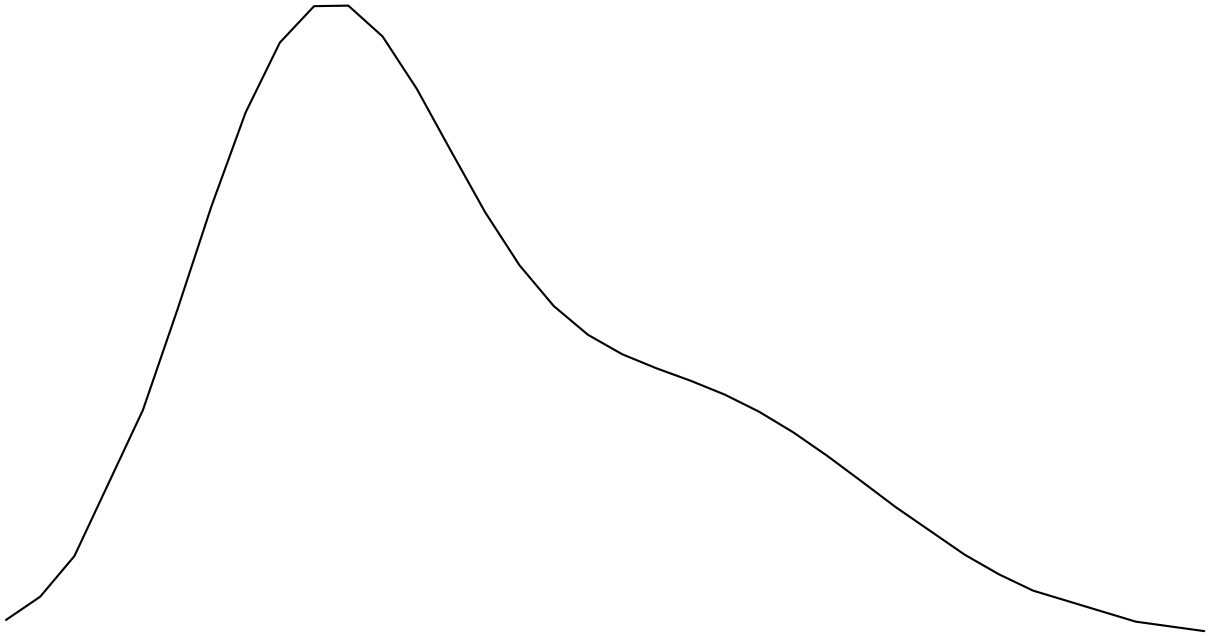


Les chaînes de Markov à variable cachée

Université de Montréal

Gabriel Lemyre
Stagiaire d'été

supervisé par
Professeur Maciej AUGUSTYNIAK

19 août 2016

Table des matières

Notations et abréviations utilisées

m	Nombre de composantes (distributions) considérées dans le modèle [2] ou le nombre d'états de la chaîne de Markov. [3]
$\bar{x}$	Moyenne échantillonnale.
s^2	Variance échantillonnale.
$\{X_t\}$	Processus observé.
$X^{(t)}$	Histoire (valeurs passées) du processus observé jusqu'au temps t.
$\{C_t\}$	Processus non observé.
$C^{(t)}$	Histoire (valeurs passées) du processus non observé jusqu'au temps t.
$\boldsymbol{\delta}$	$[\delta_1 \ \dots \ \delta_m]$ Distribution stationnaire d'une chaîne de Markov.
δ_i	$\Pr[C_t = i]$, la fraction moyenne de temps à long terme passée à l'état i.
Θ_i	Vecteur de paramètre(s) associé à la $i^{ème}$ distribution de probabilité.
λ_i	Paramètre de la loi de Poisson associé à la $i^{ème}$ distribution de probabilité.
$p(x_t)$	$\Pr[X_t = x_t]$.
$p_i(x_t)$	$\Pr[X_t = x_t \,\vert\, C_t = i]$.
$\mathbb{S}_m$	Espace d'états d'une chaîne de Markov à m états distincts.
$\mathbb{1}_{\{A\}}$	La fonction indicatrice de l'événement A.
γ_{ij}	$\Pr[C_{t+1} = j \,\vert\, C_t = i]$.
Γ	La matrice de transition de la chaîne de Markov.
f_{ii}	$\Pr[\text{retour éventuel en } i \,\vert\, \text{départ en } i]$.
μ_i	$\mathbb{E}[\text{temps de premier retour à } i \,\vert\, \text{départ en } i]$
$\mathrm{d}(i)$	$pgcd\left\{n > 0 : \gamma_{ii}^{(n)} > 0\right\}$
$u(t)$	$(\Pr[C_t = 1], \dots, \Pr[C_t = m]) = [u_1(t) \ \dots \ u_m(t)]$.
$\mathbf{P}(x_t)$	Matrice dont la diagonale contient les $p_i(x_t), \quad \forall\, i \in \mathbb{S}_m$.
$\mathbf{1}$	$[1\ 1\ 1\ \dots\ 1]$
μ_i	$\mathbb{E}[X_t \,\vert\, C_t = i]$.
σ_i^2	$\mathbb{V}\mathrm{ar}[X_t \,\vert\, C_t = i]$.
$\boldsymbol{\mu}$	$[\mu_1 \ \dots \ \mu_m]$.
$\mathbf{M}$	Matrice dont la diagonale contient $\mu_i, \quad \forall\, i \in \mathbb{S}_m$.
$\rho[X_t X_{t+k}]$	Corrélation de X_t et X_{t+k}.
$\mathrm{Cov}[X_t, X_{t+k}]$	Covariance de X_t et X_{t+k}.
$\alpha_t(i)$	$\Pr[X_1 = x_1, X_2 = x_2, \dots, X_t = x_t, C_t = i]$
$\boldsymbol{\alpha}_t$	$[\alpha_t(1) \dots \alpha_t(m)]$
L_T	Fonction de vraisemblance d'un échantillon de T observations.
η_i	$\log \lambda_i$.
τ_{ij}	$\log\left(\dfrac{\gamma_{ij}}{\gamma_{ii}}\right)$.
ϕ_t	$\dfrac{\boldsymbol{\alpha}_t}{\omega_t}$.
ω_t	$\boldsymbol{\alpha}_t \mathbf{1}'$.
l_T	Fonction de log-vraisemblance $\ln(L_T)$.

iid	indépendant et identiquement distribué
c-à-d	c'est-à-dire
MMC	Modèle de Markov caché

[↑]

1 Distribution classique

Dans le présent rapport, il sera question de la modélisation d'une série temporelle de variables aléatoires discrètes. Commençons avec un échantillon simple comportant n observations. Notons $\{X_t\}_{t=1}^n = \{x_1, x_2, \ldots, x_n\}$ la série temporelle en question. À titre d'exemple, nous considérerons le nombre de tremblements de terre (l'évènement) répertoriés par années (la période de temps) sur la planète. Nous utiliserons, à partir de maintenant, les données recueillies entre 1900 et 2006 recensant les tremblements de terre de magnitude 7 ou plus observés par années à travers le monde.

TABLE 1 – Nombre de tremblements de terre par an, tel que retrouvé sur le site http ://-neic.usgs.gov/neis/eqlist (à lire sur les lignes)

13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21	21	14	8	11	14	23	18	17	19	20	22	19
13	26	13	14	22	24	21	22	26	21	23	24	27	41	31	27	35	26	28	36	39	21	17	22	17	19	15	34	10	15
22	18	15	20	15	22	19	16	30	27	29	23	20	16	21	21	25	16	18	15	18	14	10	15	8	15	6	11	8	7
18	16	13	12	13	20	15	16	12	18	15	16	13	15	16	11	11													

La distribution classique utilisée pour modéliser un dénombrement non borné est la distribution de Poisson. Cette distribution n'a qu'un paramètre noté λ. C'est ce paramètre qui définit en entier la distribution puisque si X est une variable aléatoire de loi Poisson avec paramètre λ (dénoté $X \sim \text{Poisson}(\lambda)$), sa fonction de masse est

$$\Pr[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$$

et

$$\mathbb{E}[X] = \mathbb{V}\text{ar}[X] = \lambda.$$

Graphiquement, l'histogramme de notre échantillon, superposé avec une distribution de Poisson($\lambda = 19.36449$), ressemble à la Figure 1.

Tout d'abord, le graphique illustre la forte dispersion des valeurs possibles de l'échantillon. En effet, la moyenne de notre échantillon $\overline{x} \approx 19$ et sa variance $s^2 \approx 52$ sont très loin l'un de l'autre. Ceci pose problème puisqu'une distribution de Poisson classique telle que détaillée plus haut ne permet pas un tel écart. Il faut donc utiliser une modélisation plus flexible pour décrire cet échantillon. Comment est-il donc possible de modéliser un échantillon présentant ce type de dispersion ? Pour ce faire, une distribution permettant des queues de distribution plus larges sera nécessaire.

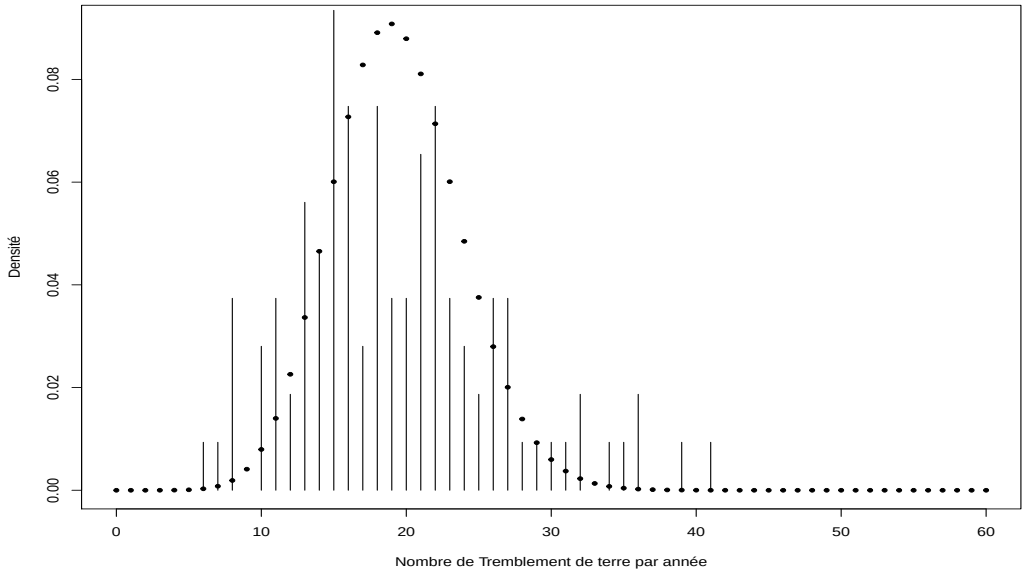


FIGURE 1 – Histogramme du nombre de tremblements de terre par année entre 1900-2006 et distribution de Poisson associée

[↑]

2 Les modèles de mélange indépendants

2.1 Propriétés des modèles de mélange indépendants

Il est possible que nous ne puissions pas supposer que la distribution du nombre de tremblements de terre par année soit la même chaque année. La technique généralement utilisée pour permettre d'accommoder ce type de situation est de considérer plus d'une distribution de probabilité. On fait alors appel à un modèle de mélange.

