Modélisation de séries temporelles

V. Monbet - 2011

Table des matières

1 Rappels - Modèles ARMA **4**

1.1 Introduction . . . 4
1.1.1 Exemples dévellopés dans ce cours . . . 4
1.1.2 Objectifs . . . 4
1.1.3 Eléments consitutifs d'une série temporelle . . . 5
1.2 Analyse de la tendance . . . 5
1.2.1 Moyenne mobile . . . 5
1.2.2 Différenciation . . . 6
1.2.3 Modèle additif avec tendance paramétrique . . . 6
1.2.4 Comparaison des différentes méthodes . . . 8
1.2.5 Conclusion . . . 8
1.3 Analyse des composantes saisonnières . . . 8
1.3.1 Calcul des moyennes saisonnières . . . 9
1.3.2 Vérification du modèle . . . 9
1.3.3 Différenciation . . . 11
1.3.4 Comparaison des différentes méthodes . . . 11
1.4 Modélisation de la composante stationnaire . . . 11
1.4.1 Généralités sur les processus stationnaires . . . 11
1.4.2 Modèles autorégressifs d'ordre 1 . . . 13
1.4.3 Modèles autorégressifs d'ordre p . . . 14
1.4.4 Estimation de la moyenne et de la fonction d'autocovariance d'un processus stationnaire . . . 16
1.4.5 Inférence statistique pour les modèles autorégressifs . . . 16
1.4.6 Prédiction dans les modèles autorégressifs . . . 18
1.4.7 Sélection de modèle . . . 19

2 Modèles à changement de régimes markovien **21**

2.1 Introduction . . . 21
2.1.1 Généralités, exemples . . . 21
2.1.2 Modèle . . . 22
2.2 Modèles de mélange . . . 24
2.2.1 Introduction . . . 24
2.2.2 Modèle . . . 25
2.2.3 Inférence . . . 26
2.2.4 Algorithme EM . . . 27
2.2.5 Cas du mélange de lois de Gauss . . . 29
2.2.6 Propriétés de l'algorithme EM . . . 30
2.3 Chaînes de Markov discrètes . . . 31
2.3.1 Définitions et exemple . . . 31

2.3.2 Inférence pour les probabilités de transition 34

3 Inférence dans les modèles à changement de régimes markovien 36
3.1 Fonction de vraisemblance 36
3.2 Apprentissage supervisé : *étape M* 37
3.3 Classification : *étape E* 38
3.3.1 Algorithme de Viterbi 38
3.3.2 Algorithme Forward-Backward (FB) 39
3.4 Algorithme EM 40

4 Validation de modèles 41
4.1 Introduction 41
4.2 Tests d'adéquation 41
4.2.1 Observations continues 41
4.2.2 Observation discrètes 42
4.3 Critères basés sur la vraisemblance 42
4.3.1 Entropy 43
4.4 Méthodes de Monte Carlo 43

Chapitre 1

Rappels - Modèles ARMA

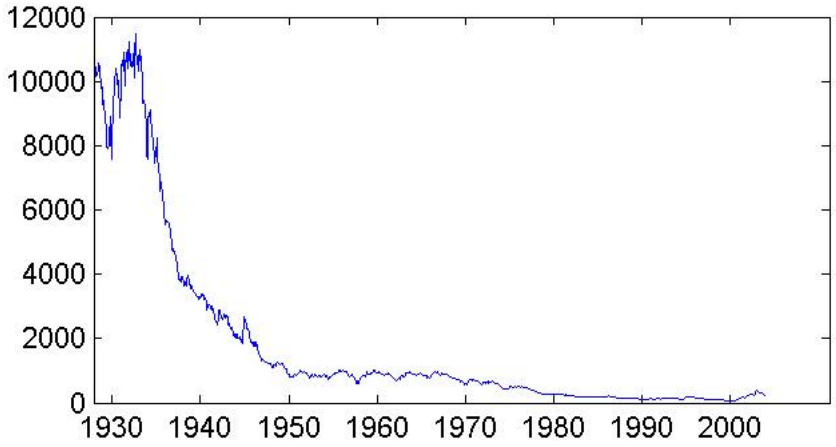
1.1 Introduction

Définition - Une série temporelle (ou chronologique) est une suite d'observations $x_1, x_2, \cdots, x_n$ indexée par le temps. On supposera qu'il s'agit d'une réalisation d'un processus X, c'est à dire d'une suite $\{X_i\}$ de variables aléatoires.

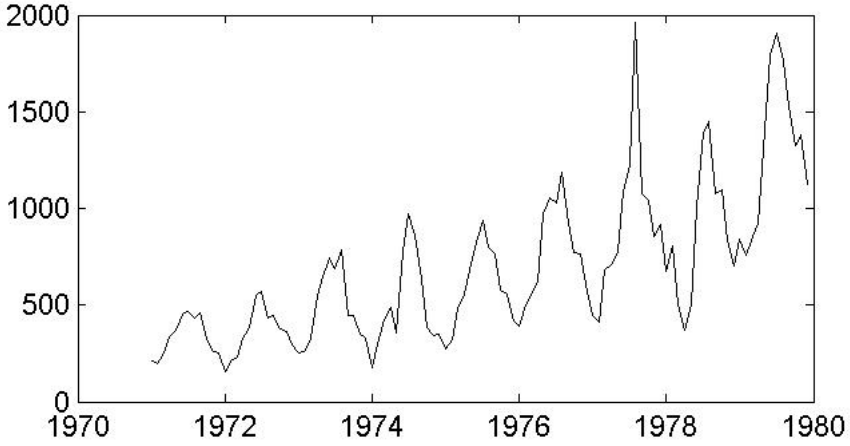
1.1.1 Exemples dévellopés dans ce cours

– **Economie**

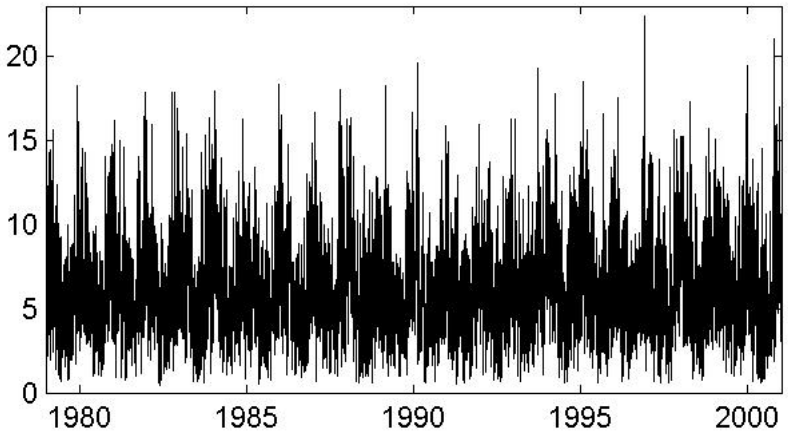
1. évolution du cours du Dow Jones entre 1928 et 2004, données mensuelles



2. production de poissons, en milliers de francs, en Finistère nord (Brest, Morlaix, Paimpol) entre 1971 et 1979, données mensuelles



– **Environement** - évolution de l'intensité du vent, en m/s, au large de la Bretagne entre 1979 et 2001, données journalières.



– **Démographie** - Exemple 4 : évolution de la population des Etats Unis, en millions d'habitants, entre 1790-1980, données décénales.

1.1.2 Objectifs

Les principaux objectifs de la modélisation des séries temporelles sont les suivants.

– Decrire. Par exemple,
  – en économétrie, détecter puis ananlyser les périodes de crises et croissances ;
  – en reconnaissance vocale, reconnaitre les mots dans des signaux ;
  – dans le séquençage du génome, détecter les parties de l'ADN qui contiennent de l'information.
– Comparer deux séries temporelles. Par exemple, l'évolution démographique de deux régions ou deux séquences d'ADN.
– Prédire l'évolution future de la série temporelle à partir de celles qui ont été observées. Par exemple, la température ou le cours d'une action du lendemain ou l'évolution de la population mondiale au cours du siècle prochain.

1.1.3 Eléments consitutifs d'une série temporelle

Une série temporelle est généralement constituée de plusieurs éléments.

- Tendance : représente l'évolution à long terme de la série (échelle interanuelle). Exemples : croissance économique, évolution climatologique à long terme (cyclique ou non)
- Saisonnalité : évolution se répétant régulierement tous les ans. Exemples :
  - En météorologie, température plus faibles en hiver qu'en été.
  - En économie, saisonnalité induite par les périodes de vacances, les périodes de fêtes, le climat...
- Composante stationnaire (ou résiduelle) : ce qui reste lorsque l'on a enlevé les autres composantes. Décrit l'évolution à court terme de la série (échelle journalière).

Notion de série temporelle stationnaire définie plus précisemment dans la suite. Cette hyppothèse jouera un rôle fondamentale dans la suite, et remplacera l'hyppothèse usuelle des v.a i.i.d. (ici, il peut exister une dépendance entre deux valeurs successives prises par la série observée).

Le modèle le plus usuel consiste à supposer que la série initiale s'écrit sous la forme (modèle additif)

$$X_t = T_t + S_t + Y_t \text{ pour tout } t \in \{1, \cdots, n\}$$

avec X_t la tendance, S_t la composante saisonnière (fonction périodique de période un an) et Y_t la composante stationnaire.

Les différentes étapes de la modélisation sont alors

1. modéliser la tendance
2. modéliser la composante saisonnière
3. modéliser la série résiduelle

1.2 Analyse de la tendance

exemple : poissons

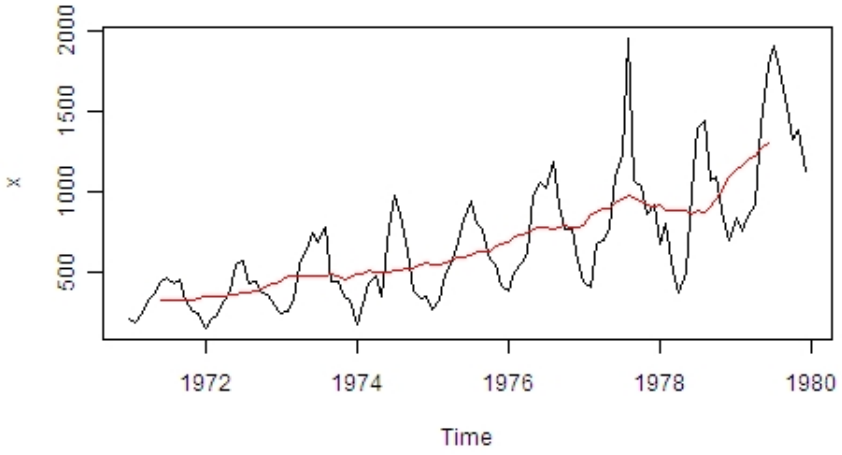
1.2.1 Moyenne mobile

Si $\{x_i\}_{i\in\{1,\cdots,n\}}$ est une série temporelle, alors la moyenne mobile d'ordre p associée est la série temporelle définie pour $t \in \{p+1, \cdots, n-p\}$ par

$$\hat{x}_t = \sum_{i=t-p}^{p} x_i$$

avec $2p+1$ la largeur de la fenêtre.

Exemple : Poissons



1.2.2 Différenciation

On peut éliminer la tendance par différenciation.

Définition - Notons ∇ l'opérateur de différenciation, c'est à dire l'opérateur défini par

$$\nabla x_t = x_t - x_{t-1}$$

pour tout $t \geq 2$.

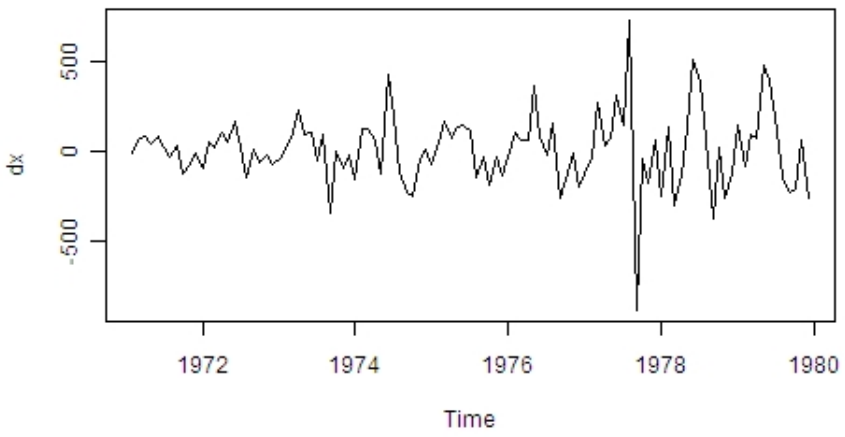
Définition - On définit l'opératuer de différenciation d'ordre k par la formule de récurrence

$$\nabla^{(k)} x_t = \nabla(\nabla^{(k-1)} x_t)$$

Propriétés -

1. Soient u_t et v_t deux séries temporelles, $\nabla(u_t + v_t) = \nabla u_t + \nabla v_t$.
2. Soit x_t une série temporelle et $\lambda \in \mathbb{R}$, $\nabla(\lambda x_t) = \lambda \nabla x_t$.
3. Soit $y_t = a_0 + a_1 t + \cdots + a_k t^k$, alors $\nabla^{(k)} y_t = k! a_k$ et $\nabla^{(k+1)} y_t = 0$.

La 3ème propriété implique que si $X_t = T_t + z_t$ avec T_t une tendance polynomiale, alors on peut supprimer la tendance en appliquant successivement plusieurs fois l'opérateur ∇.



1.2.3 Modèle additif avec tendance paramétrique

Définition On dit qu'une série temporelle suit un modèle de tendance additif lorsqu'elle peut se décomposer sous la forme

$$X_t = T_t + Z_t$$

avec Z_t une série temporelle sans tendance.

On dit qu'une série temporelle a une tendance linéaire lorsque

$$T_t = a \times t + b$$

On dit qu'une série temporelle a une tendance polynomiale lorsque

$$T_t = a_0 + a_1 t + a_2 t^2 + ...$$

Ajustement du modèle

On suppose que les observations $\{x_1, \cdots, x_n\}$ suivent un modèle additif avec une tendance paramétrique représentée par une fonction f de paramètres θ, c'est à dire vérifie

$$x_t = f(t; \theta) + z_t$$

avec z_t une série temporelle sans tendance.

