la longueur de la chaine markov cachee a simuler;
# @p_trans : la matrice des probabilites de la chaine cache;
# @p_emiss : la matrice des probabilites de la chaine observee;
# @int_cm_probs : les probabilites intiales de la chaines de markov cachee.
```

```
#initialiser une matrice vide
init_mat = matrix(0, ncol = 6, nrow = nb_vect_sim)
# simulation des p1 et p2
init_mat[,1] = runif(nb_vect_sim, min = 0.01, max = 0.99)
init_mat[,2] = runif(nb_vect_sim, min = 0.01, max = 0.99)
# simulation de p3 et p4 sous les contraintes p3+p4+p5 = 1 et pi>0
# meme chose pour p6, p7 et p8
for (j in 1:2) {
p1_init = vector()
p2_init = vector()
i = 0
while (i <= nb_vect_sim) {
a = runif(1, min = 0.01, max = 0.99)
y = runif(1, min = 0.01, max=0.99)
if(a + y < 1){ p1_init[i] = a; p2_init[i] = y
i = i+1
}
}
init_mat[, 2+2*j-1] = p1_init
init_mat[, 3+2*j-1] = p2_init
}
# simuler une chaine de markov cachee
sim = simulate_hmm(n_sequences     = 1, transition_probs    = p_trans,
emission_probs  = p_emiss, initial_probs = int_cm_probs,
sequence_length = length_seq_cmc)
#transfromer la sequence de CMC avec les - en un vecteur
seq_cmc_sim = as.numeric(gsub('\\-', ' ', sim$observations))
return(list(init_mat = init_mat, seq_cmc_sim = seq_cmc_sim))
}

# exemple

p_trans = t(matrix(c(0.7, 0.2,
0.3, 0.8), nrow=2, ncol=2))

p_emiss = t(matrix(c(0.6, 0.3, 0.1,
0.25, 0.55, 0.3), nrow=3, ncol=2))

test_init_input = init_input(100, length_seq_cmc = 50, p_trans,
                p_emiss, int_cm_probs = c(1/2, 1/2))
#-----------------------------------------
# Fin simulation des vecteurs initiales
#-----------------------------------------
```

B.1.2 Calcul de la fonction de vraisemblance

```
#-----------------------------------------
# Debut de la fonction de calcul de la log
#          vraisemblance
#-----------------------------------------


llh = function(p,y){

# input;
# @p : le vecteur des probabilites;
# @y : le vecteur des observations de la chaine de markov cachee.

trans_mat = t(matrix(c(p[1], 1-p[1],
1-p[2], p[2]), nrow=2, ncol = 2))
emiss_mat = t(matrix(c(p[3], p[4], 1-p[3]-p[4],
p[5], p[6], 1-p[5]-p[6]), nrow=3,ncol=2))
# la longueur de la chaine
T  = length(y)
Pr = numeric(T)

# distribution initiale
V = c(1/2,1/2)

# calcul de la vraisemblance
for(c in 1:T){
w = numeric(nrow(trans_mat))
#i = Y[c-1]
j = y[c]
for (k in 1:nrow(trans_mat)) {
w[k] = 0
for (l in 1:nrow(trans_mat)) {
w[k] = w[k] + V[l] * trans_mat[l, k] * emiss_mat[l, j]
}
}
Pr[c] = sum(w)
V     = w/sum(w)
}
bjf = - sum(log(Pr))
if (is.finite(bjf)){
return(bjf)
}
return(10^10);
}
```

```
# exemple
p = c(p_trans[1,1], p_trans[2,2], p_emiss[1,1], p_emiss[1,2],
      p_emiss[2,1], p_emiss[2,2])
test_llh = llh(p , y =  test_init_input$seq_cmc_sim)
#-----------------------------------------
# Fin de la fonction de calcul de la log
#         vraisemblance
#-----------------------------------------
```

B.1.3 Calcul et sélection des *top* valeurs de départ

```
#-----------------------------------------
# Debut de fct qui calcul la valeur de llh
#   pour les valeurs initiales simulees
#-----------------------------------------

fct_top_vect_max_llh  = function(n_top_init_vec,
                                 liste_fct1){
# input;
# @n_top_init_vec : le nombre de vecteur initiale qu on
                    veut garder;
# @liste_fct1 : la liste des objets retourne par la fonction 1.