Définition 1. Soit m, le nombre de distributions considérées dans le mélange et t le temps. Un modèle de mélange considère 2 processus stochastiques [a]. Le premier processus est noté $\{C_t\}$, est considéré *iid* et n'est pas observé. Ce processus génère une valeur $c_t \in 1, \ldots, m$. C'est la valeur que prend C_t qui régit la distribution du second processus au temps t. Ce second processus, noté $\{X_t\}_{t=1}^{n}$ est celui qui génère l'observation x_t.

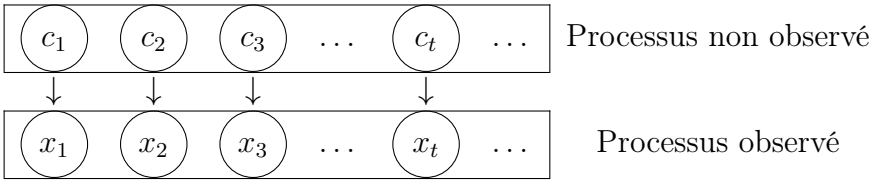


TABLE 2 – Graphe orienté d'un modèle de mélange

Si $c_t = i$, x_t est tiré à partir de la distribution associée à la composante i. Conditionnellement à la valeur prise par C_t, on suppose que X_t est indépendant de $(X_1, X_2, \ldots, X_{t-1}, X_{t+1}, \ldots, X_n)$.

Ceci implique que le processus $\{X_t\}$ est donc iid, ceci sera démontré à la fin de cette section.

a. Un processus stochastique est une suite de variables aléatoires indexées par le temps.

Ce type de modèle permet donc une plus grande flexibilité puisque les différentes distributions permettent d'élargir la distribution du mélange et de compenser la dispersion.

Dans le cas avec 2 composantes, la distribution du mélange dépend de deux probabilités, nous noterons :

Composante	1	2
Probabilité	$p_1(x_t) = \Pr[X_t = x_t \mid \Theta_1]$	$p_2(x_t) = \Pr[X_t = x_t \mid \Theta_2]$

où Θ_i est le vecteur de paramètres de la i^{ieme} distribution/composante. Nous définirons les valeurs de C_t de la manière suivante :

$$C_t = \begin{cases} 1 & \text{avec probabilité } \delta_1 \\ 2 & \text{avec probabilité } \delta_2 = 1 - \delta_1. \end{cases}$$

Si C_t=1, x_t est tiré de la distribution $p_1(x)$, et si $C_t = 2$, la valeur de x_t est tirée de celle de $p_2(x)$. Nous supposons par contre ne pas connaître les valeurs de C_t ayant générées les observations x_t.

Étendre ceci à un nombre de composantes plus élevé ($m > 2$) est assez simple :

Définition 2. Soient $\delta_1, \ldots, \delta_m$, les probabilités effectuant le mélange et $p_1(x_t), \ldots, p_m(x_t)$ leurs fonctions de masse associées. Ces probabilités sont définies comme suit :

$$\delta_i = \Pr[C_t = i]$$

et les conditions habituelles sur les probabilités s'appliquent :

$$\begin{aligned} 0 <& \delta_i < 1 \qquad \forall\, i \in \{1, \ldots, m\}, \\ &\sum_{i=1}^{m} \delta_i = 1. \end{aligned}$$

En utilisant ces outils, la distribution marginale de X_t (simplement notée $p(x_t)$=$\Pr[X_t = x_t]$) est trouvée en utilisant la formule de probabilité totale et en conditionnant sur la valeur de C_t :

$$\begin{aligned} p(x_t) = \Pr[X_t = x_t] &= \sum_{i=1}^{m} \Pr[X_t = x_t \mid C_t = i] \Pr[C_t = i], \\ &= \sum_{i=1}^{m} \delta_i p_i(x_t). \end{aligned} \tag{1}$$

Pour obtenir les distributions conjointes de plus grand ordre ($\Pr[X_t, X_{t+k}]$), nous aurons besoin d'utiliser l'indépendance du processus $\{X_t\}$. Il sera tout d'abord nécessaire de prouver ce résultat.

Proposition 1.
Notons

$$p(x_1, x_2, \ldots, x_n) = \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]$$

et

$$q(c_1, c_2, \ldots, c_n) = \Pr[C_1 = c_1, C_2 = c_2, \ldots, C_n = c_n]\,.$$

Pour un échantillon de taille n $\in \mathbb{N}$

$$p(x_1, x_2, \ldots, x_n) = \Pr[X_1 = x_1] \cdot \Pr[X_2 = x_2] \cdot \ldots \cdot \Pr[X_n = x_n] = \prod_{i=1}^{n} \Pr[X_i = x_i]$$

et donc $\{X_t\}$ est indépendant.

Preuve. Tout d'abord, puisque $\sum_B \Pr[A, B] = \Pr[A]$,

$p(x_1,x_2, \ldots, x_n)$

$$= \sum_{c_1, \ldots, c_n} \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, C_1 = c_1, C_2 = c_2, \ldots, C_n = c_n]$$

$$= \sum_{c_1, \ldots, c_n} \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, C_1 = c_1, C_2 = c_2, \ldots, C_n = c_n] \cdot q(c_1, c_2, \ldots, c_n).$$

Puisque le processus $\{C_t\}$ est indépendant et que les X_t sont indépendants conditionnellement aux valeurs prises par C_t,

$$= \sum_{c_1, \ldots, c_n} \Pr[X_1 = x_1 \,|\, C_1 = c_1]\, q(c_1) \Pr[X_2 = x_2 \,|\, C_2 = c_2]\, q(c_2) \ldots \Pr[X_n = x_n \,|\, C_n = c_n]\, q(c_n)$$

$$= \sum_{c_1} \Pr[X_1 = x_1 \,|\, C_1 = c_1]\, q(c_1) \sum_{c_2} \Pr[X_2 = x_2 \,|\, C_2 = c_2]\, q(c_2) \ldots \sum_{c_n} \Pr[X_n = x_n \,|\, C_n = c_n]\, q(c_n)$$

$$= \prod_{i=1}^{n} \sum_{c_i} \Pr[X_i = x_i \,|\, C_i = c_i]\, q(c_i)$$

$$= \prod_{i=1}^{n} \sum_{c_i} \Pr[X_i = x_i, C_i = c_i]$$

$$= \prod_{i=1}^{n} \Pr[X_i = x_i].$$

Le processus $\{X_t\}$ est donc indépendant. □

Les distributions conjointes de plus grands ordres sont donc le produit des distributions marginales d'ordre 1.

2.2 Les moments de X_t

L'espérance de X_t peut être calculée en conditionnant sur les espérances de X_t sous les différentes valeurs que peut prendre C_t. Ainsi :

$$\begin{aligned}\mathbb{E}[X_t] &= \sum_{i=1}^{m} \Pr[C_t = i]\ \mathbb{E}[X_t \,|\, C_t = i] \\ &= \sum_{i=1}^{m} \delta_i \mathbb{E}[X_t \,|\, C_t = i]\,.\end{aligned} \tag{2}$$

En définissant Y_i comme étant la variable aléatoire de distribution $p_i(x_t)$ on obtient que

$$\mathbb{E}[X_t] = \sum_{i=1}^{m} \delta_i \mathbb{E}[Y_i].$$

Les moments de plus grands ordres, le k^{ieme} par exemple, se calculent de manière équivalente comme une combinaison linéaire des moments de Y_i du même ordre :

$$\mathbb{E}\big[X_t^k\big] = \sum_{i=1}^{m} \delta_i \mathbb{E}\big[Y_i^k\big]\,, \qquad k \in \mathbb{N}. \tag{3}$$

Il est par contre faux de penser que la variance de X_t est une combinaison linéaire des variances des composantes Y_i. Dans le cas à 2 composantes, on obtient la variance de X_t comme suit :

$$\mathbb{V}\mathrm{ar}[X_t] = \delta_1 \mathbb{V}\mathrm{ar}[Y_1] + \delta_2 \mathbb{V}\mathrm{ar}[Y_2] + \delta_1\delta_2(\mathbb{E}[Y_1] - \mathbb{E}[Y_2])^2. \tag{4}$$

La preuve de ce résultat nécessite de connaître le théorème de la variance totale.

Théorème 1 (Théorème de la variance totale)**.** Si U et V sont 2 variables aléatoires rattachées à un même espace de probabilités et si $\mathbb{V}\mathrm{ar}[U]$ est finie, alors

$$\mathbb{V}\mathrm{ar}[U] = \mathbb{E}[\mathbb{V}\mathrm{ar}[U|V]] + \mathbb{V}\mathrm{ar}[\mathbb{E}[U|V]]\,.$$

Avec ce théorème en main, il est maintenant possible de prouver l'équation (4).