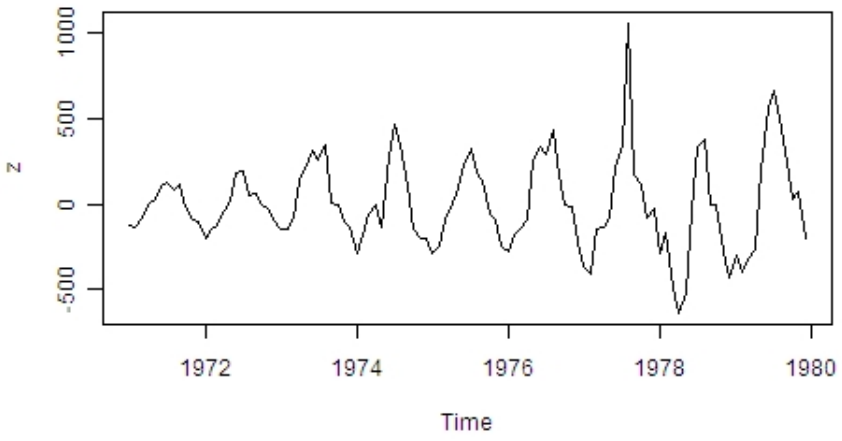
On cherche alors à estimer les paramètres inconnus θ. On utilise généralement la méthode des mointes carrés.

Vérification du modèle et autres modèles paramètriques
Afin de vérifier l'adéquation du modèle aux observations, on calcule les estimations $\hat{\theta}$ puis on trace le graphique

$$t \longmapsto x_t - f(t; \hat{\theta})$$

Si le modèle est valable, il ne reste plus de tendance sur cette nouvelle série temporelle. Sinon on cherche un modèle plus adapté.

Série résiduelle après avoir retiré la tendance estimée (polynôme d'ordre 2)



Il reste une tendance visible. Sur la série initiale, il semble que les fluctuations saisonnières sont proportionnelles à la tendance. Dans ce cas, un modèle additif n'est pas adapté. On peut alors tester un modèle multiplicatif.

Définition - On dit qu'une série temporelle suit un modèle de tendance multiplicatif lorsqu'elle peut se décomposer sous la forme

$$X_t = T_t Z_t$$

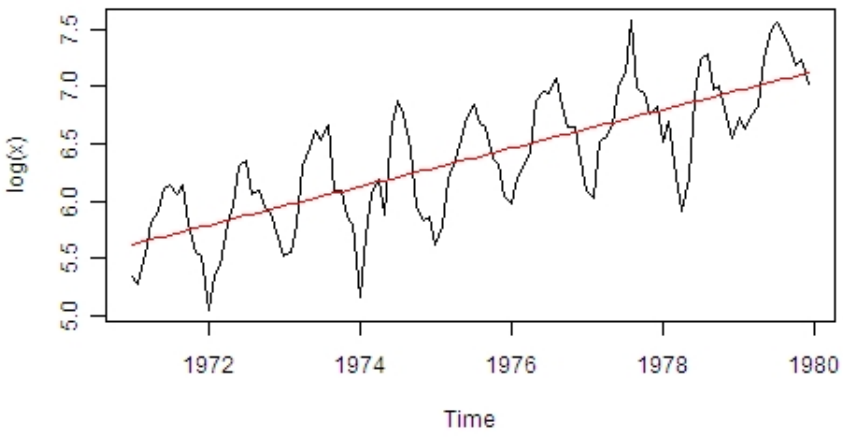
avec Z_t une série temporelle sans tendance.

En pratique, on se ramène à un modèle additif par passage au logarithme, puisqu'on a alors

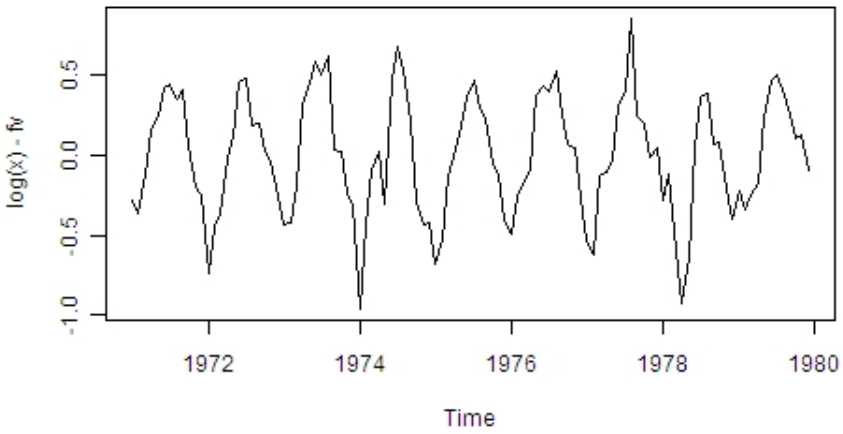
$$\log(X_t) = \log(T_t) + \log(Z_t)$$

Retour sur l'exemple de la production de poisson en finistère nord entre 1971 et 1979.

Logarithme de la série initiale et tendance linéaire ajustée



Série résiduelle après avoir retirée la tendance estimée



1.2.4 Comparaison des différentes méthodes

Les principaux avantages des méthodes paramètriques sur les deux autres méthodes sont

- qu'elles fournissent un modèle facilement interprétable et qui permet, par exemple, de comparer plusieurs séries temporelles entre elles.
- qu'elles peuvent être utilisées en prévision, ce qui n'est pas le cas des autres méthodes.

L'avantage principal des deux premières méthodes (moyenne mobile et différenciation) est qu'elles s'ajustent à de nombreuses séries temporelles sans modifications, alors que pour les modèles paramètriques il peut être difficile de choisir le bon modèle.

1.2.5 Conclusion

Pour modéliser la tendance dans une série temporelle, les différentes étapes sont :

1. Tracer la série temporelle
2. Utiliser ce graphique pour identifier le modèle en répondant aux quetions suivantes :
3. Modèle additif ou multiplicatif ? Pour cela, on regarde si les composantes saisonnières et stationnaires sont proportionnelles à la tendance ou non. Lorsque le modèle est additif, on travaille sur la série afin de se ramener à un modèle additif.
4. Méthode paramétrique ou non-paramétrique ? Pour cela, on regarde si il existe un modèle simple permettant de décrire la tendance observée (modèle linéaire, polynomiale, exponentiel...). Si on ne trouve pas de tel modèle, on peut utiliser une méthode non paramètrique (MM ou différentiation).
5. Vérifier le modèle en traçant la série "résiduelle", c'est à dire la série dans laquelle on a enlevé la tendance en utilisant la méthode choisie.

1.3 Analyse des composantes saisonnières

Dans ce paragraphe, X_t désigne une série temporelle sans tendance. En pratique, lorsqu'il existe une tendance sur la série initiale X_t, on commence par l'enlever en utilisant l'une des méthodes décrites dans la section précédente.

On supposera dans la suite de cette section, sauf mention contraire, que la série X_t suit un modèle de saisonnalité additif, c'est à dire que

$$X_t = S_t + Z_t$$

avec S_t une fonction périodique de période τ, avec τ le nombre de données par année, c'est à dire vérifiant $S_t = S_{t+\tau}$ et Z_t une série temporelle stationnaire.

De nombreuses séries temporelles observées dans la nature suivent un modèle multiplicatif, c'est à dire vérifient

$$X_t = S_t Z_t$$

Dans ce cas, les fluctuations de la série autour de la composante saisonnière sont proportionnelles à celle-ci. On se ramène alors à un modèle additif par passage au logarithme.

Dans le cadre des modèles additifs, plusieurs méthodes peuvent être utilisées pour estimer la fonction $t \longmapsto S_t$. C'est l'objet des paragraphes ci-dessous.

1.3.1 Calcul des moyennes saisonnières

On renumérote la série X_t sous la forme X_{jk}, $j = 1, \cdots, T_{an}$, $k = 1, \cdots, N_{an}$ avec T_{an} la longueur de l'année et N_{an} le nombre d'années de mesure. Le modèle se réécrit alors sous la forme :

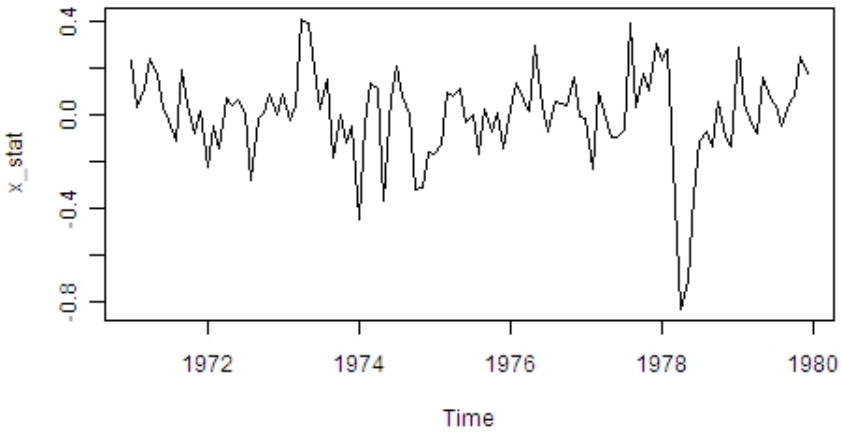
$$X_{jk} = S_k + Z_{jk}$$

Un estimateur naturel de S_k est donné par $\frac{1}{T}\sum_{j=1}^{N_{an}} Z_{jk}$, c'est à dire par la moyenne des observations disponibles à la date k pour chaque année. On appelle généralement S_k les "moyennes saisonnières".

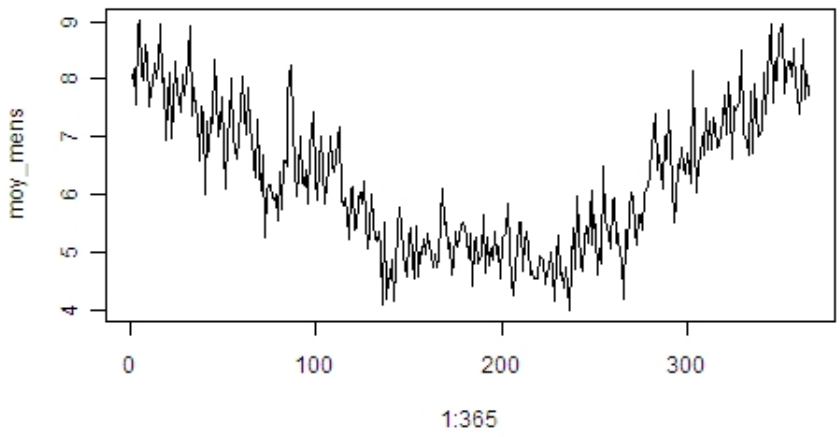
1.3.2 Vérification du modèle

Afin de vérifier l'adéquation du modèle, on trace la série résiduelle Z_t. Si le modèle convient, il ne reste plus de saisonnalité apparente sur cette série résiduelle, généralement appelée "série corrigée des variations saisonnières".

Retour sur l'exemple de la production de poissons.
Première étape : traitement de la tendance (cf section 2). On suppose que la série suit un modèle multiplicatif, i.e $X_t = T_t Z_t$. On suppose en outre que $\log(T_t) = \alpha t + \beta_t$. On note a et b les estimateurs des moindres carrés de α et β.
Deuxième étape : traitement de la saisonnalité. On suppose que la série $\log(Z_t)$ suit un modèle de saisonnalité additif, c'est à dire que $\log(Z_t) = S_t + Y_t$. On estime S_t en calculant les moyennes saisonnières associés à la série $\log(Z_t)$. On a $T_{an} = 12$, données mensuelles.



Retour sur l'exemple des données de vent.
On suppose qu'il n'y a pas de tendance dans cette série temporelle. On a $N = 22 \times 365$ (22 ans de données) et $T_{an} = 365$ (données journalières). La séquence S_k représente alors la moyenne associée au jour k de l'année calculée sur les 22 ans de données.



Sur cet exemple, la fonction obtenue n'est pas un bon estimateur de la fonction S_k : les fluctuations rapides observées sont dues au fait qu'on a seulement 22 données pour estimer la valeur de $\hat{s}_k$ pour un k donné. Dans ce cas, on peut lisser la série obtenue en utilisant les moyennes mobiles introduite précédemment. Remarque : ici on cherche à estimer une fonction périodique supposée être périodique à partir d'une observation "bruitée" x_t.

Modèle paramètrique
Lorsque l'on veut décrire la composante saisonnière grâce à un modèle paramétrique simple, on utilise généralement un polynôme trigonomètrique. On suppose alors que :

$$X_t = \mu + \alpha_c \cos(\omega t) + \alpha_s \sin(\omega t) + Z_t$$

Pour estimer α_c et α_s, on peut utiliser la méthode des moindres carrés.

Proposition - Les estimateurs des moindres carré de sont solutions du système linéaire :

$$\begin{vmatrix} T & \sum_{t=1}^T \cos(\omega t) & \sum_{t=1}^T \sin(\omega t) \\ \sum_{t=1}^T \cos(\omega t) & \sum_{t=1}^T (\cos(\omega t))^2 & \sum_{t=1}^T \cos(\omega t)\sin(\omega t) \\ \sum_{t=1}^T \sin(\omega t) & \sum_{t=1}^T \sin(\omega t)\cos(\omega t) & \sum_{t=1}^T (\sin(\omega t))^2 \end{vmatrix} \begin{bmatrix} \mu \\ \alpha_c \\ \alpha_s \end{bmatrix} = \begin{vmatrix} \sum_{t=1}^T x(t) \\ \sum_{t=1}^T x(t)\cos(\omega t) \\ \sum_{t=1}^T x(t)\sin(\omega t) \end{vmatrix}$$

Preuve : en exercice. Il suffit d'écrire le coût aux moindres carrés et de dériver.