# calcul de la vraisemblance pour la vraisemblance
resl_ini_llh = apply(liste_fct1$init_mat, 1,
function(x) llh(x, y = liste_fct1$seq_cmc_sim))

# ordonner les valeurs de la vraisemblance
sort_int     = sort(resl_ini_llh, index.return = TRUE,
                    decreasing = FALSE)

# retourner les vecteurs initiaux qui maximisent llh
val_prb_init = liste_fct1$init_mat[head(sort_int$ix, n_top_init_vec),]

# retourner une matrice de vecteur initiales et la sequence
# de CMC simule
return(list(val_prb_init = val_prb_init,
                     seq_cmc_sim = liste_fct1$seq_cmc_sim))
}

#-----------------------------------------
#  Fin de fct qui calcul la valeur de llh
#   pour les valeurs initiales simulees
#-----------------------------------------
```

B.1.4 Utilisation de EM pour le calcul des probabilités initiales

```
#-----------------------------------------
# premiere etape de maximisation : calcul
#   des valeurs initiales avec EM
#-----------------------------------------


fct_valeur_init = function(liste_top_vect_seq, int_cm_probs){

mat_prob = liste_top_vect_seq$val_prb_init
seq_cmc  = liste_top_vect_seq$seq_cmc_sim
n_row    = nrow(mat_prob)
n_col    = ncol(mat_prob)

# initialization
res_mle  = matrix(0, ncol = n_col, nrow = n_row)

# lancer EM pour chaque vecteur des n.top vecteurs
# de probs initiale
for (c in 1:n_row) {
A   = t(matrix(c(mat_prob[c, 1], 1 - mat_prob[c, 1],
1 - mat_prob[c, 2], mat_prob[c, 2]),
nrow = 2, ncol = 2))
B   = t(matrix(c(mat_prob[c, 3], mat_prob[c, 4],
1 - mat_prob[c, 3] - mat_prob[c, 4],
mat_prob[c, 5], mat_prob[c, 6],
1 - mat_prob[c, 5] - mat_prob[c, 6]),
nrow = 3, ncol = 2))

hmm = initHMM(c("A", "B"), c(1, 2, 3),
startProbs = int_cm_probs,
transProbs = A, emissionProbs = B)

true_out = baumWelch(hmm, seq_cmc,
 maxIterations = 100, pseudoCount=0)

res_mle[c,] = c(true_out$hmm$transProbs[1,1],
true_out$hmm$transProbs[2, 2],
true_out$hmm$emissionProbs[1, 1],
true_out$hmm$emissionProbs[1, 2],
true_out$hmm$emissionProbs[2, 1],
true_out$hmm$emissionProbs[2, 2])
}
colnames(res_mle) = c("p1","p2","p3","p4","p5","p6")
```

```
# compter la frequence des estimateurs
t = count(as.data.frame(round(res_mle,2)),
             c("p1","p2","p3","p4","p5","p6"),)

# je recupere le mle le plus frequent
t = t[order(t$freq, decreasing = TRUE), ]

best_mle      = as.matrix(t)
prob_init_opt = best_mle[1,-7]

return(list(mle_depart = prob_init_opt, rank_mle = best_mle))
}
```

B.2 Lancement de la routine d'optimisation avec `nlminb`

Cette fonction retourne une liste contenant plusieurs objets : un tableau pour chaque paramètre qui contient les valeurs des estimés, la variance calculée de chaque estimé, son intervalle de confiance et un indicateur de couverture de la vraie valeur (qui sert à calculer les probabilités de couverture).

```
#-----------------------------------------
# Debut de l inference sur le MLE:
#   1- calcul du MLE;
#   2- la matrice de variance covariance;
#   3- intervalle de confiance du MLE.
#-----------------------------------------

fct_optm_infer_mle = function(n_rep, vect_prob_init, length_seq_cmc,
deg_conv_der, p_trans, p_emiss,
int_cm_probs){

# input;
# @n-rep : le nombre de MLE a simuler;
# @vect_prob_init: le vecteur de probabilites initiales
#                  pour l optimisation;
# @length_seq_cmc: la longueur de la cmc a simuler;
# @deg_conv_der: taux de convergence (exemple: derivee = 0.001);
# @p_trans : matrice de transition de la chaine de Markov;
# @p_emiss : matrice de transition de la chaine observee;
# @int_cm_probs: proba initiale de la chaine de markov.