Preuve. Commençons par définir

$$\begin{aligned}\mu_1 &= \mathbb{E}[Y_1] = \mathbb{E}[X_t|C_t = 1]\,, \\ \mu_2 &= \mathbb{E}[Y_2] = \mathbb{E}[X_t|C_t = 2]\,, \\ \sigma_1^2 &= \mathbb{V}\mathrm{ar}[Y_1] = \mathbb{V}\mathrm{ar}[X_t|C_t = 1]\,, \\ \sigma_2^2 &= \mathbb{V}\mathrm{ar}[Y_2] = \mathbb{V}\mathrm{ar}[X_t|C_t = 2]\,.\end{aligned}$$

Il en découle de cette notation que

$$\begin{aligned}\mathbb{E}[\mathbb{V}\mathrm{ar}[X|C_t]] &= \mathbb{E}\big[\mathbb{1}_{\{C_t=1\}}\mathbb{V}\mathrm{ar}[X|C_t = 1] + \mathbb{1}_{\{C_t=2\}}\mathbb{V}\mathrm{ar}[X|C_t = 2]\big] \\ &= \delta_1 \mathbb{V}\mathrm{ar}[X|C_t = 1] + \delta_2 \mathbb{V}\mathrm{ar}[X|C_t = 2] \\ &= \delta_1\sigma_1^2 + \delta_2\sigma_2^2.\end{aligned}$$

Il ne manque maintenant que

$$\mathbb{V}\mathrm{ar}[\mathbb{E}[X|C_t]]\,.$$

On sait que

$$\mathbb{E}[\mathbb{E}[X|C_t]] = \mathbb{E}[X]\,,$$

et que

$$\mathbb{1}_{\{x\}}^2 = \mathbb{1}_{\{x\}},$$

et donc

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\mathbb{E}[X|C_t]] &= \mathbb{E}\big[\mathbb{E}[X|C_t]^2\big] - (\mathbb{E}[\mathbb{E}[X|C_t]])^2 \\
&= \mathbb{E}\big[\mathbb{E}[X|C_t]^2\big] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}\big[\big[\mathbb{E}[X|C_t=1]\,\mathbb{1}_{\{C_t=1\}} + \mathbb{E}[X|C_t=2]\,\mathbb{1}_{\{C_t=2\}}\big]^2\big] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}\big[\big[\mu_1\mathbb{1}_{\{C_t=1\}} + \mu_2\mathbb{1}_{\{C_t=2\}}\big]^2\big] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}\big[\big[\mu_1^2\mathbb{1}_{\{C_t=1\}}^2 + \mu_1\mu_2(\mathbb{1}_{\{C_t=1\}}\mathbb{1}_{\{C_t=2\}}) + \mu_2^2\mathbb{1}_{\{C_t=2\}}^2\big]\big] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}\big[\big[\mu_1^2\mathbb{1}_{\{C_t=1\}}^2 + 0 + \mu_2^2\mathbb{1}_{\{C_t=2\}}^2\big]\big] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}\big[\big[\mu_1^2\mathbb{1}_{\{C_t=1\}} + \mu_2^2\mathbb{1}_{\{C_t=2\}}\big]\big] - (\mathbb{E}[X])^2 \\
&= (\delta_1\mu_1^2 + \delta_2\mu_2^2) - (\mathbb{E}[X])^2 \\
&= (\delta_1\mu_1^2 + \delta_2\mu_2^2) - (\delta_1\mu_1 + \delta_2\mu_2)^2 \\
&= \delta_1\mu_1^2 + \delta_2\mu_2^2 - \delta_1^2\mu_1^2 - \delta_1\delta_2\mu_1\mu_2 - \delta_2^2\mu_2^2 \\
&= \delta_1\mu_1^2(1-\delta_1) + \delta_2\mu_2^2(1-\delta_2) - \delta_1\delta_2\mu_1\mu_2 \\
&= \delta_1\delta_2\mu_1^2 + \delta_1\delta_2\mu_2^2 - \delta_1\delta_2\mu_1\mu_2 \\
&= \delta_1\delta_2(\mu_1^2 - \mu_1\mu_2 + \mu_2^2) \\
&= \delta_1\delta_2(\mu_1 - \mu_2)^2.
\end{aligned}$$

Et donc,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X_t] &= \delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2 \\
&= \delta_1\mathbb{V}\mathrm{ar}[Y_1] + \delta_2\mathbb{V}\mathrm{ar}[Y_2] + \delta_1\delta_2(\mathbb{E}[Y_1] - \mathbb{E}[Y_2])^2.
\end{aligned}$$

□

2.3 Estimation des paramètres d'un modèle de mélange

L'estimation des paramètres d'un modèle de mélange est généralement effectuée par l'entremise de sa fonction de vraisemblance. C'est en maximisant cette fonction que l'on obtient une paramétrisation optimale du processus sous-jacent. Prenons en exemple un modèle de mélange à m composantes, une série de n observations $\{x_1, x_2, \ldots, x_n\}$, les vecteurs de paramètres $\theta_1,\ldots,\theta_m$ associés aux distributions de ces m composantes et $\delta_1,\ldots,\delta_m$ les probabilités effectuant le mélange. Dans ce cas, la fonction de vraisemblance est définie comme suit :

$$L(\Theta_1, \ldots, \Theta_m, \delta_1, \ldots, \delta_m | x_1, \ldots, x_n) = \prod_{j=1}^{n}\sum_{i=1}^{m}\delta_i p_i(x_j|\Theta_i). \tag{5}$$

Pour la suite du présent rapport, il sera question de modèles de mélange ayant pour composantes des lois de Poisson(λ_i). Un tel modèle à m composantes admet $(2m-1)$ paramètres indépendants à estimer, l'un des δ_i (disons δ_m) étant trouvé par soustraction : $\delta_m = 1 - \sum_{i=1}^{m-1}\delta_i$. La maximisation de cette fonction n'est vraiment pas simple puisqu'il n'existe pas à ce jour de

technique analytique pour résoudre un tel problème de maximisation.

Pour illustrer la complexité de cette maximisation, supposons que les observations de l'échantillon des tremblements de terre soient issues de 2 lois de Poisson de paramètres λ_1 et λ_2 et que ces paramètres soient choisis aléatoirement par un processus non observé. Les paramètres δ_1 et δ_2 restent les mêmes et les vecteurs de paramètres Θ_1 et Θ_2 sont associés aux paramètres de Poisson : $(\Theta_1 = \lambda_1)$ et $(\Theta_2 = \lambda_2)$. La distribution de mélange de ce modèle $(p(x_t))$ est alors

$$p(x_t) = \delta_1 \frac{e^{-\lambda_1}\lambda_1^{x_t}}{x_t!} + (1-\delta_1)\frac{e^{-\lambda_2}\lambda_2^{x_t}}{x_t!}. \tag{6}$$

Il y a donc 3 paramètres à estimer et la fonction de vraisemblance est simplement le produit des $p(x_t)$ sur $t \in \{1, \ldots, n\}$:

$$L(\lambda_1, \lambda_2, \delta_1 | x_1, \ldots, x_n) = \prod_{t=1}^{n} \left[\delta_1 \frac{e^{-\lambda_1}\lambda_1^{x_t}}{x_t!} + (1-\delta_1)\frac{e^{-\lambda_2}\lambda_2^{x_t}}{x_t!} \right]. \tag{7}$$

et la fonction de log-vraisemblance est

$$\begin{aligned} l(\lambda_1, \lambda_2, \delta_1 | x_1, \ldots, x_n) &= \ln\left(L(\lambda_1, \lambda_2, \delta_1 | x_1, \ldots, x_n)\right) \\ &= \sum_{t=1}^{n} \ln\left[\delta_1 \frac{e^{-\lambda_1}\lambda_1^{x_t}}{x_t!} + (1-\delta_1)\frac{e^{-\lambda_2}\lambda_2^{x_t}}{x_t!} \right] \\ &= \sum_{t=1}^{n} \left[-\ln(x_t!) + \ln\left[\delta_1\, e^{-\lambda_1}\lambda_1^{x_t} + (1-\delta_1)\, e^{-\lambda_2}\lambda_2^{x_t}\right]\right]. \end{aligned} \tag{8}$$

La maximisation de l'équation (7) par rapport aux 3 paramètres est plutôt désagréable, L est tout de même le produit de n éléments, chacun d'entre eux une somme de 2 éléments. Appliquer le logarithme naturel $\ln(L)$ comme en (8) ne permet pas non plus de maximiser la vraisemblance analytiquement. Il est donc adéquat, pour l'instant, d'utiliser des méthodes numériques dans R permettant d'effectuer ce genre de calcul (les fonctions *nlm* et *optim* par exemple). Voici un exemple de code R permettant de prendre une série d'observation X (ligne 3), y trouver les valeurs distinctes (ligne 4) et leurs fréquences d'apparition (ligne 5) puis écrire la fonction de log-vraisemblance *logg* (ligne 11 à 14).

```
1  Vraisemblance_avec_2_composantes=function(v)
2  {
3    X=as.numeric(read.table("CountSerie.txt",header=F,sep=" "))
4    x=sort(unique(X),decreasing=F)
5    N=matrix(as.numeric(table(X)),nrow=1)
6    logg=matrix(nrow=length(x),ncol=1)
7    d1=v[1]
8    Lambda1=v[2]
9    Lambda2=v[3]
10
11   logg=log(
12           (d1*exp(-Lambda1)*(Lambda1^x)/factorial(x))
13              +((1-d1)*exp(-Lambda2)*(Lambda2^x)/factorial(x))
14   )
15
16   if (d1<=1){
17     if(0<=d1){
18
19            mLLH=-(N%*%logg)
```

```
20
21      }else{mLLH= 10000}
22    }else{mLLH= 10000}
23
24    mLLH
25  }
26
27  ME=optim(c(0.5,18,22),Vraisemblance_avec_2_composantes)
```

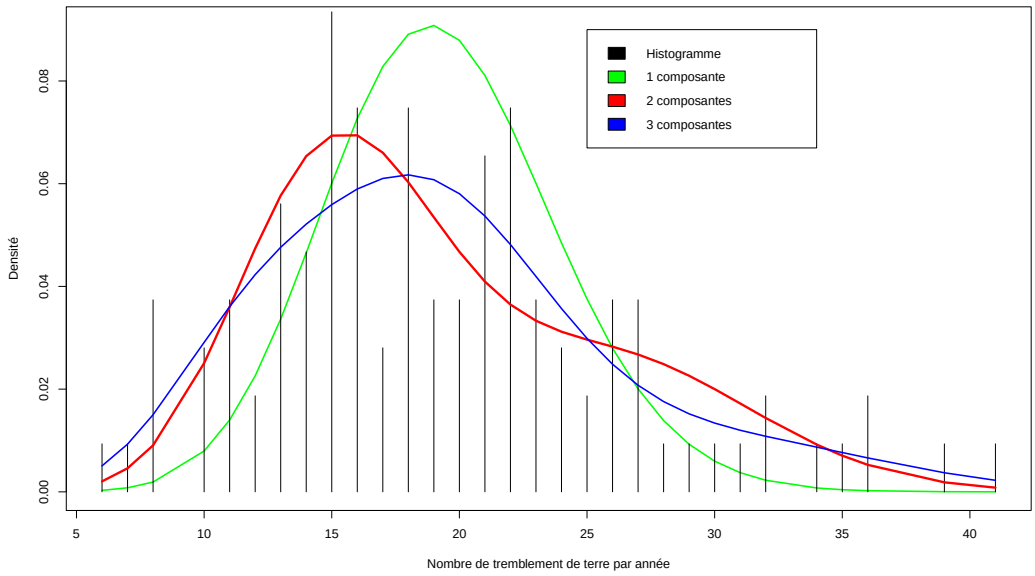
Les conditions énoncées aux lignes 16 et 17 permettent de restreindre le domaine dans lequel la fonction est optimisée. C'est ainsi qu'on assure que les probabilités soient toutes positives et inférieures à 1 et que leur somme vaut 1. Cette technique n'est cependant pas du tout efficace pour un modèle à plus de 2 composantes. Dans la Section 3.3.2, il sera question de re-paramétrisation, une technique beaucoup plus simple et concise permettant les mêmes conditions sur les δ_i . À la ligne 27, la fonction *optim* de R est utilisée, le premier argument de la fonction est un vecteur contenant les valeurs initiales $(\delta_1, \lambda_1, \lambda_2)$ sur lesquelles la fonction *optim* se base pour commencer l'optimisation.