Remarque : la solution analytique de ce système est relativement complexe (résolution numérique sous Matlab). Par contre, lorsque T est grand, une solution approchée est donnée par

$$\hat{\mu} \frac{1}{T} \sum_{t=1}^T x(t)$$

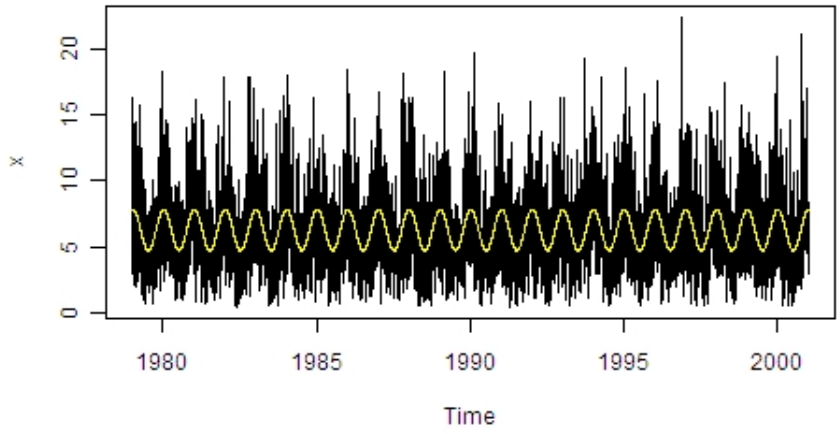
$$\hat{\alpha}_c = \frac{2}{T} \sum_{t=1}^T x(t) \cos(\omega t)$$

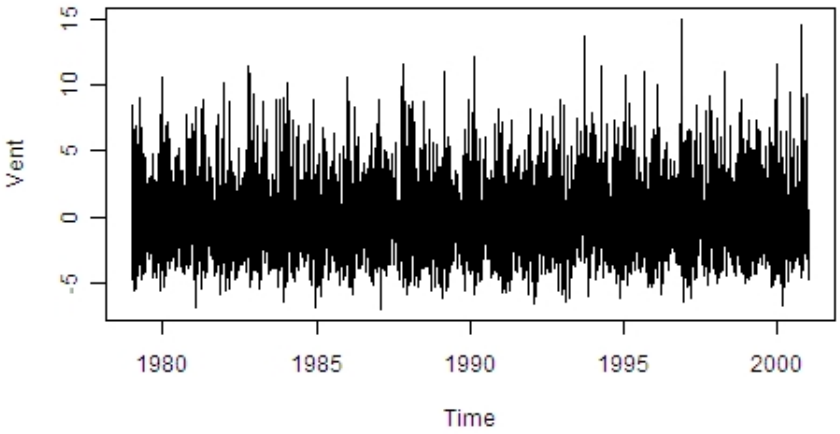
et

$$\hat{\alpha}_s = \frac{2}{T} \sum_{t=1}^T x(t) \sin(\omega t)$$

On reconnait les coefficients de Fourrier....
Ces estimateurs sont asymptotiquement sans biais et convergents.

La résolution du système précédent pour les données de vent donne $\hat{\mu} = 29.3$, $\alpha_c = 21,8$ et $\alpha_s = -7.7$.





Comparaison des différentes méthodes (p=2 mois)
Etude de la série résiduelle (modèle paramètrique) Il reste une forte saisonnalité. En fait, ici un modèle le modèle additif n'est pas adapté (cf TP ?).

1.3.3 Différenciation

Définition - On appelle différence d'ordre τ de la série X_t les quantités

$$\nabla_t^{(\tau)} = X_{t+\pi} - X_t$$

définies pour $\pi \in \mathbb{N}$. Attention : ne pas confondre cet opérateur avec l'opérateur défini à la section 2 (modélisation de la tendance).

Si on suppose que la série $\{X_t\}$ suit un modèle de saisonnalité additif, c'est à dire que

$$X_t = S_t + Z_t$$

avec S_t une fonction périodique de période τ, avec τ le nombre de données par année, c'est à dire vérifiant $S_{t+\tau} = S_t$ et Z_t une série temporelle stationnaire (cf paragraphe 3). Alors il est facile de vérifier que $\nabla_\tau x_t = x_t - x_{t-\tau} = z_t - z_{t-tau}$ est une série stationnaire (cf chapitre 3).

Retour sur l'exemple 2 (poissons).

1.3.4 Comparaison des différentes méthodes

Avantages des méthodes décrites au 2.a et 2.b sur la méthode du 2.c :
- Fournit un modèle facilement interprétable et qui permet, par exemple, de comparer plusieurs séries temporelles entre elles.
- Peut-être utilisé en prévision.

Les méthodes 2.a et 2.c s'ajustent à de nombreuses séries temporelles sans modification.

1.4 Modélisation de la composante stationnaire

1.4.1 Généralités sur les processus stationnaires

Définiton - On appelle processus à temps discret une suite $\{X_1, \cdots, X_T\}$ de variables aléatoires.

Remarque - Dans la suite, sauf mention contraire, on supposera que X_t est à valeurs réelles.

Définiton - Soit $X = \{X_t\}_{t\in\mathbb{Z}}$ un processus tel que $E[X_t^2] < \infty$ pour $t \in mathbbZ$ On dira que ce processus est (faiblement) stationnaire (ou stationnaire d'ordre 2) si les 2 conditions suivantes sont vérifiées pour tout t :
- $E(X_t) = m$ pour tout $t \in \{1, \cdots, T\}$
- $E(X_t X_{t+h}) = \gamma(h)$ est indépendant de t

On appelera alors *fonction d'autocovariance* de X_t la fonction R définie par

$$C(k) = E(X_t X_{t+k}) - E(X_t)E(X_{t+k})$$

et fonction d'autocorrélation la fonction R définie par

$$C(k) = \frac{E(X_t X_{t+k}) - E(X_t)E(X_{t+k})}{\sqrt{E(X_t^2)E(X_{t+k}^2)}}$$

.

Remarques :

1. On a $C(t) = C(-t)$ (fonction paire) donc il suffit de connaitre la fonction d'autocovariance sur $[0, T]$. De même pour R.
2. Pour tout k, $|C(k)| \leq C(0)$.
3. La fonction d'autocovaraince est définie positive.

Par la suite le terme stationnaire fera référence à la stationnarité faible.

Quelque exemples classiques de processus.

Bruit blanc
Définition - On dira que le processus ϵ_t est un "bruit blanc" s'il forme une suite de variables indépendantes et identiquement distribuées. Il sera dit centré si $E[\epsilon_t] = 0$ et réduit si $Var[\epsilon_t] = 1$.

Proposition - Soit X_t un bruit blanc vérifiant $E[X_t] = m$ et $Var[X_t] = \sigma^2$. Alors X_t est un processus stationnaire.

La première condition signifie tout simplement que l'espérance du processus est indépendante du temps. La seconde condition implique bien entendu l'indépendance de la fonction d'autocovariance par rapport au temps (stationnarité). Mais elle implique en outre que les termes d'autocovariance (pour $h \geq 0$) sont tous nuls. Seule la variance est non nulle. Autrement dit, cela signifie que les bruits blancs sont des processus stationnaires particuliers sans "mémoire". Le niveau de la série considéré aujourd'hui n'a aucune incidence sur son niveau de demain, tout comme le niveau d'hier n'a aucune incidence sur le niveau d'aujourd'hui.

Processus moyenne mobile
Définition : On dira que le processus $\{Z_t\}$ suit un modele "moyenne mobile d'ordre q" (MA(q)) s'il existe un bruit blanc centré $\{\epsilon_t\}$ et des constantes $\beta_k, k = 0, \cdots, q$ tels que

$$Z_t = \sum_{k=0}^{q} \beta_k \epsilon_{t-k}$$

Proposition : Soit $\{X_t\}$ un processus suivant un modèle moyenne mobile, avec $\{\epsilon_t\}$ un bruit blanc centré vérifiant $E[\epsilon_t] = \sigma^2 < \infty$, alors le processus $\{X_t\}$ est stationnaire. Sa fonction d'autocovariance est donnée par

$$C(h) = \sum_{k=0}^{q-|k|} \beta_k^2 \sigma^2 \text{ si } |h| \in 0, \cdots, q \text{ et } 0 \text{ sinon}$$

Preuve : Un tel processus est stationnaire : écrire la moyenne et la covariance.

Théorème : Décomposition de Wold) Tout processus stationnaire d'ordre deux $(X_t; t \in \mathbb{Z})$ peut être représenté sous la forme :

$$X_t = \sum_{k=1}^{\infty} \beta_k \epsilon_k + \kappa_t$$

où les paramètres β_k satisfont $\beta_0 = 1$, $\beta_k \in \mathbb{R}$ pour $k > 0$, $\sum_{k=1}^{\infty} \beta_k^2 < \infty$ et où ϵ_t est un bruit blanc.On dit que la somme des chocs passés correspond à la composante linéaire stochastique de X_t. Le terme κ_t désigne la composante linéaire déterministe telle que $Cov(\kappa_t, \epsilon_{t-k}) = 0$ pour tout $k \in \mathbb{Z}$.

Ainsi, d'après le théorème de Wold, si l'on omet la composante déterministe κ_t, tout processus stationnaire peut s'écrire comme une somme pondérée infinie de chocs passés, ces chocs étant

représentés par un bruit blanc de variance finie. L'implication forte de ce théorème est que, si l'on connaît les pondérations $\beta_k, k \in \mathbb{N}$, et si l'on connaît la variance du bruit blanc, on est mesure de proposer une représentation de n'importe quel processus stationnaire. Cette représentation est aussi qualifiée de représentation moyenne mobile infinie. Reste à comprendre ce que peut être cette composante linéaire déterministe κ_t. La condition $Cov(\kappa_t, \epsilon_{t-k}) = 0$ implique que ce terme est, par définition (déterministe), indépendant des chocs. Alors le cas le plus simple est celui d'un processus stationnaire $(X_t; t \in \mathbb{Z})$ d'espérance non nulle, tel que $E(X_t) = m$. Puisque le bruit blanc est par définition un processus centré, une somme pondérée de ces chocs est elle-même centrée. Par conséquent, la représentation de Wold du processus $(X_t; t \in \mathbb{Z})$ suppose que l'on ajoute à cette somme pondérée des chocs passés, une composante déterministe qui n'est autre que l'espérance du processus.

Définition - On dira que le processus X_t suit un modele autorégressif d'ordre p $(AR(p))$ s'il existe un bruit blanc centré réduit $\{\epsilon_t\}$, tel que ϵ_t soit indépendant de $X_0, \cdots, X_{t-1}$ et des constantes $\alpha_1, \cdots, \alpha_p$ et σ tels que pour $t \in \{1, \cdots, T\}$ on ait

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \alpha_2(X_{t-2} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + \sigma\epsilon_t$$

1.4.2 Modèles autorégressifs d'ordre 1

Quitte à considérer le processus $X_t - \mu$, on peut supposer que $E[X_t] = 0$. On suppose donc que X_t est centré.

Notons

$$X_t = \alpha_1 X_{t-1} + \epsilon_t$$

avec ϵ_t un bruit blanc centré de variance σ^2. On vérifie aisément que

$$X_t = \epsilon_t + \alpha_1\epsilon_{t-1} + \cdots + \alpha_1^k\epsilon_{t-k} + X_{t-k-1}$$

En particulier, on a

$$\left| X_t - \sum_{j=0}^{k} \alpha_1^j \epsilon_{t-j} \right|^2 = \alpha_1^{2k+2} |X_{t-k-1}| \text{ si } |alpha_1| < 1$$

d'où X_t admet une représentation en moyenne mobile

$$X_t = \sum k = 1^{\infty} \alpha_1^k \epsilon_{t-k}$$

On en déduit que $E[X_t] = 0$ pour tout t et

$$\begin{aligned} Cov(X_{t+k}X_t) &= \lim_{n\to\infty} E\left[\left(\sum_{j=0}^{n} \alpha_1^j \epsilon_{t+k-j} \right) \left(\sum_{j=0}^{n} \alpha_1^j \epsilon_{t-j} \right) \right] \\ &= \sigma^2 \alpha^{|k|} \sum_{j=0}^{\infty} \alpha_1^{2j} \\ &= \sigma^2 \frac{\alpha^{|k|}}{1 - \alpha_1^2} \end{aligned}$$

Rappel : somme d'une suite géomètrique $s_n = \sum_{k=0}{}^{N} - 1q^k = \frac{1-q^{n+1}}{1-q} \to \frac{1}{1-q}$ si $|q| < 1$. Si $|q| \geq 1$ il ne peut exister de solution stationnaire (processus explosif).

Théorème - Le modèle défini par l'équation $X_t = \alpha X_{t-1} + \sigma\epsilon$ pour possède une solution stationnare si et seulement si $|\alpha| < 1$. Dans ce cas, la solution stationnaire vérifie $E[X_t] = \alpha^t x_0$ et $E[X_t^2] = \frac{\sigma^2}{1-\alpha^2}$ et sa fonction d'autocovariance est donnée par $C(k) = \frac{\sigma^2}{1-\alpha^2}\alpha^{|k|}$ avec $h > 0$.

On remarque que $\alpha^k = e^{k\log(\alpha)}$. La décroissance de la fonction d'autocavariance d'un processus autorégressif d'ordre 1 est donc exponentielle.

Preuve du théorème
CNS d'existence d'une solution stationnaire : admis. Supposons que X_t soit stationnaire et vérifie $E[X_t] = 0$ et $E[X_t^2] = E[X_{t'}^2]$. . On a alors

$$E[X_t^2] = \alpha^2 E[X_t^2] + \sigma^2$$

donc

$$E[X_t^2] = \frac{\sigma^2}{1-\alpha^2}$$

. On a de plus

$$C(k) = E[X_t X_{t-k}] = \alpha E[X_{t-1}X_{t-k}] = \alpha^2 E[X_{t-2}X_{t-k}]$$

. On montre ainsi, par itération, que $C(k) = \frac{\sigma^2}{1-\alpha^2}\alpha^k$.

1.4.3 Modèles autorégressifs d'ordre p

Théorème - Le modèle défini par l'équation

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \sigma^2\epsilon_t \\ X_0 &= 0 \end{aligned}$$

pour $t \geq p$ possède une solution stationnaire si et seulement si les racines (complexes) du polynôme

$$\alpha_1 x + \cdots + \alpha_p x^p$$

sont de module strictement supérieur à 1 .

Remarque : modèle d'ordre $p = 1$: on retrouve la condition du théorème précédent.

.

Exercice - Donner les conditions de stationnarité d'un processus AR(2).