len          = length(vect_prob_init)
```

```
mle_sim      = matrix(0,nrow = n_rep, ncol=len)
mat_var_para = matrix(0,nrow = n_rep, ncol=len)
mat_der_abs  = matrix(0,nrow = n_rep, ncol=len)
prob_reel    = c(p_trans[1,1], p_trans[2,2], p_emiss[1,1],
p_emiss[1,2], p_emiss[2,1], p_emiss[2,2])

k = 1

while (k <= n_rep) {
# simuler une CMC
sim = simulate_hmm(n_sequence       = 1,
transition_probs    = p_trans,
emission_probs  = p_emiss,
initial_probs = int_cm_probs,
sequence_length = length_seq_cmc)

# transfromer la sequence de CMC avec les - en un vecteur
seq_cmc_sim = as.numeric(gsub('\\-', ' ', sim$observations))

# optimiser llh
res_nlminb = nlminb(vect_prob_init, objective = llh,
y = seq_cmc_sim, scale  = 1,
control = list( iter.max = 250,
eval.max = 300, abs.tol = 10^(-15),
x.tol = .0001),
lower   = rep(0.01,6),
upper = rep(0.99,6))


# verifier la derivee si elle est proche de 0
derv_abs = abs(grad(llh,res_nlminb$par, y = seq_cmc_sim,
method = "complex",method.args =
list(eps = 1e-8, d = 0.00001,
zero.tol = sqrt(.Machine$double.eps/7e-8),
r = 4, v = 2, show.details = FALSE)))

# contraintes p3+p4<1 et p5+p6<1
if( res_nlminb$par[3] + res_nlminb$par[4] < 1 &&
res_nlminb$par[5] + res_nlminb$par[6] < 1
&& sum(res_nlminb$par != rep(0.01)) == 6 &&
sum(res_nlminb$par != rep(0.99)) == 6
&& derv_abs < rep(deg_conv_der, 6)){

print(k)
mle_sim[k,] = res_nlminb$par
```

```
# derivee numerique de llh pour le k ieme mle
hess = hessian(llh, x = mle_sim[k,], y = seq_cmc_sim,"complex",
method.args=list(eps = 1e-10, d = 0.00001,
zero.tol =
sqrt(.Machine$double.eps/7e-10), r = 6, v = 4))

# avant de calculer la variance (1/I(theta)) on verifie qu'on a
# pas de NANs et que la mat hess est inversible
if(sum(is.nan(hess)) == 0 && diag(hess) >= 0){
mat_var_para[k,]  = diag(solve(hess))
}
k = k + 1
}
}

colnames(mle_sim)      = c("p1", "p2", "p3", "p4", "p5", "p6")
colnames(mat_var_para) = c("p1", "p2", "p3", "p4", "p5", "p6")

#2- calcul des intervalles de confiance et les flag de couverture

seq <- matrix(c("p1", "p2", "p3", "p4", "p5", "p6"), ncol = 6)

# un array pour stocker les informations
array_obs <- array(0, dim   = c(nrow(mle_sim), 5, 6),
dimnames = list(c(), c("mle", "var.fish",
"lower.b", "upper.b", "appat.ic"),
c("p1", "p2", "p3", "p4", "p5", "p6")))

for (j in 1:length(seq[1, ])) {
for (c in 1 : length(mle_sim[, j])) {
array_obs[c, "var.fish", seq[, j]]  <- mat_var_para[c, j]
if(array_obs[c, "var.fish", seq[, j]] > 0) {

array_obs[c, "lower.b", seq[, j]]  = mle_sim[c, j] -
2 * (sqrt(mat_var_para[c, j])) #/ sqrt(length.seq.CMC))
array_obs[c, "upper.b", seq[, j]]  = mle_sim[c, j] +
2 * (sqrt(mat_var_para[c, j])) #/ sqrt(length.seq.CMC))
array_obs[c, "appat.ic", seq[, j]] =
as.numeric(prob_reel[j] >=
array_obs[c, "lower.b", seq[, j]] &&
prob_reel[j] <= array_obs[c, "upper.b",
seq[, j]])
}
array_obs[c, "mle", seq[,j]]         = mle_sim[c, j]
}
```