Exemple 1. Utilisation du code ci-haut à notre échantillon

Lorsque nous étudions l'échantillon de la table 1 grâce à la fonction *Vraisemblance_avec_2_composantes*, nous obtenons comme valeurs estimées des paramètres :

$$\widehat{\delta_1} = 0{,}3242506 \quad \widehat{\delta_2} = 0{,}6757494 \quad \widehat{\lambda_1} = 15{,}7775813 \quad \widehat{\lambda_2} = 26{,}8416205$$

Graphiquement, la distribution du modèle à 2 composantes ainsi que les distributions à 1 et 3 composantes ressemblent à :



Cet exemple illustre bien le fait que d'augmenter le nombre de composantes étire les queues de la distribution.

Ce type de modèle est approprié lorsque l'on ne peut pas détecter de dépendance temporelle au sein de notre échantillon. Certains échantillons présentent cependant une auto-corrélation non nulle. La figure 2 illustre cette auto-corrélation à l'intérieur de notre échantillon.

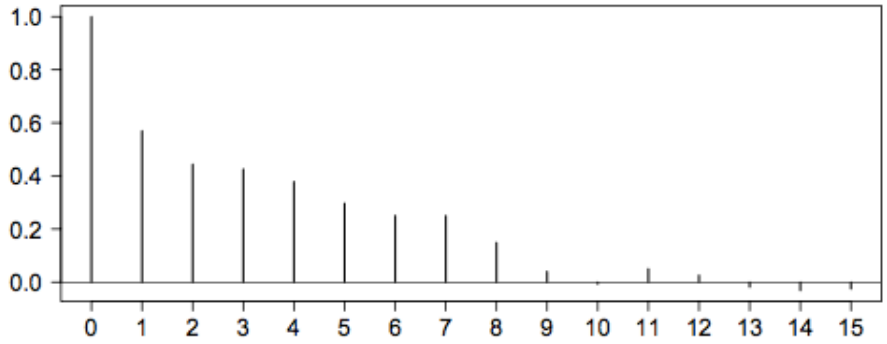


FIGURE 2 – Fonction d'auto-corrélation par rapport à l'écart entre les observations de notre échantillon

La figure 2 suggère que les observations de la suite $\{C_t\}$ ne sont pas indépendantes. Il sera donc nécessaire de s'intéresser à un type de modèle qui puisse tenir compte de cette dépendance à l'intérieur du processus. Nous tournons alors notre attention vers les modèles de Markov cachés. [↑]

3 Les modèles de Markov cachés (MMC)

Les modèles de Markov cachés supposent toujours l'indépendance du processus $\{X_t\}$ conditionnellement aux valeurs prises par $\{C_t\}$, telle que dans la définition 1. Cependant, on ne supposera plus que $\{C_t\}$ est iid et donc $\{X_t\}$ lui-même ne sera plus iid non plus. Les modèles de Markov cachés sont un type de modèle à changement de régime.

C'est grâce à ce type de modèle qu'il sera possible de tenir compte de la dépendance entre les observations du processus. Les modèles à changement de régime ont attiré l'attention des économistes suite aux travaux de l'économétricien James Douglas Hamilton dans les années 90. Il propose l'utilisation de tels modèles pour lier l'état de l'économie et le produit intérieur brut américain. Dans ce contexte, l'état dans lequel se retrouve le processus stochastique permet de modéliser les différentes structures sous-jacentes de l'économie. Ces modèles permettraient de tenir compte du fait qu'en période de récession économique, la volatilité des rendements est généralement plus élevée qu'en période de prospérité économique. Dans cet exemple, le modèle comporterait 2 composantes, une composante à moyenne négative et variance élevée et une à moyenne positive et variance faible.

Hamilton choisit d'utiliser les chaînes de Markov pour modéliser le processus $\{C_t\}$, c'est ce type de processus stochastique qui sera discuté dans la section suivante.

3.1 Les chaînes de Markov à temps discret

Dans cette section, il sera question des propriétés des chaînes de Markov étant nécessaires dans les modèles de Markov cachés.

Propriété Markovienne

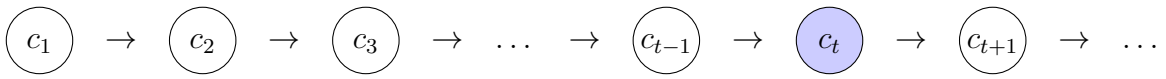
Une suite de variables aléatoires discrètes $\{C_t\}$ prenant des valeurs dans l'espace d'état $\mathbb{S}_m = \{1, \ldots, m\}$ est appelée chaîne de Markov si $\forall\, t \in \mathbb{N}$ elle respecte la propriété Markovienne (c-à-d l'absence de mémoire à l'intérieur du processus). Cette propriété est notée :

$$\Pr[C_{t+1} \,|\, C_t, \ldots, C_1] = \Pr[C_{t+1} \,|\, C_t]\,.$$

En notant $C^{(t)} = (C_t, \ldots, C_1)$, la propriété markovienne devient

$$\Pr[C_{t+1}|C^{(t)}] = \Pr[C_{t+1}|C_t]. \tag{9}$$

La probabilité sur C_t, conditionnellement aux valeurs précédemment prises par la suite, ne dépend donc que de la valeur de C_{t-1}. L'observation au temps t dépend seulement de l'observation au temps $t-1$.



Un des aspects importants des chaînes de Markov est l'existence de ces probabilités conditionnelles que nous appellerons maintenant **probabilités de transitions**. Pour une chaîne de Markov admettant m états distincts, ces probabilités de transition sont :

$$\Pr[C_{t+1} = j \,|\, C_t = i]\,, \quad i, j \in \mathbb{S}_m.$$

Homogénéité

Si les probabilités de transitions ne dépendent pas de $t \in \mathbb{N}$, alors la chaîne $\{C_t\}$ est dite *homogène* et on note

$$\gamma_{ij} = \Pr[C_{t+1} = j \,|\, C_t = i]\,, \quad \forall\, t \in \mathbb{N} \quad \forall\, i, j \in \mathbb{S}_m. \tag{10}$$

La relation (10) peut être étendue à un nombre de pas (transitions) $k \geq 1$. Nous noterons à l'avenir

$$\gamma_{ij}^{(k)} = \Pr[C_{t+k} = j \,|\, C_t = i]\,, \quad \forall\, t \in \mathbb{N} \quad \forall\, i, j \in \mathbb{S}_m, \tag{11}$$

la probabilité de transition en k pas.

Nous ferons toujours cette hypothèse.

Matrice de transition

La matrice de transition en k pas, notée $\Gamma^{(k)}$, est la matrice contenant, au croisement de la ligne i et de la colonne j, la probabilité $\gamma_{ij}^{(k)}$ $(\forall\, i, j \in \mathbb{S}_m)$. C'est cette matrice qui définit en entier la chaîne en question. En notation matricielle,

$$\Gamma^{(k)} = \begin{pmatrix} \gamma_{11}^{(k)} & \cdots & \gamma_{1m}^{(k)} \\ \vdots & \gamma_{ij}^{(k)} & \vdots \\ \gamma_{m1}^{(k)} & \cdots & \gamma_{mm}^{(k)} \end{pmatrix}.$$

Cette matrice, ainsi que son contenu, respecte les propriétés suivantes :

Propriété 1. $0 \leq \gamma_{ij}^{(k)} \leq 1, \quad \forall\, i, j \in \mathbb{S}_m,$

Propriété 2. La somme des valeurs sur une ligne de la matrice $\Gamma^{(k)}$ doit être égale à 1

$$\sum_j \gamma_{ij}^{(k)} = 1, \quad \forall\, i \in \mathbb{S}_m,$$

Propriété 3 (Équation de Chapman-Kolmogorov)**.**

$$\gamma_{ij}^{(n)} = \sum_b \gamma_{ib}^{(k)} \gamma_{bj}^{(n-k)}, \quad \forall\, i, j \in \mathbb{S}_m \quad \forall\, k \in \{1, \ldots, n\}, \tag{12}$$

avec

$$\gamma_{ij}^{(0)} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon.} \end{cases}$$

La même propriété s'applique à la matrice de transition $\Gamma^{(n)}$ et on obtient l'égalité

$$\Gamma^{(n)} = \Gamma^{(k)}\, \Gamma^{(n-k)}, \qquad \forall\, k \in \{0, \ldots, n\}.$$

avec $\Gamma^{(0)} = \mathrm{I}$, où I est la matrice identité.

Propriété 4. $\Gamma^{(n)} = \Gamma^n, \quad \forall\ n \geq 1.$

Classification des états

Les chaînes de Markov admettent plusieurs types d'états. La classification de ces différents types est essentielle à la compréhension du comportement asymptotique desdites chaînes.

(a) Un état i est dit **récurrent** si

$$f_{ii} = \Pr[\text{retour éventuel en } i \,|\, \text{départ en } i] = 1.$$

Si $f_{ii} < 1$ alors l'état est dit **transient**

(b) Un état récurrent i est dit **récurrent positif** si

$$\mu_i = \mathbb{E}[\text{temps de premier retour à } i \,|\, \text{départ en } i] < \infty.$$

Sinon ($\mu_i = \infty$), l'état est dit **récurrent nul**.

(c) La période $\mathrm{d}(i)$ d'un état i est définie comme suit

$$\mathrm{d}(i) = pgcd\left\{n > 0 : \gamma_{ii}^{(n)} > 0\right\}.$$

Un état i est **périodique** si $\mathrm{d}(i) > 1$ et **apériodique** si $\mathrm{d}(i) \in \{0, 1\}$.

(d) Un état i tel que $\gamma_{ii}^{(1)} = 1$ est dit **absorbant**.

(e) Un état récurrent positif apériodique est dit **ergodique**. Par exemple, un état absorbant est ergodique.

L'espace d'état $\mathbb{S}_m$ d'une chaîne de Markov peut être subdivisé en classes d'états de même type et de même période.