Proposition - Soit X_t un processus stationnaire suivant un modèle $AR(p)$ de paramètres $\alpha_1, \cdots, \alpha_p, \sigma^2$. On a alors $E[X_t^2] = \sum_{k=1}^{p} \alpha_k C(-k) + \sigma^2$ et la fonction d'autocovariance $C(k)$ vérifie les relations ci-dessous (équations de Yule Walker) :

$$\begin{bmatrix} C(1) \\ C(2) \\ C(3) \\ \vdots \end{bmatrix} = \begin{bmatrix} C(0) & C(-1) & C(-2) & \cdots & \\ C(1) & C(0) & C(-1) & \cdots & 1 \\ C(2) & C(1) & C(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \end{bmatrix}$$

Démonstration
L'équation définissant le processus AR est

$$X_t = \sum_{i=1}^{p} \varphi_i \, X_{t-i} + \varepsilon_t$$

En multipliant les deux membres par X_{t-j} et en prenant l'espérance, on obtient

$$E[X_t X_{t-j}] = E\left[\sum_{i=1}^{p} \varphi_i \, X_{t-i} X_{t-j}\right] + E[\varepsilon_t X_{t-j}]$$

Or, il se trouve que $E[X_t X_{t-j}] = C(j)$ par définition de la fonction d'autocovariance. Les termes du bruit blanc sont indépendants les uns des autres et, de plus, X_{t-j} est indépendant de ε_t où j est plus grand que zéro. Pour $j > 0$, $E[\varepsilon_t X_{t-j}] = 0$. Pour $j = 0$,

$$E[\varepsilon_t X_t] = E\left[\varepsilon_t \left(\sum_{i=1}^{p} \varphi_i \, X_{t-i} + \varepsilon_t\right)\right] = \sum_{i=1}^{p} \varphi_i \, E[\varepsilon_t \, X_{t-i}] + E[\varepsilon_t^2] = 0 + \sigma_\varepsilon^2$$

Maintenant, on a pour $j \geq 0$,

$$C(j) = E\left[\sum_{i=1}^{p} \varphi_i \, X_{t-i} X_{t-j}\right] + \sigma_\varepsilon^2 \delta_j$$

Par ailleurs,

$$E\left[\sum_{i=1}^{p} \varphi_i \, X_{t-i} X_{t-j}\right] = \sum_{i=1}^{p} \varphi_i \, E[X_t X_{t-j+i}] = \sum_{i=1}^{p} \varphi_i \, C(j-i)$$

qui donne les équations de Yule-Walker :

$$C(m) = \sum_{i=1}^{p} \varphi_i \gamma_{j-i} + \sigma_\epsilon^2 \delta_j$$

pour $j \geq 0$. Pour $j < 0$,

$$C(j) = C(-j) = \sum_{i=1}^{p} \varphi_i C(|j| - i) + \sigma^2 \delta_j$$

Remarque : Ces formules peuvent être utilisées pour calculer récursivement la fonction d'autocovariance.

Exercice : On considère le modele AR(2) défini par

$$X_t = 0.9 X_{t-1} - 0.3 X_{t-2} + 0.1 \epsilon_t$$

a. Vérifier que ce processus admet une solution stationnaire
b. Calculer les valeurs de $E[X_t]$, $C(1), \cdots, C(5)$ du modèle précédent.

1.4.4 Estimation de la moyenne et de la fonction d'autocovariance d'un processus stationnaire

Soit X_t un processus stationnaire de moyenne μ et de fonction d'autocovariance $C(h)$. Soit $\{x_1, \cdots, x_T\}$ une réalisation de ce processus.

On estime alors généralement μ par la moyenne empirique, c'est à dire la quantité $m = \frac{1}{n}\sum_{t=1}^{T} x_t$ et par la fonction d'autocovariance empirique la fonction $\hat{C}$ définie par

$$\hat{C}(h) = \frac{1}{n}\sum_{t=1}^{T-h}(x_t - m)(x_{t+h} - m)$$

pour $k \in 0, \cdots, T-k$.

Exercice
On considère la série temporelle suivante
0.4 1.2 0.8 1.0 0.6 2.2 -1.6 2.6 1.0 1.7
Calculer la moyenne et les quantités $\hat{C}(0)$,$C(5)$,$\hat{C}(10)$.

Exercice
Soit $\{X_t\}$ un processus stationnaire de moyenne μ et de fonction d'autocovariance $C(h)$. On pose $M_t = \frac{1}{T}\sum_{t=1}^{T} X_t$. a. Calculer $E[M_t]$ b. Exprimer $var(M_t)$ en fonction de C On suppose maintenant que le processus $\{X_t\}$ suit un modèle AR(1), c'est à dire vérifie $X_t - \mu = \alpha(X_{t-1} - \mu) + \sigma\epsilon_t$ avec ϵ un bruit blanc centré et réduit. c. Exprimer $var(M_t)$ en fonction de α et σ^2 d. En déduire un procédé permettant de construire un intervalle de confiance pour μ

1.4.5 Inférence statistique pour les modèles autorégressifs

Estimation des paramètres
On dispose d'une observation $\{x_0, \cdots, x_T\}$ de longueur $T+1$ d'un processus stationnaire X_t supposé suivre un modèle AR(p), c'est à dire vérifiant

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \sigma\epsilon_t$$

avec $t \in \mathbb{Z}$ et $X_0 = x_0$ des paramètres inconnus. On cherche alors à estimer ces paramètres à l'aide des observations disponibles.

Pour cela différentes méthodes sont possibles
- Moindres carrés (non étudié)
- Maximum de vraisemblance (non étudié). L'estimation d'un modèle AR(P) par la méthode du maximum de vraisemblance est délicate car la fonction de vraisemblance est très complexe et n'a pas de dérivée analytique. Cette difficutlé provient de l'interdépendance des valeurs, ainsi que du fait que les observations antérieures ne sont pasd toutes disponibles pour les p premières valeurs.
- Méthode des moments...utilisation des équations de Yule-Walker

Estimateurs de Yule-Walker
La méthode consiste à reprendre les équations de Yule-Walker en inversant les relations : on exprime les coefficients en fonction des autocovariances. On applique alors le raisonnement de la Méthode des moments : on trouve les paramètres estimés d'après les autocovariances estimées.

On a vu précedemment que $\mu = E[X_1]$, $C(0) = \alpha_1 C(1) + \cdots + \alpha_p C(p) + \sigma^2$ et

$$C(k) = \alpha_1 C(k-1) + \cdots + \alpha_p C(k-p)$$

En prenant l'équation sous sa forme matricielle

$$\begin{bmatrix} C(0) \\ C(1) \\ C(2) \\ C(3) \\ \vdots \end{bmatrix} = \begin{bmatrix} C(-1) & C(-2) & C(-3) & \dots & 1 \\ C(0) & C(-1) & C(-2) & \dots & 0 \\ C(1) & C(0) & C(-1) & \dots & 0 \\ C(2) & C(1) & C(0) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \sigma_\varepsilon^2 \end{bmatrix}$$

que l'on écrira aussi $R_p \alpha = C$ avec $R_p = (C(i-j)$ et $C = (C(i))$. Le vecteur des paramètres $\hat{\theta} = \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\sigma}_\varepsilon^2 \end{pmatrix}$ peut alors être obtenu.

Remarque : les estimateurs de Yule-Walker sont tels que la moyenne et les $p+1$ premières valeurs de la fonction d'autocovariance du modèle estimé coincident avec celles estimées sur les données. Cas particuliers : Modèles d'ordre $p = 1$. On obtient $\hat{R}_1 = \hat{C}(0)$ alors $\alpha_1 = \hat{C}(1)/\hat{C}(0)$ et $\hat{\sigma}^2 = \hat{C}(0) - \hat{C}(1)^2/\hat{C}(0)$ Modèles d'ordre $p = 2$. On obtient alors , donc , ,

Les estimateurs de Yule-Walker sont évidemment consistants dès que les $\hat{C}(k)$ sont des estimateurs consistants des covariances $C(k)$. Le résultat suivant donne plus de précisions sur le comportement asymptotique de ces estimateurs. Il est utile même dans le cas où on aurait résolu le système des équations de Yule-Walker avec une valeur de p qui ne serait pas la bonne.

Proposition - Soit (X_t) un processus $AR(p)$ stationnaire, soit $q \geq p$ et notons $\alpha \in \mathbb{R}^q$ le vecteur $\alpha = (\alpha_1, \dots, \alpha_p, 0, \dots 0)$. Notons $\hat{\alpha}^{(n)}$ la solution de l'équation de Yule-Walker empirique dans $\mathbb{R}^q$ soit $R_m^{(n)} \hat{\alpha}^{(n)} = C_{X,m}^{(n)}$. Alors

$$\sqrt{n}(\hat{\alpha}^{(n)} - \alpha) \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} N(0, \sigma^2 R_q^{-1}) \quad (6.7)$$

Remarque : La matrice de covariance dépend de σ^2 de façon linéaire (est égale à σ^2 multiplié par la matrice de covariance pour un bruit de variance 1). Donc la matrice de covariance asymptotique $\sigma^2 R_m^{-1}$ ne dépend pas de σ^2.

La Proposition précise la variance asymptotique des estimateurs de Yule-Walker. Ces estimateurs ont la même variance asymptotique que ceux du maximum de vraisemblance. Mais, évidemment, ils ne sont utiles que lorsque on sait déjà que l'on a affaire avec un processus AR (et donc que la partie MA est nulle).

Dans le cadre d'un processus AR, la proposition fournit une approche à l'étude de la meilleure valeur de p. En effet, si $m > p$ et $\hat{\alpha}^{(n)} = (\hat{\alpha}_1^{(n)}, \dots, \hat{\alpha}_m^{(n)})$, (6.7) affirme que $\sqrt{n}\, \hat{\alpha}_m^{(n)} \to N(0, \lambda)$ pour $n \to \infty$, où λ est égal a σ^2 multiplié par l'élément diagonal d'indice m, m de la matrice R_m^{-1}. Or il est remarquable que cette quantité vaut toujours 1.

Remarque : ces formules peuvent être utilisées pour construire des intervalles de confiance asymptotiques en remplaçant σ et R_p par leurs estimations. Plus précisemment, pour T grand, on a $\sqrt{T}(\hat{\alpha} - \alpha) \sim \mathcal{N}(0, \Sigma)$, donc en particulier $\sqrt{T}(\hat{\alpha}_i - \alpha_i) \sim \mathcal{N}(0, \Sigma_{ii})$ avec Σ_{ii} le ième terme

diagonal de la matrice $\Sigma = \sigma^2 R_p^{-1}$, puis

$$P(-q_{1-\alpha}\sqrt{\Sigma_{ii}} \leq \sqrt{T}(\hat{\alpha}_i - \alpha_i) \leq q_{1-\alpha}\sqrt{\Sigma_{ii}}$$

avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Gauss centrée et réduite (+ dessin).

Exercice
On considère une série temporelle $\{x_t\}$ de longueur $T = 200$ telle que $\sum x_t = 13.1$,$\sum x_t^2 = 164,2$,$\sum x_t x_{t+1} = 6.1$, $\sum x_t x_{t+2} = -9.98$,$\sum x_t x_{t+3} = 13.4$,

$$\sum x_t x_{t+4} = 11.8$$

.
1. On suppose que cette série temporelle suit un modèle d'ordre 1
a. Ecrire les éqiuations de Yule-Walker
b. En déduire une estimation des paramètres du modèle
c. Donner un intervalle de confiance pour le paramètre $alpha_1$
2. On suppose que cette série temporelle suit un modèle d'ordre 2.
a. Ecrire les éqiuations de Yule-Walker
b. En déduire une estimation des paramètres du modèle
c. Donner un intervalle de confiance pour le paramètre $alpha_1$ puis pour α_2.

1.4.6 Prédiction dans les modèles autorégressifs

Dans ce paragraphe, on suppose que $\{X_t\}$ est un processus stationnaire qui suit un modèle AR(p), c'est à dire vérifie

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + \sigma\epsilon_t$$

avec ϵ_t un bruit blanc centré réduit.

Objectif : on cherche à prédire la valeur prise par le processus aux instants $t+1, t+2, \cdots$ à partir de la connaissance des valeurs prises par ce processus jusqu'à l'instant t, c'est à dire de $x_0, \cdots, x_t$.

Le modèle se réécrit sous la forme

$$X_t = \alpha_1(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + \mu + \sigma\epsilon_t$$

En général, on utilise les quantités suivantes :

$$\hat{X}_{t+1|t} = \alpha_1(X_t - \mu) + \cdots + \alpha_p(X_{t-p+} - \mu) + \mu$$

pour prédire X_{t+1} à partir de $X_1, \cdots, X_t$.

$$hatX_{t+2|t} = \alpha_1(\hat{X}_{t+1|t} - \mu) + \cdots + \alpha_p(X_{t-p+2} - \mu) + \mu$$

Et de façon générale

$$hatX_{t+k|t} = \alpha_1(\hat{X}_{t+k-1|t} - \mu) + \cdots + \alpha_p(X_{t+k-p|t} - \mu) + \mu$$

Remarque - Dans le cas des modèles d'ordre 1, on a $\hat{X}_{t+1|t} - \mu = \alpha_1(X_t - \mu)$, $\hat{X}_{t+2|t} - \mu = \alpha_1(\hat{X}_{t+1|} - \mu) = \alpha_1^2(X_t - \mu), \cdots$. On vérifie aisément par récurrence que $\hat{X}_{t+t|t} - \mu = \alpha_1^k(X_t - mu)$, donc en particuler que $\hat{X}_{t+k|t} \to \mu$ quand k tend vers l'infini.

Proposition (admise) - Si $\{X_t\}$ est un processus stationnaire qui suit un modèle AR(p), alors la meilleure prédiction, au sens des moindres carrés, de X_{t+1} connaissant $x_t, ..., x_{t-p+1}$ est donnée par $\hat{x}_t$ défini par

$$\hat{x}_{t+1|t} = \alpha_0 + \alpha_1 x_t + \cdots + \alpha_p x_{t-p+1} + \mu$$

Qualité de la prédiction

Prédiction à un pas de temps On a par définition $\hat{X}_{t+1|t} - X_{t+1} = \sigma\epsilon_{t+1}$. ϵ_{t+1} représente donc l'erreur de prédiction à un pas de temps. En particulier, l'estimation est non-biaisée et la variance de l'erreur d'estimation est σ^2. En général, afin de construire des intervalles de prédiction, on est ammené à supposer que ϵ suit une loi de Gauss centrée et réduite. On en déduit alors que .... L'intervalle [] est appelé intervalle de prédiction à 95%. En pratique, les quantités et sont inconnus. On les estime.