```
}

# l'inference avec la variance empirique
array_reel <- array(0, dim   = c(nrow(mle_sim), 5, 6),
dimnames = list(c(), c("mle", "var.emp",
"lower.b", "upper.b", "appat.ic"),
c("p1", "p2", "p3", "p4", "p5", "p6")))

biais = vector()
for (j in 1:6) {
array_reel[, "mle"     , j] = array_obs[, "mle", j]
array_reel[, "var.emp" , j] = var(array_obs[, "mle", j])
array_reel[, "lower.b" , j] = array_reel[, "mle", j] -
2 * sqrt(array_reel[, "var.emp", j])
array_reel[, "upper.b" , j] = array_reel[, "mle", j] +
2 * sqrt(array_reel[, "var.emp", j])
array_reel[, "appat.ic", j] = as.numeric(prob_reel[j] >=
                              array_reel[, "lower.b", j] &
prob_reel[j] <= array_reel[, "upper.b", j])
biais[j] = abs((mean(array_reel[, "mle", j]) -
                      prob_reel[j])/prob_reel[j])
}


data_mle_to_plot <- data.frame(
param   = c( rep("p1", n_rep), rep("p2", n_rep), rep("p3", n_rep),
rep("p4", n_rep), rep("p5", n_rep)
, rep("p6", n_rep)),
valeurs = c( mle_sim[,"p1"], mle_sim[,"p2"], mle_sim[,"p3"],
mle_sim[,"p4"], mle_sim[,"p5"],
mle_sim[,"p6"])
)

data_var_obs_to_plot <- data.frame(
param   = c( rep("var.obs.p1", n_rep), rep("var.obs.p2", n_rep),
rep("var.obs.p3", n_rep),
rep("var.obs.p4", n_rep), rep("var.obs.p5", n_rep),
rep("var.obs.p6", n_rep)  ),
valeurs = c(mat_var_para[,1], mat_var_para[,2], mat_var_para[,3],
mat_var_para[,4],
mat_var_para[,5], mat_var_para[,6])
)

return(list(array_obs = array_obs, array_reel = array_reel ,
data_mle_to_plot = data_mle_to_plot,
data_var_obs_to_plot = data_var_obs_to_plot,
```

```
biais = biais))

}

#-----------------------------------------
# Fin de l inference sur le MLE.
#-----------------------------------------
```

B.3 Exemple de déroulement du code avec toutes les fonctions

```
# 1- simulation des vecteurs de probabilite initiale;

p_trans = t(matrix(c(0.8, 0.2,
                     0.1, 0.9), nrow=2, ncol=2))

p_emiss = t(matrix(c(0.95, 0.03, 0.02,
                     0.07, 0.9, 0.03), nrow=3, ncol=2))

liste_init = init_input(nb_vect_sim = 1000, length_seq_cmc = 100,
                        p_trans, p_emiss,
                        int_cm_probs = c(1/2, 1/2))

# 2- calcul de la vraisemblance et selection des N top vecteur qui la
#    maximise

liste_top_vect_seq = fct_top_vect_max_llh(n_top_init_vec = 100,
                     liste_fct1 =    liste_init)

# 3- Cacul des valeurs initiales finales avec l'algorithme EM

liste_etp_optm_em  = fct_valeur_init(liste_top_vect_seq,
                                     int_cm_probs = c(1/2, 1/2))
vect_prob_init     = liste_etp_optm_em$mle_depart

# 4- Lancement de la fonction pour infere sur les MLEs

test_50  = fct_optm_infer_mle(n_rep=1000, vect_prob_init,
                              length_seq_cmc = 50,
                              p_trans, p_emiss,
                              deg_conv_der = 0.0001,
                              int_cm_probs = c(1/2, 1/2))
```

RÉFÉRENCES

Aitchison, J. et Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, *29*(3), 813–828.

Baum, E. et Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, *36*(6), 1554–1563.

Baum, L. E. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Billingsley, P. (1961). *Statistical inferences for Markov processes*. Midway reprint.

Efron, B. et Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator : Observed versus expected fisher information. *Biometrika*, *65*(3), 457–482.

Ferguson, T. (1996). *A course in large sample theory*. Chapman and Hall/CRC.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *12*(1), 155–160.

Gilbert, E. J. (1959). On the identifiability problem for functions of finite markov chains. *Annals of Mathematical Statistics*, *30*(3), 688–697.

Hamilton, J. D. (1987). Rational expectations econometric analysis of change in regime. *Journal of Economic Dynamics and Control*, 384–423.

J. Kiefer, J. W. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Kalbfleisch, J. D. et Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B*, *32*(2), 175–208.

LeCam, L. (1960). On some asymptotic propreties of maximum likelihood estimates and related estimates. *Matematika*, *4*(2), 69–120.

Lévy, P. (1937). *Théorie de L'addition des Variables Aléatoires*. Monographies des Probabilités, publiés sous la direction de E. Borel.

Neyman, J. et Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.

Omura, J. (1969). On the viterbi decoding algorithm. *IEEE Transactions on Information Theory*, *15*(1), 177–179.

Rabiner, L. et Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine (Volume : 3)*, *3*(1), 4–16.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, *20*(4), 595–601.