1. Une classe dont tous les états sont récurrents est dite *récurrente* ou *fermée*.
2. Une chaîne sur une seule classe d'états est dite *irréductible*.
3. Une chaîne irréductible (ou toute classe fermée) sur un espace d'états *fini* est récurrente positive.

Probabilités inconditionnelles

Il est important, pour la compréhension de la section suivante, de s'intéresser aux probabilités marginales $\Pr[C_t = i]$ de se retrouver dans un état i au temps t. Ces probabilités seront contenues dans un vecteur ligne noté

$$\begin{aligned} u(t) &= (\Pr[C_t = 1], \ldots, \Pr[C_t = m]) \\ &= [u_1(t) \ \ldots \ u_m(t)]. \end{aligned}$$

avec $u(1)$ la **distribution initiale** de la chaîne.

La distribution marginale au temps $(t+1)$ est obtenue en multipliant la distribution inconditionnelle au temps t par la matrice de transition Γ. Ainsi,

$$u(t+1) = u(t)\Gamma,$$

et donc,

$$u(t+1) = u(1)\Gamma^t. \tag{13}$$

Distribution stationnaire

Une chaîne irréductible sur un nombre fini d'états (espace d'états fini) admet une distribution stationnaire $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ unique. Cette **distribution stationnaire** doit respecter les conditions suivantes :

(a) $0 < \delta_j < 1 \quad \forall\, j \in \mathbb{S}_m$,

(b) $\sum_j \delta_j = 1$,

(c) $\delta_j = \sum_i \delta_i \gamma_{ij} \quad \forall\, j \in \mathbb{S}_m$ est appelée l'**équation de stationnarité**.

En notation matricielle, cette équation devient $\boldsymbol{\delta}\Gamma = \boldsymbol{\delta}$. En conséquence, puisque $u(t+1) = u(t)\Gamma$, une chaîne ayant pour distribution initiale sa distribution stationnaire ($u(1) = \boldsymbol{\delta}$) vérifie

$$u(t) = u(1) \quad \forall\, t \in \mathbb{N}.$$

Cette distribution stationnaire est également donnée par $\delta_j = \mu_j^{-1}$. Elle représente la fraction moyenne de temps passé à long terme dans l'état j.

La matrice de transition en k pas d'une chaîne irréductible sur un espace d'états fini $\mathbb{S}_m$ satisfait

$$\lim_{k\to\infty} \Gamma^{(k)} = \begin{pmatrix} \delta_1 & \delta_2 & \dots & \delta_m \\ \vdots & \vdots & \ddots & \vdots \\ \delta_1 & \delta_2 & \dots & \delta_m \end{pmatrix}.$$

3.2 Propriétés des modèles de Markov cachés

Tel que mentionné dans la Section 2.3, certains échantillons ont une auto-corrélation non nulle. On ne peut donc pas supposer l'indépendance du processus $\{X_t\}$. Les modèles de Markov cachés fonctionnent de la manière suivante :

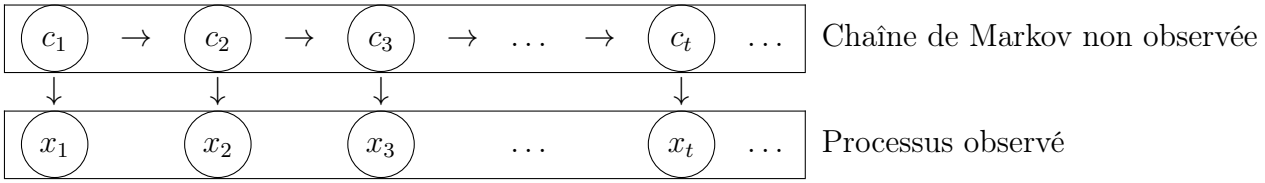


TABLE 3 – Graphe orienté d'un modèle de Markov cachée

Rappelons que, puisque $\{C_t\}$ est une chaîne de Markov, elle respecte toujours l'absence de mémoire du processus, c-à-d,

$$\Pr\left[C_t \,|\, C^{(t-1)}\right] = \Pr[C_t \,|\, C_{t-1}] \quad \forall\, t \in \{2, 3, \ldots\}.$$

Il est à noter qu'ici cette probabilité ne s'applique pas au premier état (C_1) pris par la chaîne $\{C_t\}$. Il sera donc nécessaire de faire une hypothèse sur la distribution initiale $u(1)$ de $\{C_t\}$ lors du conditionnement sur la valeur du premier état.

Cette absence de mémoire s'applique aussi au processus $\{X_t\}$, mais seulement lorsqu'on conditionne sur $\{C_t\}$. Tel que mentionné au début de la section 3, $\{X_t\}$ est indépendant conditionnellement aux valeurs prises par $\{C_t\}$. On obtient donc

$$\Pr\left[X_t \,|\, X^{(t-1)}, C^{(t)}\right] = \Pr[X_t \,|\, C_t], \quad \forall\, t \in \mathbb{N}. \tag{14}$$

Dans la présente section, nous aurons besoin de la distribution marginale de X_t ainsi que des distributions conjointes de plus grands ordres (i.e., $\Pr[X_t, X_{t+k}]$).

Distribution marginale de X_t

(a) En ne considérant pas nécessairement la chaîne de Markov comme *stationnaire*,

$$
\begin{aligned}
p(x_t) = \Pr[X_t = x_t] &= \sum_{i=1}^{m} \Pr[C_t = i] \Pr[X_t = x_t \,|\, C_t = i] \\
&= \sum_{i=1}^{m} u_i(t) p_i(x_t).
\end{aligned}
\tag{15}
$$

Puisqu'ici $\{C_t\}$ est une chaîne de Markov, $\mu_i(t)$ est obtenue en conditionnant sur la première transition de la chaîne. Ainsi

$$u(t) = u(1)\Gamma^{t-1}.$$

En notant $\mathbf{P}(x_t) = \begin{pmatrix} p_1(x_t) & & 0 \\ & \ddots & \\ 0 & & p_m(x_t) \end{pmatrix}$ et $\mathbf{1}' = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, l'équation (15) devient

$$
\begin{aligned}
p(x_t) &= u(t)\mathbf{P}(x_t)\mathbf{1}' \\
&= u(1)\Gamma^{t-1}\mathbf{P}(x_t)\mathbf{1}'.
\end{aligned}
\tag{16}
$$

(b) Si la chaîne de Markov est *stationnaire*, on obtient

$$
\begin{aligned}
p(x_t) &= \boldsymbol{\delta}\,\Gamma^{t-1}\mathbf{P}(x_t)\mathbf{1}' \\
&= \boldsymbol{\delta}\,\mathbf{P}(x_t)\mathbf{1}',
\end{aligned}
$$

où δ est la distribution stationnaire de la chaîne.

Distributions conjointes de plus grands ordres

Dans la preuve 2.1, nous considérions que $\{C_t\}$ était indépendant et donc que

$$\Pr[C_t = c_t, C_{t+k} = c_{t+k}] = \Pr[C_t = c_t] \Pr[C_{t+k} = c_{t+k}].$$

Ici, le processus $\{C_t\}$ n'est plus indépendant. Dans ce cas,

$$\Pr[C_t = c_t, C_{t+k} = c_{t+k}] = \Pr[C_t = c_t] \Pr[C_{t+k} = c_{t+k} \,|\, C_t = c_t].$$

Intéressons nous maintenant à la distribution de $[X_t, X_{t+k}]$

(a) En ne considérant pas nécessairement la chaîne de Markov comme *stationnaire*, on

obtient que

$$
\begin{aligned}
&\Pr[X_t = a, X_{t+k} = b] \\
&\quad = \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr[X_t = a, X_{t+k} = b, C_t = i, C_{t+k} = j] \\
&\quad = \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr[X_t = a, X_{t+k} = b \,|\, C_t = i, C_{t+k} = j] \Pr[C_t = i, C_{t+k} = j] \\
&\quad = \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr[X_t = a \,|\, C_t = i] \Pr[X_{t+k} = b \,|\, C_{t+k} = j] \Pr[C_t = i] \Pr[C_{t+k} = j \,|\, C_t = i] \\
&\quad = \sum_{i=1}^{m}\sum_{j=1}^{m} \underbrace{\Pr[C_t = i]}_{u_i(t)} \underbrace{\Pr[X_t = a \,|\, C_t = i]}_{p_i(a)} \underbrace{\Pr[C_{t+k} = j \,|\, C_t = i]}_{\gamma_{ij}^{(k)}} \underbrace{\Pr[X_{t+k} = b \,|\, C_{t+k} = j]}_{p_j(b)} \\
&\quad = \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t) p_i(a) \gamma_{ij}^{(k)} p_j(b).
\end{aligned}
$$

En notation matricielle, cette double somme devient

$$
\begin{aligned}
\Pr[X_t = a, X_{t+k} = b] &= u(t)\, \mathbf{P}(a)\, \Gamma^k\, \mathbf{P}(b)\, \mathbf{1}' \\
&= u(1)\, \Gamma^k\, \mathbf{P}(a)\, \Gamma^k\, \mathbf{P}(b)\, \mathbf{1}'. \qquad (17)
\end{aligned}
$$

(b) Si la chaîne de Markov est *stationnaire*, on obtient

$$
\begin{aligned}
\Pr[X_t = a, X_{t+k} = b] &= u(1)\, \Gamma^k\, \mathbf{P}(a)\, \Gamma^k\, \mathbf{P}(b)\, \mathbf{1}' \\
&= \boldsymbol{\delta}\, \mathbf{P}(a)\, \Gamma^k\, \mathbf{P}(b)\, \mathbf{1}'. \qquad (18)
\end{aligned}
$$

Ce résultat peut être étendu aux distributions conjointes de plus grands ordres. Ainsi, lorsque la chaîne est stationnaire, les distributions conjointes d'ordre 3 prennent la forme

$$\Pr[X_t = a, X_{t+k} = b, X_{t+k+n} = c] = \boldsymbol{\delta}\, \mathbf{P}(a)\, \Gamma^k\, \mathbf{P}(b)\, \Gamma^n\, \mathbf{P}(c)\, \mathbf{1}'.$$

3.2.1 Les moments de X_t dans un modèle de Markov caché

Les moments de $\{X_t\}$ sous un modèle de Markov caché sont sensiblement les mêmes que sous un modèle de mélange (section 2.2). Cependant, dans un modèle de Markov caché, le vecteur $\boldsymbol{\delta}$ est la distribution stationnaire du processus $\{C_t\}$.

Proposition 2.