Prédiction à deux pas de temps On a par définition

$$\hat{X}_{t+2|t} - X_{t+2} = \sigma\epsilon_{t+2} + \alpha_1\sigma\epsilon_{t+1}$$

En particulier, l'estimation est non-biaisé et la variance de l'erreur d'estimation est $(1 + \alpha_1^2)\sigma^2$ On peut aussi construire des IP.

Remarque - La généralisation au cas général est complexe, sauf dans le cas des modèles d'ordre $p = 1$. On a alors (donner formule avec k=1, k=2, puis k=3).

$$\hat{X}_{t+k|t} - X_{t+k} = \sigma\epsilon_{t+k} + \alpha_1\sigma\epsilon_{t+k-1} + \cdots + \alpha_1^{k-1}\sigma\epsilon_{t+1}$$

En particulier, on a l'estimation est non-biaisée et la variance de l'erreur d'estimation est $\sigma^2 \sum_{j=0}^{k-1} \alpha^{2j}$.. On peut aussi construire des IP dans le cas gaussien puisque...On vérifie aisément que la qualité de la prédiction se dégrade lorsque k augmente.

1.4.7 Sélection de modèle

Jusqu'à présent, nous avons supposé que l'ordre du modèle AR est connu. Cependant, en pratique, cette quantité est généralement inconnue, et on cherche alors à estimer sa valeur à partir des observations, ce qu'on appele généralement le probleme de la selection de modèle.

Une première méthode consiste à réaliser des tests en utilisant le résultat ci-dessous.

Proposition - supposons que $\{X_t\}$ suit un modèle AR(p), et qu'on ajuste un modèle d'ordre $q > p$ à ce processus. Notons l'estimateur correpondant de sous . On a alors pour "grand". On en déduit le test suivant : On veut tester : suit un modèle AR(p) contre H1 : suit un modèle AR(p+1) avec un risque de première espèce . On ajuste un modèle d'ordre . Si , on accepte , sinon, on refuse .

Pour sélectionner la meilleur valeur de , on peut alors utiliser des tests successifs. On ajuste alors des modèle d'ordre , puis p=2, ...jusqu'à ce que le dernier coefficient du modèle ajusté soit inférieur à .

Une autre idée consiste à utiliser le modèle qui minimise l'erreur de prédiction à un pas de temps, c'est à dire la valeur de . Cependant, en pratique, plus on agmente la valeur de , plus l'erreur de prédiction diminue. On peut alors utiliser le critère BIC. On choisit alors le modèle qui minimise la quantité

On obtient alors généralement un modèle pour lequel on a un bon compromis entre le nombre de paramètre et l'erreur de prediction. En pratique, on obtient généralement des modeles parsimonieux qui s'ajuste bien aux données.

Validation de modèle

Lorsque l'on veut vérifier l'adéquation d'un modèle autorégressif à des données observées, on utilise généralement le test de Portmenteau. Supposons tout d'abord que l'on veuille tester H0 : est un bruit blanc : n'est pas un bruit blanc Pour cela, on dispose d'une réalisation de . Notons la fonction d'autocovariance de ce processus et sa version paramétrique. Si est vraie, alors si . On s'attend donc à ce que si . On peut montrer que, si est suffisamment grand par rapport à h, alors . On utilise alors la statistique de test avec h "petit" par rapport à T (par exemple, h=T/10). Sous , suit une loi du chi2 à h ddl. Pour un risque de première espéce , on accepte donc si . Sinon, on refuse .

Supposons maintenant que l'on veuille tester : suit le modèle AR(p) avec b.b : ne suit pas ce modèle On dispose pour cela d'une réalisation . On calcule les résidus conditionnels , et on test si cette séquence est un b.b avec le test de potmenteau.

Exemple : Sur une série observée, de longueur , on a , , , , et . a. Trouver le "meilleur" modèle AR qui a un ordre r compris entre 0 et 4. (en vrai AR(2) avec et) reponse : cf ordi

Chapitre 2

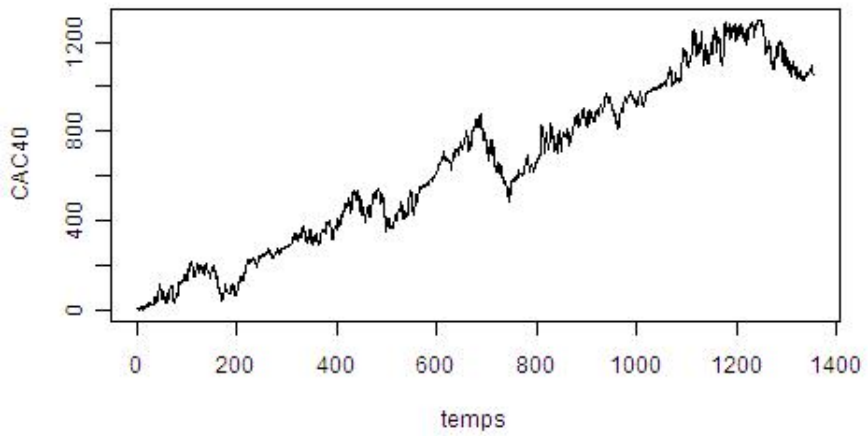
Modèles à changement de régimes markovien

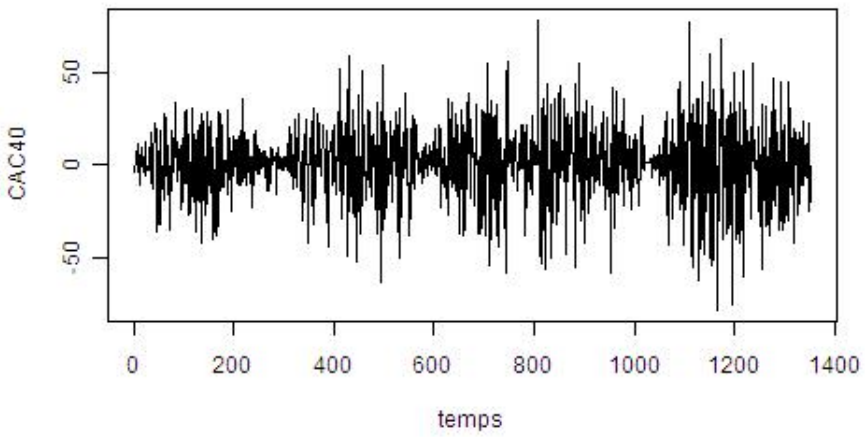
2.1 Introduction

2.1.1 Généralités, exemples

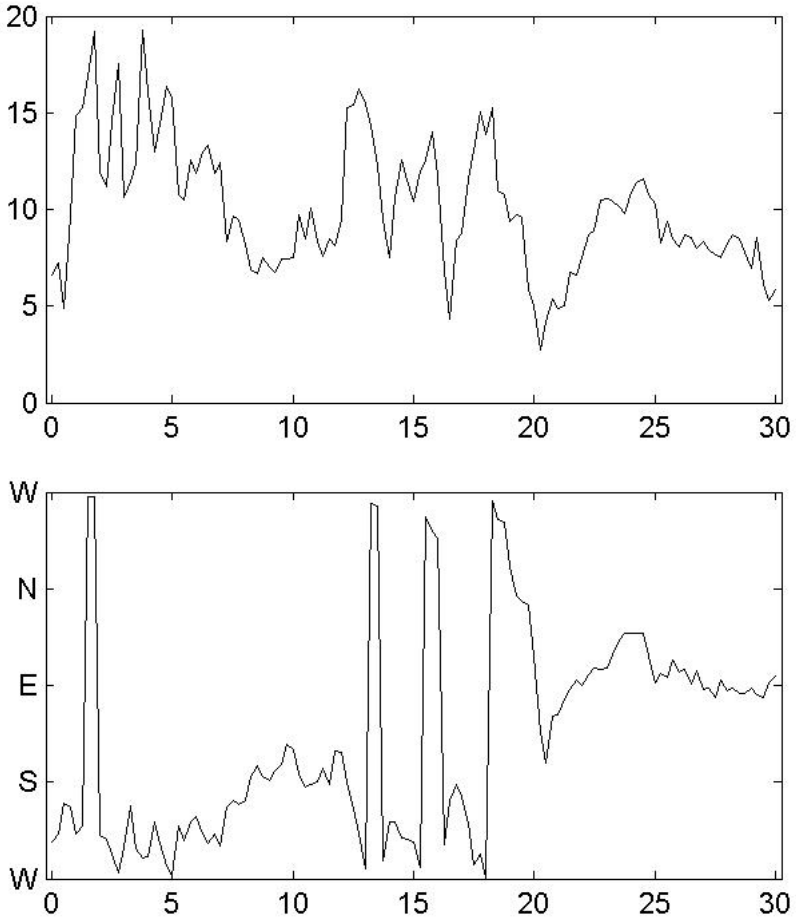
Les modèles AR, ARMA ou ARIMA permettent de modéliser des processus stationnaires ie dont les caractéristiques (moyenne, variance, etc) ne varient pas avec le temps. Dans certaines situations, ces modèles ne sont pas suffisants (voir les exemaples ci-dessous). On doit alors utiliser d'autres approches. Les modèles à changement de régime markovien sont classiquement utilisés pour modéliser un processus aléatoire ayant des caractéristiques (ou paramètres) qui changent au cours du temps. Nous verrons dans la suite que ces modèles sont très flexibles et permettent de modéliser des séries temporelles complexes.

Exemple 1 - On montre par exemple ci-dessous la série temporelle de l'évolution journalière du CAC40 entre le 01/11/1995 et le 18/02/2000. La première figure représente l'indice brut et dans la seconde nous avons retiré une tendance linéaire $y_t = x_t - x_{t-1}$. On observe des variations de la volatilité : à certaines périodes le CAC40 varie faiblement autour de sa moyenne tandis qu'à d'autres périodes les variations sont très importantes.

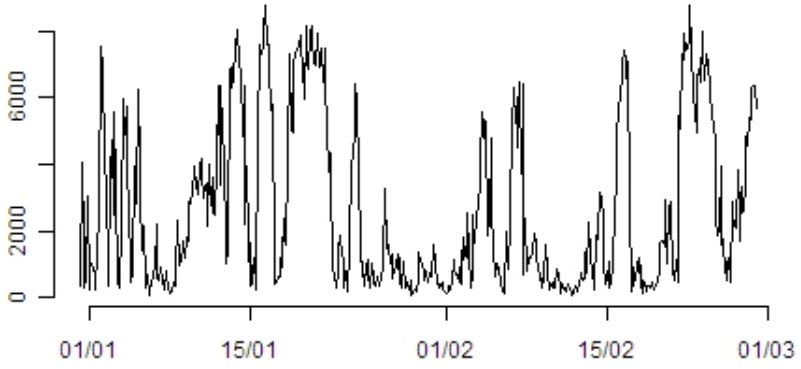


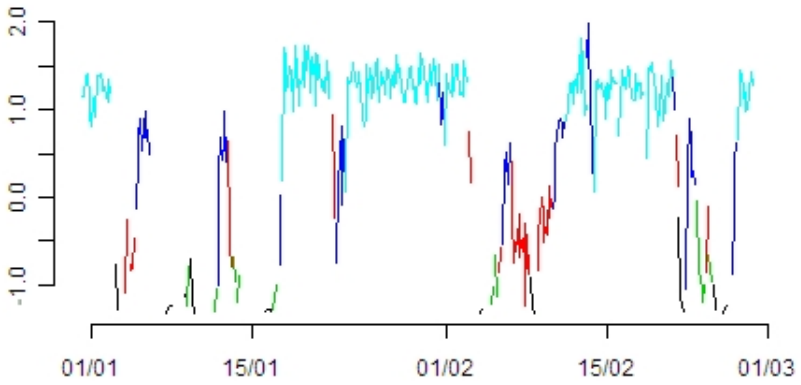


Exemple 2 - Intensité et direction du vent. On observe que l'intensité et la direction du vent varient en moyenne et en variance en fonction du régime météorologique (dépression, anticyclone, ...). Le régime météorologique n'est pas observé directement.

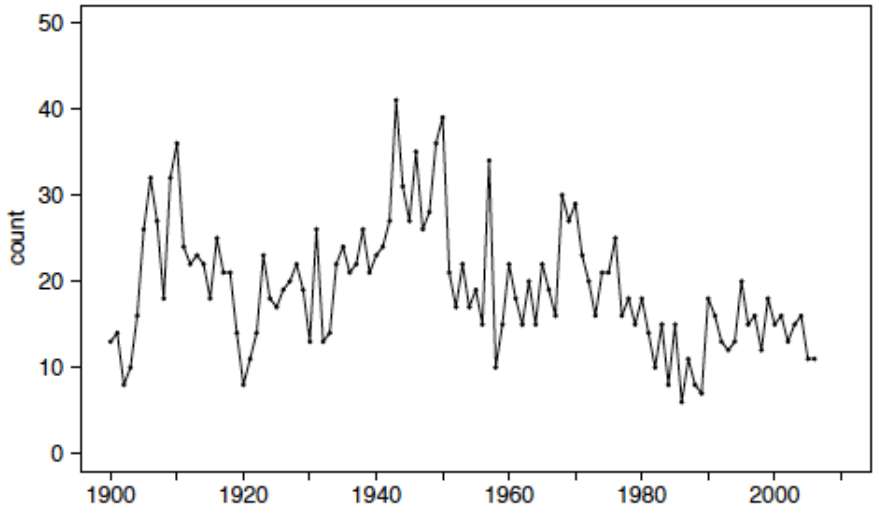


Exemple 3 - Production d'électricité éolienne. On observe des périodes de production forte ou de production faible, etc.





Exemple 4 - Nombre de tremblements de terre de magnitude élevée (supérieure ou égale à 7) dans le monde de 1900 à 2006. Cet exemple permetra d'illustrer les modèles HMM Poisson. En effet la loi de Poisson est la loi la plus usuelle pour modéliser des variables aléatoires de comptage (définies sur $\mathbb{N}$).