Notons tout d'abord

$$\mu_i = \mathbb{E}[X_t \,|\, C_t = i]\,, \qquad \sigma_i^2 = \mathbb{V}\mathrm{ar}[X_t \,|\, C_t = i]\,,$$

$$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_m]\,, \qquad \boldsymbol{\delta} = [\delta_1, \ldots, \delta_m]\,,$$

$$\mathbf{M} = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_m \end{pmatrix}.$$

Alors

$$\mathbb{E}[X_t] = \boldsymbol{\delta}\boldsymbol{\mu}', \tag{19}$$

$$\mathbb{E}\left[X_t^2\right] = \sum_{i=1}^{m} \delta_i \left(\mu_i^2 + \sigma_i^2\right), \tag{20}$$

$$\mathbb{V}\mathrm{ar}[X_t] = \left[\sum_{i=1}^{m} \delta_i \left(\mu_i^2 + \sigma_i^2\right)\right] - \left[\sum_{i=1}^{m} \delta_i \mu_i\right]^2, \tag{21}$$

et l'espérance du produit de 2 réalisations du processus $\{X_t\}$ est

$$\mathbb{E}[X_t X_{t+k}] = \boldsymbol{\delta}\,\mathbf{M}\,\Gamma^{\,k}\boldsymbol{\mu}'. \tag{22}$$

Preuve.

(i) Espérance de X_t

$$\begin{aligned}
\mathbb{E}[X_t] &= \sum_{i=1}^{m} \Pr[C_t = i]\ \mathbb{E}[X_t \,|\, C_t = i] \\
&= \sum_{i=1}^{m} \delta_i \mathbb{E}[X_t \,|\, C_t = i] \\
&= \sum_{i=1}^{m} \delta_i \mu_i \\
&= \boldsymbol{\delta}\boldsymbol{\mu}'.
\end{aligned}$$

Avec $\boldsymbol{\mu}'$ la transposée du vecteur $\boldsymbol{\mu}$.

(ii) Espérance de X_t^2

$$\begin{aligned}
\mathbb{E}\big[X_t^2\big] &= \mathbb{E}\big[\mathbb{E}\big[X_t^2 \,|\, Ct\big]\big] \\
&= \sum_{i=1}^{m} \Pr[C_t = i]\ \mathbb{E}\big[X_t^2 \,|\, C_t = i\big]\,. \\
&= \sum_{i=1}^{m} \delta_i\ \mathbb{E}\big[X_t^2 \,|\, C_t = i\big]\,, \\
&= \sum_{i=1}^{m} \delta_i \left(\mu_i^2 + \sigma_i^2\right),
\end{aligned}$$

car

$$\begin{aligned}
\mathbb{E}\big[X_t^2 \,|\, C_t = i\big] &= \mathbb{V}\mathrm{ar}[X_t \,|\, C_t = i] + (\mathbb{E}[X_t \,|\, C_t = i])^2 \\
&= \mu_i^2 + \sigma_i^2.
\end{aligned}$$

(iii) Variance de X_t

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X_t] &= \mathbb{E}\big[X_t^2\big] - (\mathbb{E}[X_t])^2 \\
&= \left[\sum_{i=1}^{m} \delta_i \left(\mu_i^2 + \sigma_i^2\right)\right] - \left[\sum_{i=1}^{m} \delta_i \mu_i\right]^2.
\end{aligned}$$

(iv) Espérance du produit de 2 réalisations

$$\begin{aligned}
\mathbb{E}[X_t X_{t+k}] &= \sum_{i,j=1}^{m} \mathbb{E}[X_t X_{t+k} \,|\, C_t = i, C_{t+k} = j] \Pr[C_t = i, C_{t+k} = j] \\
&= \sum_{i,j=1}^{m} \Pr[C_t = i]\, \mathbb{E}[X_t \,|\, C_t = i] \Pr[C_{t+k} \,|\, C_t]\, \mathbb{E}[X_{t+k} \,|\, C_{t+k} = j] \\
&= \sum_{i,j=1}^{m} \delta_i\, \mu_i\, \gamma_{ij}^{(k)}\, \mu_j,
\end{aligned}$$

et donc,

$$\mathbb{E}[X_t X_{t+k}] = \boldsymbol{\delta}\, \mathbf{M}\, \Gamma^{\,k} \boldsymbol{\mu}'.$$

□

3.2.2 Fonction d'auto-correlation

La fonction d'auto-corrélation du processus $\{X_t\}$, que nous noterons $\rho[X_t X_{t+k}]$, est un indicateur de dépendance.

$$\begin{aligned}\rho[X_t, X_{t+k}] &= \frac{\mathrm{Cov}[X_t, X_{t+k}]}{\sqrt{\mathbb{V}\mathrm{ar}[X_t]\,\mathbb{V}\mathrm{ar}[X_{t+k}]}} \\ &= \frac{\mathrm{Cov}[X_t, X_{t+k}]}{\mathbb{V}\mathrm{ar}[X_t]} \\ &= \frac{\left(\boldsymbol{\delta}\,\mathbf{M}\,\Gamma^{\,k}\boldsymbol{\mu}'\right) - \left(\boldsymbol{\delta}\boldsymbol{\mu}'\right)^2}{\left[\sum_{i=1}^{m}\delta_i\left(\mu_i^2+\sigma_i^2\right)\right] - \left[\sum_{i=1}^{m}\delta_i\mu_i\right]^2},\end{aligned} \tag{23}$$

car

$$\mathrm{Cov}[X_t, X_{t+k}] = \mathbb{E}[X_t X_{t+k}] - \mathbb{E}[X_t]\,\mathbb{E}[X_{t+k}]\,.$$

3.3 Fonction de vraisemblance et optimisation d'un modèle de Markov caché

Dans la Section 2.3, il était question de la fonction de vraisemblance d'un modèle de mélange. Il sera maintenant question de la fonction de vraisemblance d'un modèle de Markov caché. Soit un échantillon de T observations $\{x_1, x_2, \ldots, x_T\}$. On suppose que celui-ci a été généré par un modèle de Markov caché. Puisque la fonction de vraisemblance de cet échantillon, notée L_T, est en fait la distribution conjointe de toutes les observations de l'échantillon, nous utiliserons les résultats obtenus dans la Section 3.2.

3.3.1 Fonction de vraisemblance

Proposition 3.
Soit

$$\alpha_t(i) = \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t, C_t = i]$$

et le vecteur

$$\boldsymbol{\alpha}_t = [\alpha_t(1)\ldots\alpha_t(m)]\,. \tag{24}$$

Alors,

$$\boldsymbol{\alpha}_t = u(1)\,\mathbf{P}(x_1)\,\Gamma\,\mathbf{P}(x_2)\ldots\Gamma\,\mathbf{P}(x_t) \tag{25}$$

et la vraisemblance du modèle est donnée par

$$L_T = \boldsymbol{\alpha}_T\mathbf{1}'. \tag{26}$$

Preuve.

(i) Considérons tout d'abord le cas $t = 1$.

$$\begin{aligned}\alpha_1(i) &= \Pr[X_1 = x_1, C_1 = i] \\ &= \Pr[C_1 = i]\ \Pr[X_1 = x_1 \,|\, C_1 = i]\end{aligned}$$

ce qui implique que

$$\boldsymbol{\alpha}_1 = u(1)\,\mathbf{P}(x_1).$$

Si $u(1) = \boldsymbol{\delta}$, alors on peut aussi écrire

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\,\Gamma\,\mathbf{P}(x_1).$$

(ii) Montrons maintenant que

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\,\Gamma\,\mathbf{P}(x_t). \tag{27}$$

Tout d'abord,

$$\begin{aligned}\alpha_t(i) &= \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t, C_t = i] \\ &= \Pr\big[X^{(t)} = x^{(t)}, C_t = i\big] \\ &= \sum_j \Pr\big[X^{(t)} = x^{(t)}, C_t = i, C_{t-1} = j\big] \\ &= \sum_j \Pr\big[X^{(t-1)} = x^{(t-1)}, C_{t-1} = j\big]\ \Pr\big[X_t = x_t, C_t = i \,|\, C_{t-1} = j, X^{(t-1)} = x^{(t-1)}\big] \\ &= \sum_j \alpha_{t-1}(j)\ \Pr\big[X_t = x_t, C_t = i \,|\, C_{t-1} = j, X^{(t-1)} = x^{(t-1)}\big] \\ &= \sum_j \alpha_{t-1}(j)\ \Pr[X_t = x_t, C_t = i \,|\, C_{t-1} = j] \\ &= \sum_j \alpha_{t-1}(j)\ \frac{\Pr[X_t = x_t, C_t = i, C_{t-1} = j]}{\Pr[C_{t-1} = j]} \\ &= \sum_j \alpha_{t-1}(j)\ \frac{\Pr[X_t = x_t \,|\, C_t = i, C_{t-1} = j]\,\Pr[C_t = i, C_{t-1} = j]}{\Pr[C_{t-1} = j]} \\ &= \sum_j \alpha_{t-1}(j)\ \frac{\Pr[X_t = x_t \,|\, C_t = i]\,\Pr[C_t = i, C_{t-1} = j]}{\Pr[C_{t-1} = j]} \\ &= \sum_j \alpha_{t-1}(j)\ \frac{\Pr[X_t = x_t \,|\, C_t = i]\,\Pr[C_t = i \,|\, C_{t-1} = j]\,\Pr[C_{t-1} = j]}{\Pr[C_{t-1} = j]} \\ &= \sum_j \alpha_{t-1}(j)\ \Pr[X_t = x_t \,|\, C_t = i]\,\Pr[C_t = i \,|\, C_{t-1} = j] \\ &= \sum_j \alpha_{t-1}(j)\,\gamma_{ji}\ p_i(x_t)\end{aligned}$$

et donc

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\,\Gamma\,\mathbf{P}(x_t).$$

(iii) Par induction, on obtient que

$$\boldsymbol{\alpha}_T = \boldsymbol{\delta}\,\mathbf{P}(x_1)\,\Gamma\,\mathbf{P}(x_2)\ldots\Gamma\,\mathbf{P}(x_T)$$

et donc que

$$L_T = \boldsymbol{\alpha}_T\mathbf{1}',$$

car

$$L_T = \sum_j \Pr\big[X^{(T)} = x^{(T)}, C_T = j\big] = \sum_j \alpha_T(j).$$

□

Remarque

Si la chaîne de Markov est *stationnaire* et l'échantillon est incomplet (c-à-d qu'il manque des données), on doit tenir compte du nombre de transitions d'écart entre les observations entourant ces inconnus.

Supposons par exemple que les observations x_2, x_3 et x_{T-1} soient manquantes. La fonction de vraisemblance, notée ici $L_T^{-(2,3,(T-1))}$, devient simplement,

$$L_T^{-(2,3,(T-1))} = \boldsymbol{\delta}\,\mathbf{P}(x_1)\,\Gamma^3\,\mathbf{P}(x_4)\,\Gamma\,\mathbf{P}(x_5)\ldots\Gamma\,\mathbf{P}(x_{T-2})\,\Gamma^2\,\mathbf{P}(x_T)\,\mathbf{1}'. \tag{28}$$

3.3.2 Optimisation avec contraintes

Lorsque la fonction de vraisemblance d'un modèle de mélange a été étudiée dans la Section 2.3, le code R présenté permettait de contraindre δ_1 entre 0 et 1. Tel que mentionné dans cette même section, cette technique n'est pas du tout efficace pour un nombre de composantes $m > 2$. Pour la suite de la présente section, nous considérerons un modèle de Markov caché dont les composantes sont des lois Poisson(λ_i).

Dans ce cas, les contraintes sur Γ et les λ_i sont les suivantes :

(i) $\sum_{i=1}^{m} \gamma_{ij} = 1, \quad \forall\, j \in \mathbb{S}_m,$

(ii) $\gamma_{ij} \geq 0, \quad \forall\, i, j \in \mathbb{S}_m,$

(iii) $\lambda_i \geq 0, \quad \forall\, i \in \mathbb{S}_m.$

Pour pouvoir contourner ces contraintes lors de l'optimisation, la matrice de transition Γ ainsi que les paramètres λ_i devront être reparamétrés.

Reparamétrisation

Nous utiliserons la reparamétrisation proposée dans le livre *Hidden Markov Models for Time Series : An Introduction Using R, Second Edition* [2].

1. Reparamétrer λ_i pour en assurer la positivité est simple grâce à la fonction exponentielle. Si

$$\eta_i = \log \lambda_i, \tag{29}$$

alors $\eta_i \in \mathbb{R}$ et n'est donc pas restreint. La maximisation est alors faite en fonction des paramètres η_i sans contraintes puis on applique la transformation inverse,

$$\lambda_i = e^{\eta_i}. \tag{30}$$

2. Pour reparamétrer la matrice Γ, il est utile de noter qu'elle contient m^2 entrées, mais que seulement $m(m-1)$ de celles-ci sont libres puisqu'il y a m contraintes sur ses lignes, soit

$$\gamma_{i1} + \gamma_{i2} + \ldots + \gamma_{im} = 1 \quad \forall\, i \in \mathbb{S}_m. \tag{31}$$

Prenons en exemple un modèle de Markov caché avec $m = 3$. Pour contourner les contraintes sur Γ, la matrice T sera définie comme suit :

$$T = \begin{pmatrix} 0 & \tau_{12} & \tau_{13} \\ \tau_{21} & 0 & \tau_{23} \\ \tau_{31} & \tau_{32} & 0 \end{pmatrix} \tag{32}$$

Dans notre cas, la fonction exponentielle est utilisée pour transformer les entrées γ_{ij} en τ_{ij}. Pour illustrer cette transformation, intéressons nous aux valeurs sur la première ligne de Γ,

$$\gamma_{11} = \frac{1}{1 + e^{\tau_{12}} + e^{\tau_{13}}},$$

$$\gamma_{12} = \frac{e^{\tau_{12}}}{1 + e^{\tau_{12}} + e^{\tau_{13}}}$$

et

$$\gamma_{13} = \frac{e^{\tau_{13}}}{1 + e^{\tau_{12}} + e^{\tau_{13}}}.$$

Avec, comme transformation inverse,

$$\tau_{11} = \log\left(\frac{\gamma_{11}}{\gamma_{11}}\right) = 0,$$
$$\tau_{12} = \log\left(\frac{\gamma_{12}}{\gamma_{11}}\right),$$
$$\tau_{13} = \log\left(\frac{\gamma_{13}}{\gamma_{11}}\right).$$

Ainsi, la somme sur une ligne de Γ vaut 1 et chaque entrée γ_{ij} est positive. La même transformation est bien sur effectuée sur les autres lignes de Γ.

Définition

Les paramètres η_i et τ_{ij} sont appelés les paramètres de *travail* et les paramètres λ_i et γ_{ij} sont appelés les paramètres *naturels*.

Fonctions dans R effectuant la reparamétrisation

Nous ferons toujours l'hypothèse que la chaîne de Markov est stationnaire pour la suite.

Paramètres *naturels* → Paramètres de *travail*

Voici une fonction R permettant de transformer les paramètres *naturels* en paramètres de *travail* tels que nous les avons définis :

```
1   MMC.trans.Naturels.Travail=function(m,lambda,gamma)
2   {
3       wlambda=log(lambda)
4       wgamma=NULL
5
6       if (m>1)
7       {
8           T=log(gamma/diag(gamma))
9           wgamma=as.vector(v[!diag(m)])
10
11      }
12      vecpar=c(wlambda,wgamma)
13      vecpar
14  }
```

1 - La fonction *MMC.trans.Naturels.Travail* prend comme entrées m, $lambda = [\lambda_1 \ \ldots \ \lambda_m]$ et $gamma = \Gamma$.

3 - Elle crée le vecteur $wlambda = [\eta_1 \ \ldots \ \eta_m] = \log(lambda)$.

8 - Elle construit T en calculant le logarithme naturel de la matrice Γ divisée par sa diagonale.

9 - Les valeurs non nulles de T sont misent sous forme de vecteur.

12 - La fonction concatène les vecteurs *wlambda* et *wgamma* pour former *vecpar* qui sera ensuite utilisé lors de la maximisation.

Paramètres de *travail* → Paramètres *naturels*

La fonction R permettant d'effectuer la transformation inverse est :

```
1   MMC.trans.Travail.Naturels=function(m,vecpar)
2   {
3       exppar=exp(vecpar)
4       lambda=exppar[1:m]
5       exppar=exppar[-c(1:m)]
6
7       gamma=diag(m)
8       gamma[!gamma]=exppar
9
10      gamma=gamma/apply(gamma,1,sum)
11      delta=c(rep(1,m))  %*%  solve(diag(m)-gamma+1)
12      list(lambda= lambda,gamma=gamma,delta=delta)
13  }
```

1 - La fonction *MMC.trans.Travail.Naturels* utilise le vecteur *vecpar* tel que défini dans la fonction *MMC.trans.Naturels.Travail*.

3 - Elle applique l'exponentielle à toutes les entrées de *vecpar*.

4 - Elle assigne les m premières entrées *exppar* au vecteur *lambda*.

5 - Elle retire ces m observations du vecteur *exppar*.

7 - La fonction définie *gamma* comme une matrice identité de taille $(m \times m)$.

8 - Les entrées nulles de *gamma* sont remplacées par les $m(m-1)$ valeurs de *exppar* restantes.

10 - La matrice *gamma* est divisée par la somme de ses lignes.

11 - La fonction calcule la distribution stationnaire *delta* de la matrice de transition Γ obtenue à la ligne 10.

12 - Les noms des différents paramètres sont enregistrés.

3.3.3 Construction de la fonction de vraisemblance

Régularisation de la fonction de vraisemblance

Puisque nous avons défini α_t comme un produit de probabilités, sa valeur décroit rapidement lorsque t augmente, pour être éventuellement arrondie 0. La technique proposée par le livre *Hidden Markov Models for Time Series : An Introduction Using R, Second Edition* [2] consiste à pondérer les éléments du vecteur α_t en définissant $\forall\, t \in \{1, \ldots, T\}$:

$$\phi_t = \frac{\boldsymbol{\alpha}_t}{\omega_t}, \tag{33}$$

avec

$$\omega_t = \sum_i \alpha_t(i) = \boldsymbol{\alpha}_t \mathbf{1}'. \tag{34}$$

ω_t est donc la fonction de vraisemblance des t premières observations. Ainsi, le paramètre ϕ_t est construit de sorte que lorsqu'un terme est ajouté à la fonction de vraisemblance, chacun des m termes de α_t est divisé par la fonction de vraisemblance issue des t premières observations.

Proposition 4. Sous cette notation, la fonction de vraisemblance devient

$$L_T = \prod_{t=1}^{T} \left(\frac{\omega_t}{\omega_{t-1}} \right) \tag{35}$$

et la fonction de log-vraisemblance devient

$$l_T = \sum_{t=1}^{T} \log \left(\frac{\omega_t}{\omega_{t-1}} \right) = \sum_{t=1}^{T} \log \left(\phi_{t-1}\, \Gamma\, \mathbf{P}(x_t)\, \mathbf{1}' \right). \tag{36}$$

Preuve. À partir des définitions de ϕ_t et ω_t, on obtient que

$$\begin{aligned} \omega_t\, \phi_t &= \boldsymbol{\alpha}_t \\ &= \boldsymbol{\alpha}_{t-1}\, \Gamma\, \mathbf{P}(x_t) \\ &= \omega_{t-1}\, \phi_{t-1}\, \Gamma\, \mathbf{P}(x_t) \end{aligned}$$

et donc

$$\omega_t = \omega_t\, \phi_t\, \mathbf{1}' = \omega_{t-1} \left(\phi_{t-1}\, \Gamma\, \mathbf{P}(x_t)\, \mathbf{1}' \right),$$

car

$$\phi_t\, \mathbf{1}' = \phi_t.$$

La fonction de vraisemblance

$$\begin{aligned} L_T &= \boldsymbol{\alpha}_T \, \mathbf{1}' \\ &= \omega_T \, \phi_T \, \mathbf{1}' \\ &= \omega_T \, (\phi_T \, \mathbf{1}') \\ &= \omega_T \\ &= \prod_{t=1}^{T} \left(\frac{\omega_t}{\omega_{t-1}} \right) \end{aligned}$$

et donc la fonction de log-vraisemblance, notée l_T, est

$$\begin{aligned} l_T &= \log \left(\prod_{t=1}^{T} \left(\frac{\omega_t}{\omega_{t-1}} \right) \right) \\ &= \sum_{t=1}^{T} \log \left(\frac{\omega_t}{\omega_{t-1}} \right) \\ &= \sum_{t=1}^{T} \log \left(\phi_{t-1} \, \Gamma \, \mathbf{P}(x_t) \, \mathbf{1}' \right). \end{aligned}$$