Dans ces quatre exemples, on identifie des suites de séries temporelles. Dans chaque séquence, la série temporelle à des caractéristiques propres (moyenne, variance) et ces caractéristiques changent d'une série à l'autre.

Les domaines dans lesquels on utilise les modèles à chaîne de Markov cachée : économie et finance, bioinformatique, biologie, environnement (météo), reconnaissance de l'écriture ou de la parole, image, etc.

Les deux références les plus citées

1 Rabiner (1989) , A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, PROCEEDINGS OF THE IEEE. 77 (2), p. 257 - 286.
2 Ephraim and Merhav (2002), Hidden Markov Processes, IEEE TRANSACTIONS ON INFORMATION THEORY, 48 (6) p. 1518 - 1569.

Un livre : W. Zucchini, L. MacDonald, Hidden Markov Model for Time series, An introduction using R. Chapman and Hall. (2009)

2.1.2 Modèle

Définition - Une chaîne de Markov cachée est un processus à temps discret $(X_t)_{t\in 1,\cdots,N}$ avec deux composantes $X_t = (S_t, Y_t)$. On supposera dans ce chapitre que S_t est à valeurs dans un ensemble fini $E = 1, ..., M$ (si $E \subset \mathbb{R}^d$, on parle de "modèles à espace d'état"). Y_t est à valeurs dans un espace F qui pourra être discret ou continu selon l'application.

Ce processus vérifie de plus les propriétés d'indépendance conditionnelles ci-dessous :

- $P(S_t|S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(S_t|S_{t-1} = s_{t-1})$
- $P(Y_t|S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(Y_t|S_t = s_t)$

En pratique, généralement seul le processus $\{Y_t\}$ sera observé. S_t est une variable "cachée" ou "latente" ou "manquante".

On résume généralement ces propriétés par le graphe d'indépendance conditionnelle ci-dessous :

$$\begin{array}{lccccccccc} \textit{State} & \cdots & \rightarrow & S_{t-1} & \rightarrow & S_t & \rightarrow & S_{t+1} & \rightarrow & \cdots \\ & & & \downarrow & & \downarrow & & \downarrow & & \\ \textit{Observation} & \cdots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \cdots \end{array}$$

Interprétation - S_t est une chaîne de Markov et conditionnellement à cette chaîne de Markov les observations successives $\{Y_t\}$ sont supposés indépendantes. Toute la dynamique est dans le processus caché !

Remarque - Les processus (X_t) et (S_t) sont Markoviens. Par contre, le processus (Y_t) n'est en général pas Markovien !

Un processus de Markov caché est donc paramétré par

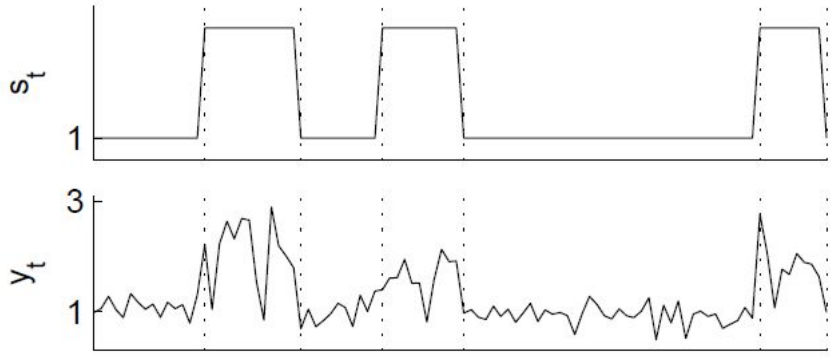
- la loi initiale $\pi = (\pi(1), ..., \pi(M))$ et la matrice de transition $Q = (Q(i,j)_{i,j\in\{1...M\}}$ de la chaîne de Markov cachée qui vérifient les contraintes usuelles.
- les probabilités d'émission $P(Y_t|S_t = s)$ pour $s \in \{1, \cdots, M\}$. On supposera que ces probabilités conditionnelles ont une densité par rapport à une certaine mesure dominante commune (mesure de comptage dans le cas où F est discret, mesure de Lebesgue dans le cas où F est continu) et on notera $g(y_t; \theta^{(s)})$ cette densité avec $\theta^{(s)} \in \Theta^{(s)}$.

Exemple - Lorsque le processus observé est à valeurs dans $F = \mathbb{R}$, on utilise généralement la loi normale pour paramétrer les probabilités d'émission. On a alors $\theta^{(s)} = (m^{(s)}, \sigma^{(s)}) \in \mathbb{R} \times \mathbb{R}^{+*}$ et l'équation d'observation peut être simulée simplement à partir de la simulation d'un bruit blanc gaussien $\{E_t\}_t \sim_{\text{iid}} \mathcal{N}(0,1)$ par l'expression :

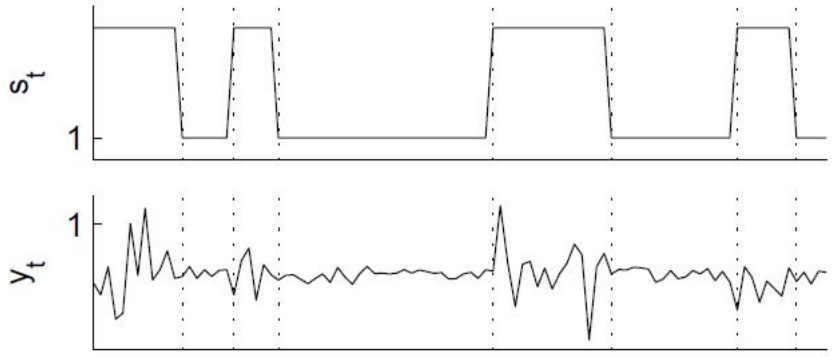
$$Y_t = m^{(S_t)} + \sigma^{(S_t)} E_t$$

Exemples de trajectoires dans le cas M = 2.

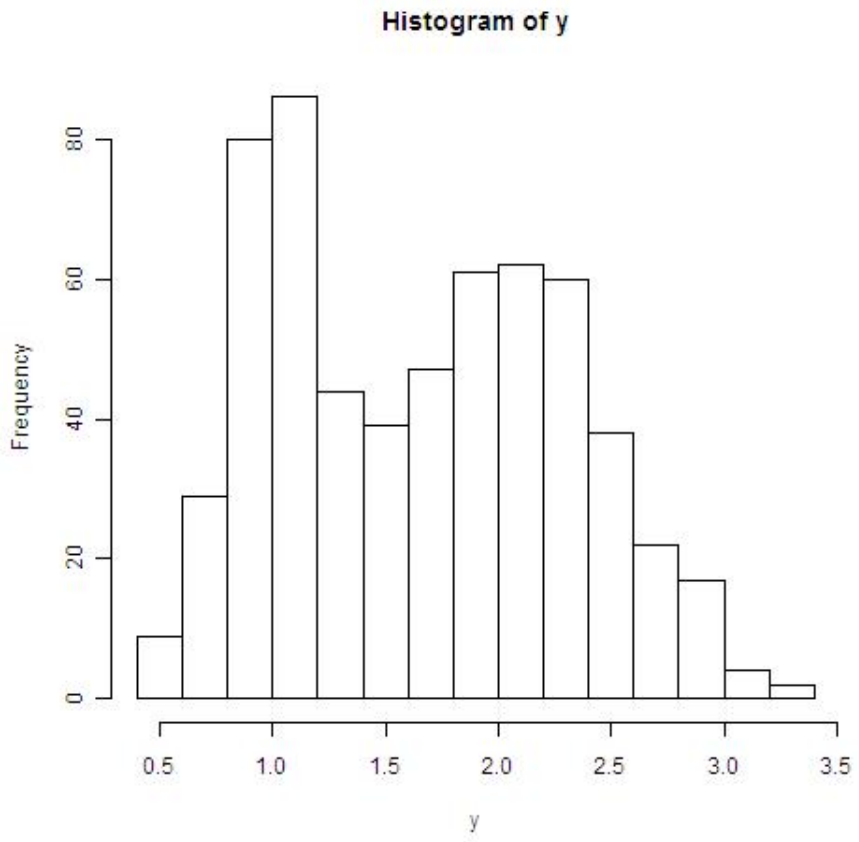
1. $m^{(1)} = 1$, $m^{(2)} = 2$, $\sigma^{(1)} = 0.2$, $\sigma^{(2)} = 0.5$, $Q = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$



2. $m^{(1)} = 0$, $m^{(2)} = 0$, $\sigma^{(1)} = 0.1$, $\sigma^{(2)} = 0.5$, $Q = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$



3. Considérons aussi le cas dégénéré où $Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. On choisit $\pi_1 = 1/3$ et $\pi_2 = 2/3$ ainsi que $m^{(1)} = 0$, $m^{(2)} = 0$, $\sigma^{(1)} = 0.1$, $\sigma^{(2)} = 0.5$. On parle alors de "modèle de mélange" et non plus de modèles à chaîne de Markov cachée. En absence de dynamique, la visualisation de la série temporelle n'est pas intéressante. En revanche, on peut tracer l'histogramme des observations y.



Dans la suite nous allons nous attarder un peu sur les modèles de mélange puis sur les chaines de Markov.

2.2 Modèles de mélange

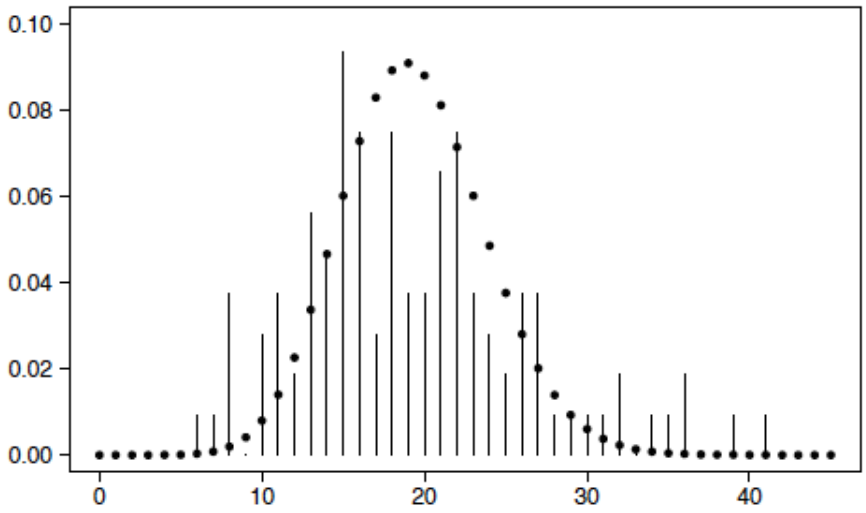
2.2.1 Introduction

Considérons de nouveau l'exemple 4 du nombre de tremblements de terre dans le monde (voir tableau ci-dessous). Un modèle classique pour les variables aléatoire entières est la distribution de Poisson dont la loi est définie par

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

et qui est telle que la variance est égale à la moyenne. Dans la série du nombre de tremblements de terre de magnitude supérieure ou égale à 7, la variance empirique, $s^2 \simeq 52$, est bien plus grande que la moyenne empirique $\bar{x} \simeq 19$. La dispersion des valeurs observées est donc plus importante que la dispersion attendue dans une loi de Poisson. Ce modèle n'est donc pas pertinent pour ces données. Ceci est confirmé par la figure ci-dessous qui représente un diagramme en batons du nombre de tremblements de terre par an.

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													



Les modèles de mélange permettent de de modéliser des données très dispersées, en particulier quand les observations présentent une bimodalité (ou plus généralement de la multimodalité). En effet, les modèles de mélange sont construits de façon à prendre en compte une hétérogénéité non observée. Implicitement, on considère que la population est constituée de plusieurs groupes non observés, chacun admettant une distribution distincte (voir aussi le cours de classification non supervisée).

Considérons par exemple le nombre, X, de paquets de cigarettes acheté par les clients d'un bureau de tabac-presse. Les clients peuvent être divisés en 3 groupes : les non fumeurs, les

fumeurs occasionnels, les fumeurs réguliers. Dans chacun des groupes, le nombre de paquets de cigarettes peut être distribué suivant une loi de Poisson ; X sera plus dispersée qu'une variable de Poisson.

Supposons maintenant que chaque nombre de tremblements de terre est généré par une des 2 lois de Poisson de moyenne λ_1 et λ_2 et que le choix de la loi est déterminé par un mécanisme aléatoire : λ_1 est choisi avec la probabilité π_1 et λ_2 avec la probabilité $\pi_2 = 1 - \pi_1$. Ainsi, en notant X le nombre de tremblements de terre et S la variable aléatoire discrète qui vaut 1 avec la probabilité π_1 et 2 avec la probabilité π_2,

$$P(X = x) = \pi_1 P(X = x|S = 1) + (1 - \pi_1)P(X = x|S = 2)$$

2.2.2 Modèle

Un modèle de mélange caractérise la distribution de la variable X d'un couple (S, X) tel que

- S est une variable aléatoire discrète définie sur $\{1, \cdots, M\}$; S n'est pas observée (on dit aussi que S est *cachée* ou *latente*).
- X est une variable aléatoire à valeurs sur $\mathbb{R}^d, d \leq 1$ (ou tout autre sous ensemble continu ou discret) telle que la loi conditionnelle $P(X|S = m)$ admet une densité (ou une fonction de probabilité) $g_m(.)$ pour tout $m \in \{1, \cdots, M\}$.