□

Fonctions dans R construisant récursivement la vraisemblance

Voici une fonction R permettant de construire la fonction vraisemblance régularisée :

```
MMC.LLH.Stationnaire.PoisDist=function(vecpar,x,m)
{
   if (m==1) return(-sum(dpois(x,exp(vecpar),log=T)))


   pn=MMC.trans.Travail.Naturels(m,vecpar)

   n=length(x)

   toutesprob=outer(x,pn$lambda,dpois)

   toutesprob=ifelse(!is.na(toutesprob),toutesprob,1)

   phi=pn$delta
   Regulog=0

   for (i in 1:n)
   {
      phi = phi  %*%  pn$gamma  *  toutesprob[i,]
      omega=sum(phi)

      Regulog=Regulog+log(omega)

      phi=phi/omega
   }
```

```
26
27     mllh=-Regulog
28     mllh
29
30 }
```

3.3.4 Optimisation numérique dans R

Vecteur de paramètres initiaux

Les fonctions dans R permettant de minimiser une fonction (notamment *nlm* et *optim*) nécessitent des valeurs initiales pour tous les paramètres à estimer. La fonction dans R ci-dessous permet de construire ce vecteur initial en fonction de l'échantillon x à modéliser. Les éléments *lambda0* et *gamma0* sont respectivement une estimation des paramètres des m distributions et une estimation des probabilités de transition.

```
1  lambda0.gamma0.construction=function(m,x)
2  {
3
4    mX=mean(x,na.rm=T)
5    sdX=sd(x,na.rm=T)
6
7      if (is.integer(m/2))
8      {
9        centre=m/2+0.5
10     }
11     else
12     {
13       centre=round(m/2)+1
14     }
15     gamma0=NULL
16
17     lambda0=NULL
18
19     for (j in 1:m)
20     {
21       lambda0[j]=mX+(j-centre)*(sdX)
22     }
23
24     lambda0=sort(lambda0,decreasing=F)
25
26     gamma0=matrix(1/m,nrow=m,ncol=m)
27     return(list(lambda0=lambda0,gamma0=gamma0))
28 }
```

1 - La fonction *lambda0.gamma0.construction* utilise l'échantillon x et le nombre d'états m.

4 - Elle calcule la moyenne de l'échantillon.

5 - Elle calcule l'écart-type de l'échantillon.

(7 - 14) - Elle trouve le point central entre 1 et m.

(19 - 22) - La fonction construit *lambda0* en fonction de la moyenne et de l'écart-type.

24 - La fonction s'assure que le vecteur *lambda0* est en ordre croissant.

26 - Toutes les entrées de la matrice *gamma0* sont estimées à $\frac{1}{m}$.

27 - Les noms des différents paramètres sont enregistrés puis ces paramètres sont renvoyés par la fonction.

Optimisation et organisation des résultats

Voici une fonction dans R permettant de calculer le vecteur initial, d'optimiser la fonction de vraisemblance puis d'organiser le retour des résultats :

```
1  pois.MMC.emv=function(x,m, nDigits)
2  {
3      parvect0naturel=lambda0.gamma0.construction(m,x)
4         lambda0=parvect0naturel$lambda0
5         gamma0=parvect0naturel$gamma0
6
7      parvect0= MMC.trans.Naturels.Travail(m,lambda0,gamma0)
8      mod=nlm(MMC.LLH.Stationnaire.PoisDist,parvect0,x=x,m=m)
9      pn= MMC.trans.Travail.Naturels(m,mod$estimate)
10     mllh =mod$minimum
11
12     list(
13       lambda=pn$lambda,
14       gamma=round(pn$gamma,digits=nDigits),
15       delta=pn$delta,
16       mllh=mllh
17      )
18 }
```

1 - La fonction *pois.MMC.emv* utilise l'échantillon x et le nombre d'états m et permet de spécifier le nombre de décimales à afficher dans la matrice de transition *gamma*.

3 - Elle applique la fonction *lambda0.gamma0.construction* pour obtenir un vecteur initial *parvect0natural*.

(4 - 5) - Elle extrait les paramètres *naturels* initiaux de *parvect0natural*.

7 - Les paramètres *naturels* initiaux obtenus sont transformés en paramètre de *travail*.

8 - La fonction *nlm* minimise l'équation définie par *MMC.LLH.Stationnaire.PoisDist* correspondant à $-L_T$.

9 - Les paramètres obtenue par minimisation sont enregistrés sous le nom *pn*.

(12 - 17) - Les résultats de l'optimisation sont enregistrés.

3.3.5 Exemple numérique

Voici les résultats obtenues pour des modèles de Markov cachés à 1, 2 et 3 états lorsque nous étudions l'échantillon de la table 1 grâce aux 5 fonctions définies plus haut :

Modèle de Markov admettant 1 état

Considérer qu'il n'y a pas indépendance dans le processus $\{C_t\}$ ne change rien quand on modélise avec un seul état. Tel que dans la Section 1,

$$\widehat{\lambda_1} = 19{,}36449.$$

La vraisemblance de ce modèle est

$$L_T = -391{,}9189.$$

Modèle de Markov admettant 2 états

$$\begin{array}{ll} \widehat{\lambda_1} = 15{,}47223\ , & \widehat{\delta_1} = 0{,}6608189\ , \\ \widehat{\lambda_2} = 26{,}12535\ , & \widehat{\delta_2} = 0{,}3391811 \end{array}$$

$$\text{et } \Gamma = \begin{pmatrix} 0{,}9340 & 0{,}0660 \\ 0{,}1285 & 0{,}8715 \end{pmatrix}.$$

La vraisemblance de ce modèle est

$$L_T = -342{,}3183.$$

Modèle de Markov admettant 3 états

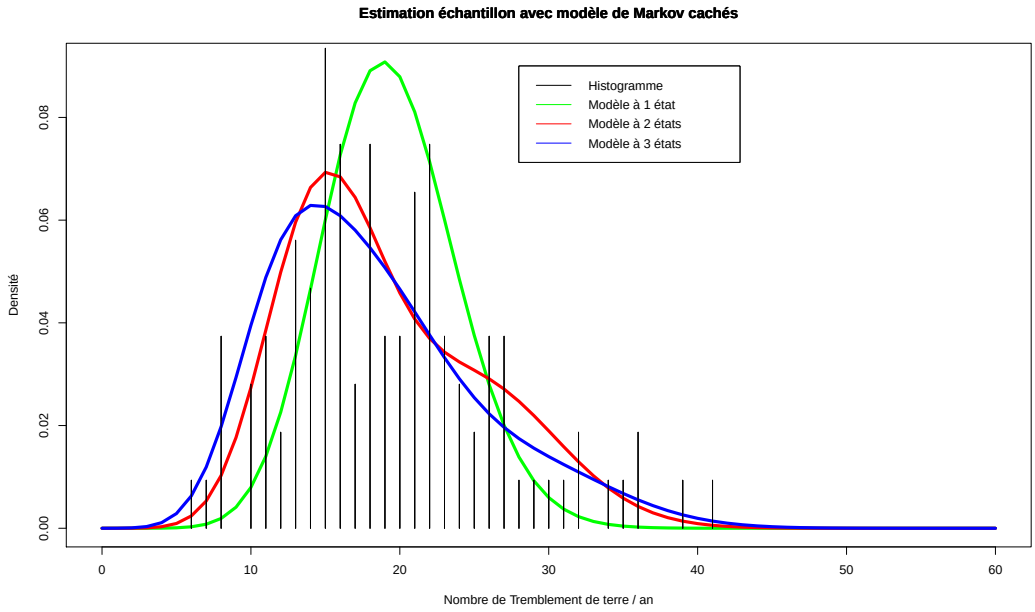
$$\begin{array}{ll} \widehat{\lambda_1} = 13{,}14573\ , & \widehat{\delta_1} = 0{,}4436431\ , \\ \widehat{\lambda_2} = 19{,}72101\ , & \widehat{\delta_2} = 0{,}4044976\ , \\ \widehat{\lambda_3} = 29{,}71437\ , & \widehat{\delta_3} = 0{,}1518593 \end{array}$$

$$\text{et } \Gamma = \begin{pmatrix} 0{,}9546 & 0{,}0244 & 0{,}0209 \\ 0{,}0498 & 0{,}8994 & 0{,}0509 \\ 0 & 0{,}1966 & 0{,}8034 \end{pmatrix}.$$

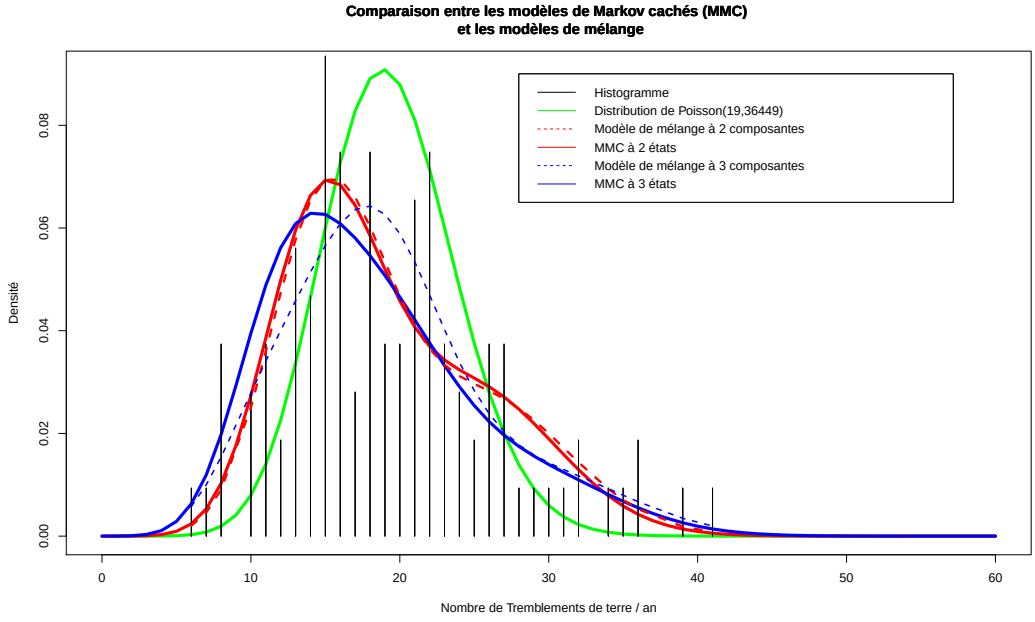
La vraisemblance de ce modèle est

$$L_T = -329{,}4603.$$

Graphiquement, la distribution des modèles à 1, 2 et 3 états ressemblent à :



Et voici un graphique illustrant la différence entre les modèles mixtes et les modèles de Markov cachés pour un même nombre de composantes :



[↑]

Références

[1] S. Lessard. *Processus stochastiques : Cours et exercices corrigés*. Références sciences. Ellipses, 2014.

[2] W. Zucchini, I.L. MacDonald, and R. Langrock. *Hidden Markov Models for Time Series : An Introduction Using R, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2016.