D'après le théorème de Bayes, pour tout ensemble A, la loi marginale de la variable X vérifie

$$P(X \in A) = \sum_{m=1}^{M} P(X \in A|S = m)P(S = m)$$

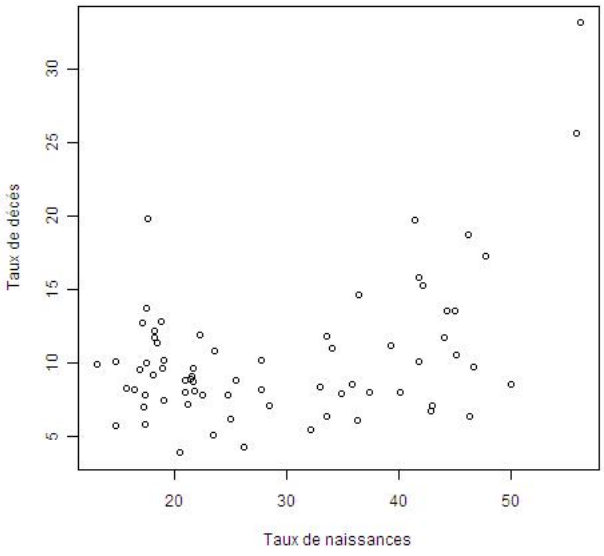
On écrit alors la densité de X comme une combinaison convexe des densités $g_m(.), m \in \{1, \cdots, M\}$

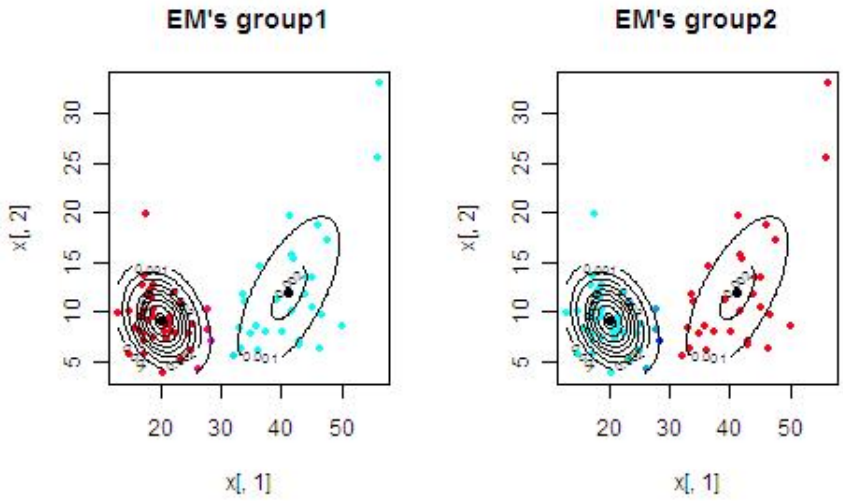
$$g(x) = \sum_{m=1}^{M} \pi_m g_m(x)$$

où $\pi_m = P(S = m)$ avec

$$\pi_1 + \cdots + \pi_M = 1$$

Exemple - Taux de naissances et de décès pour 70 pays du monde.





On note que si on définit Y_m la variable aléatoire de loi g_m, l'espérance de la loi de X est donnée par

$$E(X) = \sum_{m=1}^{M} P(S = m)E(X|S = m) = \sum_{m=1}^{M} \pi_m E(Y_m).$$

Plus généralement, le moment d'ordre k s'écrit

$$E(X^k) = \sum_{m=1}^{M} P(S = m)E(X^k|S = m) = \sum_{m=1}^{M} \pi_m E(Y_m^k).$$

Ce résultat n'est bien sûr pas vrai pour les moments centrés. En particulier, la variance n'est pas une combinaison linéaire des différentes lois. Dans le cas où $M = 2$, elle est donnée par l'expression suivante

$$Var(X) = \pi_1 Var(Y_1) + \pi_2 Var(Y_2) + \pi_1\pi_2(E(Y_1) + E(Y_2))^2 \tag{2.1}$$

Exercice - montrer l'équation (2.1).

2.2.3 Inférence

Identifiabilité

Maximum de vraisemblance

Rappelons la définition du problème d'estimation par maximum de vraisemblance. On considère qu'on a une variable Z dont la distribution admet une densité $p(z;\theta)$ avec $\theta \in \Theta$ le vecteur des paramètres. Et on suppose qu'on dispose d'une suite de n réalisations indépendantes de Z : $\mathcal{Z} = \{z_1, \cdots, z_n\}$. Alors la vraisemblance d'un paramètre θ sachant les données s'écrit :

$$\log \mathcal{L}(\theta; \mathcal{Z}) = \sum_{i=1}^{n} p(z_i; \theta)$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log \mathcal{L}(\theta; \mathcal{Z})$$

En fonction de la forme de $p(z;\theta)$ ce problème peut-être simple ou difficile à résoudre. En particulier, quand une partie des données n'est pas observée ou est manquante, ce problème n'est pas trivial et on doit avoir recours à des méthodes spécifiques.

Considérons par exemple le mélange de deux lois de Poisson de moyennes λ_1 et λ_2 avec pour paramètres de mélange π_1 et π_2, $\pi_1 + \pi_2 = 1$. La fonction de probabilité de mélange est

$$p(x) = \pi_1 \frac{\lambda_1^x e^{-\lambda_1}}{x!} + \pi_2 \frac{\lambda_2^x e^{-\lambda_2}}{x!}$$

Comme $\pi_2 = 1 - \pi_1$, il y a 3 paramètres à estimer : λ_1, λ_2 et π_1. La vraisemblance s'écrit

$$\mathcal{L}(\lambda_1, \lambda_2, \pi_1 | x_1, \cdots, x_n) = \prod_{i=1}^{n} \left(\pi_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_!} + (1 - \pi_1) \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!} \right)$$

On ne sait pas résoudre analytiquement le problème de la maximisation de $\mathcal{L}(\lambda_1, \lambda_2, \pi_1 | x_1, \cdots, x_n)$. On utilise alors des méthodes numériques ou l'algorithme EM (voir par exemple le package **R** `flexmix`).
Si on utilise un algorithme d'optimisation sans contrainte (voir par exemple la fonction `nlm`) on doit reparamétrer le modèle afin de en compte les contraintes : $\lambda_m > 0, \pi_m \in [0,1], \pi_1 + \cdots + \pi_M = 1$. On introduit alors les paramètres suivant, définis sur $\mathbb{R}$ tout entier :

$$\eta_m = \log \lambda_m \quad (m = 1, \cdots, M)$$

$$\tau_m = \log \left(\frac{\pi_m}{1 - \sum_{j=2}^{M} \pi_j} \right) \quad (m = 2, \cdots, M)$$

On retrouve les paramètres originaux par les transformations inverses

$$\lambda_m = e_m^{\eta} \quad (m = 1, \cdots, M)$$

$$\pi_m = \frac{e^{\tau_m}}{1 + \sum_{j=2}^{M} e^{\tau_m}} \quad (m = 2, \cdots, M)$$

et $\pi_1 = 1 - \sum_{j=2}^{M} \pi_j$.

unbounded likelihood...

2.2.4 Algorithme EM

L'algorithme EM (Estimation-Maximisation) a été proposé par Arthur Dempster, Nan Laird et Donald Rubin en 1977. C'est une méthode qui permet d'approcher l'estimateur du maximum de vraisemblance quand les données sont incomplètes (cas d'une variable cachée par exemple) ou quand une partie des données est manquante (cas d'une censure par exemple).

Notons $\mathcal{X}$ les données incomplètes. Si une variable est cachée, on suppose qu'il existe les données complètes (ou complétées) $\mathcal{Z} = (\mathcal{X}, \mathcal{S})$ et on considère la densité jointe

$$p(z;\theta) = p(x,s;\theta) = p(x|s;\theta)p(s;\theta)$$

On peut ainsi définir la vraisemblance des données complètes :

$$\mathcal{L}(\theta;\mathcal{Z}) = \mathcal{L}(\theta;\mathcal{X},\mathcal{S}) = p(\mathcal{X},\mathcal{S};\theta)$$

Mais il faut garder en tête que cette fonction est aléatoire puisqu'on n'observe pas S. On ne peut donc pas la calculer. En revanche on peut, dans certains cas, calculer son espérance si on fixe la valeur du paramètre θ.

L'algorithme EM est un algorithme itératif qui consiste à chaque étape k en deux sous étapes :

Etape E [1] : Calculer l'espérance de la log vraisemblance des données complètes pour une valeur fixée $\theta^{(k)}$ du paramètre

$$Q(\theta,\theta^{(k)}) = E_{\mathcal{S}}\left[\log p(\mathcal{X},\mathcal{S};\theta)|\mathcal{X},\theta^{(k)}\right] \tag{2.2}$$

Etape M [2] : Maximiser la fonction $Q : \theta \mapsto (\theta,\theta^{(k)})$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \mathcal{Q}(\theta,\theta^{(k)}) \tag{2.3}$$

Ces deux étapes sont répétées autant que nécessaire. A chaque itération, la fonction Q augmente et on peut montrer que l'algorithme converge vers un maximum local de la vraisemblance.

Revenons à l'équation (2.2). L'idée est que quand les observations $\mathcal{X}$ et le paramètre $\theta = \theta^{(k)}$ sont fixés, alors $\mathcal{S}$ est une variable aléatoire de densité $p(s|\mathcal{X};\theta^{(k)})$. On peut alors écrire

$$E_{\mathcal{S}}\left[\log p(\mathcal{X},\mathcal{S};\theta)|\mathcal{X};\theta^{(i)}\right] = \int \log(p(\mathcal{X},s;\theta))p(s|\mathcal{X};\theta^{(i)})ds$$

Si S ne prend que les valeurs discrètes cette expression devient :

$$\int \log(p(\mathcal{X},s;\theta))p(s|\mathcal{X};\theta^{(i)})ds = \sum_{s=1}^{M} \log p(\mathcal{X},s;\theta)P(S=s|\mathcal{X};\theta^{(i)}).$$

Par définition, la densité de la variable aléatoire X s'écrit alors

$$p(x;\theta) = \sum_{m=1}^{M} \pi_m p(x;\theta^{(m)}).$$

[1] Expectation
[2] Maximisation

avec θ_m le paramètre de la loi de X dans le groupe m. Et la log-vraisemblance des données complétées s'exprime

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\delta_{\{S_i=m\}}\log\left(\pi_m p(x_i;\theta_m)\right).$$

On obtient cette écriture en imitant le cas où S est observée.
Ainsi l'étape E conduit à l'écriture

$$Q(\theta^{(k)},\theta)=\sum_{i=1}^{n}\sum_{k=1}^{M}E_{S,\theta^{(k)}}(\delta_{\{S_i=m\}})\log(\pi_m p(x_i;\theta_m))$$

avec

$$\begin{aligned}E_{S,\theta}(\delta_{\{S_i=m\}}) &= P(S=m|X=x_i;\theta)=\frac{P(S=m,X=x_i;\theta)}{P(X=x_i;\theta)} \quad \text{(aux abus de notations près)}\\ &= \frac{\pi_m p(x_i;\theta_m)}{\sum_{m=1}^{M}\pi_k p(x_i;\theta_m)}=T_{m,i}^{(t)}\end{aligned}$$

L'algorithme EM prend une forme simple

1. quand S est une variable aléatoire discrête comme ci-dessus
2. quand la loi des données complètes $p(\mathcal{Z};\theta)$ est dans la famille exponentielle :

$$p(\mathcal{Z};\theta)=b(\mathcal{Z})\exp\left\{\mathbf{c}^\top(\theta)\mathbf{t}(\mathcal{Z})\right\}/a(\theta),$$

où $\mathbf{t}(\mathcal{Z})$ est un vecteur de statistiques exhaustives, $\mathbf{c}(\theta)$ est un vecteur de fonctions du paramètre, $a(\theta)$ and $b(\mathcal{Z})$ sont des fonctions scalaires. L'étape E s'écrit alors

$$Q\left(\theta,\theta^{(k)}\right)=E_{\mathcal{S}}\left[\log(b(\mathcal{Z});\theta^{(k)}\right]+\mathbf{c}^T(\theta)E_S\left[\mathbf{t}(\mathcal{Z});\theta^{(k)}\right]-\log(\mathbf{a}(\theta)) \tag{2.4}$$

On remarque que le premier terme du membre de droite de l'équation (2.4) ne dépend pas de θ. Aussi l'étape M est réduite à

$$\theta^{(i+1)}=arg\max_{\theta\in\Theta}\mathbf{c}^T(\theta)E_S\left[\mathbf{t}(\mathcal{Z});\theta^{(i)}\right]-\log(a(\theta)) \tag{2.5}$$

Exercice -

1. Écrire l'algorithme EM pour estimer les paramètres d'un mélange de deux lois normales univariées.
2. Généraliser l'algorithme au cas multivarié.

2.2.5 Cas du mélange de lois de Gauss

Nous donnons ici l'algorithme EM dans le cas de mélange de lois de Gauss. On remarque tout d'abord que S ne prend que les valeurs discrètes de telle sorte que :

$$\int\log(p(\mathcal{X},s;\theta))p(s|\mathcal{X};\theta^{(i)})ds=\sum_{s=1}^{M}\log p(\mathcal{X},s;\theta)P(S=s|\mathcal{X};\theta^{(i)})$$

Par définition, la densité de la variable aléatoire X s'écrit

$$p(x;\theta)=\sum_{m=1}^{M}\pi_m\phi(x;\mu^{(m)},\Sigma^{(m)}).$$

et la vraisemblance des données complétées (on imite ici le cas où S serait observée) s'exprime

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\delta_{\{S_i=m\}}\log\left(\pi_m\phi(x_i;\mu^{(m)},\Sigma^{(m)}\right)$$

Ainsi l'étape E conduit à l'écriture

$$Q(\theta^{(t)},\theta)=\sum_{i=1}^{n}\sum_{k=1}^{M}E_{S,\theta^{(t)}}(\delta_{\{S_i=m\}})\log(\pi_m\phi(x_i;\mu^{(m)},\Sigma^{(m)}))$$

avec

$$E_{S,\theta}(\delta_{\{S_i=m\}})=P(S=m|x_i;\theta)=\frac{\pi_m\phi(x_i;\mu^{(m)},\Sigma^{(m)})}{\sum_{m=1}^{M}\pi_k\phi(x_i;\mu^{(m)},\Sigma^{(m)})})=T_{m,i}^{(t)}$$

Et l'étape M est donnée par l'annulation des dérivées par rapport aux différentes composantes de θ de $Q(\theta',\theta)$ sous la contrainte $\sum_{m=1}^{M}\pi_m^{(t+1)}=1$. On obtient alors pour les probabilités

$$\pi_m^{(t+1)}=\frac{\sum_{i=1}^{n}T_{m,i}^{(t)}}{\sum_{i=1}^{n}\sum_{m=1}^{M}T_{1,k}^{(t)}}=\frac{1}{n}\sum_{i=1}^{n}T_{m,i}^{(t)}$$

pour les moyennes

$$\boldsymbol{\mu}_{(t+1)}^{(m)}=\frac{\sum_{i=1}^{n}T_{m,i}^{(t)}\mathbf{x}_i}{\sum_{i=1}^{n}T_{m,i}^{(t)}}$$

et pour les variances

$$\Sigma_{(t+1)}^{(m)}=\frac{\sum_{i=1}^{n}T_{m,i}^{(t)}(\mathbf{x}_i-\boldsymbol{\mu}_{(t+1)}^{(m)})(\mathbf{x}_i-\boldsymbol{\mu}_{(t+1)}^{(m)})^\top}{\sum_{i=1}^{n}T_{m,i}^{(t)}}$$

2.2.6 Propriétés de l'algorithme EM

Notons k la densité conditionnelle des données complètes sachant la variable observée :

$$k(\mathcal{Z}|\mathcal{X};\theta)=\frac{p(\mathcal{Z};\theta)}{p(\mathcal{X};\theta)}$$

On peut alors écrire la log-vraisemblance des données complètes comme suit

$$\log\mathcal{L}(\theta)=\log\mathcal{L}(\theta;\mathcal{X})+\log\left(k(\mathcal{Z}|\mathcal{X};\theta)\right)$$

où $\mathcal{L}(\theta;\mathcal{X})$ est la vraisemblance de θ sachant les observations $\mathcal{X}$. En prenant l'espérance en fonction de S, on obtient

$$Q(\theta,\theta^{(i)})=\log\mathcal{L}(\theta;\mathcal{X})+H(\theta,\theta^{(i)})$$

avec $H(\theta,\theta^{(i)})=E_S\left[\log\left(k(\mathcal{Z}|\mathcal{X};\theta)\right)\right]$. Il s'en suit

$$\begin{aligned}\log\mathcal{L}\left(\theta^{(i+1)}\right)-\log\mathcal{L}\left(\theta^{(i)}\right)&=\left\{Q\left(\theta^{(i+1)};\theta^{(i)}\right)-Q\left(\theta^{(i)};\theta^{(i)}\right)\right\}\\&\quad-\left\{H\left(\theta^{(i+1)};\theta^{(i)}\right)-H\left(\theta^{(i)};\theta^{(i)}\right)\right\}\end{aligned}$$

Par l'inégalité de Jensen, on a $H(\theta^{(i+1)};\theta^{(i)}) \leq H(\theta^{(i)};\theta^{(i)})$ or on a aussi $Q(\theta^{(i+1)};\theta^{(i)}) \leq Q(\theta^{(i)};\theta^{(i)})$. Ainsi la vraisemblance est croissante d'une itération de l'EM à la suivante :

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

Et pour une suite bornée de valeurs de vraisemblance $\{\mathcal{L}(\theta^{(i)})\}$, $\mathcal{L}(\theta^{(i)})$ converge vers un $\mathcal{L}^*$.

Avant la formulation générale de l'algorithme EM données par Dempster et al. (1977), certains auteurs ont donné des résultats de convergence pour des cas particuliers, on peut citer en particulier Baum et al. (1970) pour les chaînes de Markov cachées. Cependant, c'est Wu (1983) qui donne le premier des résultats plus généraux. Il montre que quand les données complètes appartiennent à la famille exponentielle avec un espace de paramètres et quand la fonction Q satisfiat certaine conditions de différentiabilité alors, toute séquence EM converge vers un point stationnaire de la vraisemblance (pas nécessairement un maximum). Si la vraisemblance admet plusieurs points stationnaire la convergence de l'EM vers un maximum local ou global dépend de la valeur initiale $\theta^{(0)}$ du paramètre.

Quand la vraisemblance est bornée, on peut montrer que l'estimateur du maximum de vraisemblance est un point fixe de l'algorithme EM. De plus, si une suite d'itérations de l'EM $\{\theta^{(i)}\}$ satisfait les conditions

1. $[\partial Q(\theta;\theta^{(i)})/\partial\theta]_{\theta=\theta^{(k+1)}} = \mathbf{0}$, et
2. la suite $\{\theta^{(i)}\}$ converge vers une valeur θ^* et $\log k(\mathcal{X}|\mathcal{S};\theta)$ est assez régulière,

alors on a $[\partial \log \mathcal{L}(\theta)/\partial\theta]_{\theta=\theta^*} = \mathbf{0}$; voir Little and Rubin (2002) and Wu (1983).

Vitesse de convergence

La vitesse de convergence de l'algorithme EM a aussi un intérêt pratique. La première remarque est que l'algorithme EM converge moins vite que la vitesse quadratique des algorithmes de Newton. Dempster et al. (1977) montrent que la vitesse de l'EM est linéaire et qu'elle dépend de la proportion d'information contenue dans les observations.

En pratique, on observe que l'algorithme EM est plutôt plus performant que les algorithmes de Newton quand on est loin d'un minimum local, mais qu'à proximité de la solution un algorithme de Newton est plus performant. On peut donc avoir avantage à coupler ces deux méthodes.

Propriétés de l'algorithme EM

L'algorithme EM a plusieurs propriétés intéressantes, en particulier :

1. L'algorithme est numériquement stable puisqu'à chaque itération la vraisemblance augmente.
2. Il converge vers une solution globale sous des conditions assez générales.
3. Il est facile à implémenter et require peu d'espace mémoire.
4. Le coût numérique d'une itération de l'algorithme est généralement faible ; ce qui peut compenser le fait que l'algorithme est lent.
5. On peut l'utiliser pour inférer les données manquantes.

On peut aussi citer quelques inconvénients :

1. L'algorithme EM ne fournit pas automatiquement une estimation de la matrice de covariance des paramètres estimés. Cependant on peut remédier facilement à cet inconvénient (McLachlan and Krishnan, 1997, Chap. 4).
2. L'EM peut converger très lentement dans certains cas.
3. Pour certains problèmes, il n'est pas possible de traiter analytiquement l' étape E et/o l'étape M.

2.3 Chaînes de Markov discrètes

Nous allons maintenant inroduire les chaînes de Markov qui sont le second "ingrédient" dont on a besoin pour définir des modèles à chaîne de Markov cachée. Nous nous intéressons uniquement aux chaînes de Markov discrètes. *[une bonne référence : Grimmett and Stizaker (2001, chapitre 6)]*.

2.3.1 Définitions et exemple

Une suite de variables aléatoires discrètes définies sur $E \subset \mathbb{N}$, $\{S_t, t \in \mathbb{N}\}$, est appelée **chaîne de Markov** (à temps discret) si pour tout $t \in \mathbb{N}$ elle vérifie la propriété de Markov

$$P(S_{t+1}|S_t, \cdots, S_1) = P(S_{t+1}|S_t)$$

C'est à dire que conditionner par tout le passé jusqu'au temps t est équivalent à conditionner seulement par la valeur la plus récente S_t. Autrement dit, le passé et le futur ne sont dépendants qu'à travers le présent. La propriété de Markov peut être vue comme la relaxation la plus faible de l'hypothèse d'indépendance.

Critère fondamental - Soit une suite $Y = (Y_t)_{t\geq 1}$ de variables aléatoires indépendantes et de même loi, à valeurs dans un espace F , et soit f une application mesurable de $E \times F$ dans E. Supposons que la suite$X = (X_t)_{t\geq 0}$ est définie par la relation de récurrence :

$$\forall t \geq 0, \qquad X_{t+1} = f\left(X_t, Y_{t+1}\right),$$

et supposons que la suite Y est indépendante de X_0. Alors X est une chaîne de Markov homogène.

A une chaîne de Markov, on associe des **probabilités de transition**

$$P(S_{t'+t} = j|S_{t'} = i) = \gamma_{ij}(t)$$

Si ces probabilités ne dépendent pas de t' , on dit que la chaîne de Markov est homogène. Sinon elle est non homogène. Dans le cas des chaînes de Markov homogène, on notera dans la suite $\Gamma(t)$ la matrice dont les coefficients sont les $\gamma_{ij}(t)$.

Proposition - La matrice de transition $\Gamma = (\gamma_{i,j})_{(i,j)\in E^2}$ est stochastique : la somme des termes de n'importe quelle ligne de *Gamma* donne toujours 1.

$$\forall i \in E, \quad \sum_{j\in E} \gamma_{i,j} = 1.$$

Dans le cas où $t = 1$, on note $\Gamma = \Gamma(1)$ la matrice de transition à un pas de temps,

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \cdots & & \cdots \\ \gamma_{M1} & \cdots & \gamma_{MM} \end{pmatrix}$$

Propriété de Chapman-Kolmogorov
Les chaînes de Markov homogènes à espace d'état fini vérifient l'équation de Chapman-Kolmogorov

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

La propriété de Chapman-Kolmogorov implique que pour tout $t \in \mathbb{N}$

$$\Gamma(t) = \Gamma(1)^t$$

c'est à dire que la matrice de transition correspondant à t pas de temps est égale à la matrice de transition à 1 pas de temps à la puissance t.

Pour une chaîne de Markov, on s'intéresse aussi à la probabilité $P(S_t = j)$ d'être dans l'état j au temps t. On note $u(t)$ le vecteur

$$u(t) = (P(S_t = 1), \cdots, P(S_t = M)), \quad t \in \mathbb{N}$$

On dit que $u(1)$ est la **distribution initiale** de la chaîne de Markov. Par ailleurs, on obtient la distribution $u(t+1)$ au temps $t+1$ par

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

En effet,

$$\begin{aligned} P(S_{t+1} = j) &= \sum_{i=1}^{M} P(S_{t+1} = j | S_t = i) P(S_t = i) \\ &= \sum_{i=1}^{M} \gamma_{ij} P(S_t = i) \\ &= u(t)\Gamma_{.j} \end{aligned}$$

Exemple -Imaginons que la succession de jours ensoleillés et pluvieux est telle que le temps d'un jour donné ne dépend que de celui de la veille et que les probabilités de transition sont données par

	jour $t+1$	
jour t	pluie	soleil
pluie	0.9	0.1
sunny	0.6	0.4

Le temps suit donc une chaîne de Markov à deux états avec pour matrice de transition

$$\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{pmatrix}$$

Supposons maintenant que le premier jour il fasse beau. La distribution du 1er jour est donc

$$u(1) = (P(S_1 = 1), P(S_1 = 2)) = (0, 1)$$

La distribution du lendemain et des jours suivants est obtenue en multipliant $u(1)$ à droite par Γ :

$$u(2) = (P(S_2 = 1), P(S_2 = 2)) = u(1)\Gamma = (0.6, 0.4)$$

$$u(3) = (P(S_3 = 1), P(S_3 = 2)) = u(2)\Gamma = (0.78, 0.22), \text{ etc.}$$

Distributions stationnaires
Une chaîne de Markov homogène de matrice de transition Γ admet une **distribution stationnaire** δ si

$$\delta\Gamma = \delta \tag{2.6}$$

$$\sum_{m=1}^{M} \delta_m = 1, \delta_m \geq 0 \tag{2.7}$$

L'équation (2.6) exprime la stationnarité et l'équation (2.7) permet de garantir que δ définit une loi de probabilité.

Par exemple, la chaîne de Markov associée à la matrice de transition

$$\Gamma = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

a pour distribution stationnaire $\delta = \frac{1}{32}(15, 9, 8)$.

Comme $u(t+1) = u(t)\Gamma$, une chaîne de Markov qui a pour distribution initiale sa distribution stationnaire continuera à être distribuée suivant la distribution stationnaire à tout temps t. On dit alors que c'est une chaîne de Markov stationnaire.
Remarque - L'homogénéité ne suffit pas à avoir une chaîne de markov stationnaire ; il faut en plus l'égalité des distibutions initiales et staionnaires.

Fonction d'autocorrélation
Dans la suite nous aurons besoin de comparer la fonction d'autocorrélation du modèle avec celle des données. Pour une chaîne de Markov $\{S_t\}$ à M états, la fonction d'autocorrélation est donnée par

$$Cov(S_t, S_{t+k}) = \delta V \Gamma^k v^T - (\delta v)^2$$

où $v = (1, 2, \cdots, M)$ et $V = diag(1, 2, \cdots, M)$. Si Γ est diagonalisable de valeurs propres $\omega_1, \cdots, \omega_M$, on peut montrer que pour les k positifs la covariance se met sous la forme

$$Cov(S_t, S_{t+k}) = \sum_{i=1}^{M} a_i b_i \omega_i^k$$

2.3.2 Inférence pour les probabilités de transition

Quand on a une réalisation d'une chaîne de Markov (à espace d'état fini) et qu'on veut estimer les probabilités de transition, une approche consiste à utiliser les estimateurs empiriques ; c'est à dire à compter le nombre d'occurence de chacun des transitions et d'en déduire une estimation des probabilités de transition. Par exemple, pour la réalisation suivante d'une chaîne de Markov à 3 états :

```
2332111112 3132332122 3232332222 3132332212 3232132232
3132332223 3232331232 3232331222 3232132123 3132332121
```

la matrice de countage des occurences de transition est

$$(n_{ij}) = \begin{pmatrix} 4 & 7 & 6 \\ 8 & 10 & 24 \\ 6 & 24 & 10 \end{pmatrix}$$

où on note n_{ij} le nombre de transitions observées de l'état i vers l'état j. Comme le nombre de transition de 2 vers 3 est 24 et que le nombre total de transition au départ de 2 est 8+10+24, une estimation de γ_{23} est $24/42$. On obtient de la même façon une estimation de la matrice de transition

$$(n_{ij}) = \begin{pmatrix} 4/17 & 7/17 & 6/17 \\ 8/42 & 10/42 & 24/42 \\ 6/40 & 24/40 & 10/40 \end{pmatrix}$$

Nous allons voir que cet estimateur empirique qui est assez naturel est aussi l'estimateur du maximum de vraisemblance conditionnellement à la première observation.

Supposons qu'on veuille estimer les $(m-1)m$ paramètres γ_{ij} d'une chaîne de Markov à M états à partir d'une réalisation $s_1, \cdots, s_T$. La vraisemblance conditionnée par la première observation s'écrit

$$L(\Gamma; c_1, \cdots, c_T) = \prod_{i=1}^{M} \prod_{j=1}^{M} \gamma_{ij}^{n_{ij}}$$

Et la log vraisemblance est donc

$$\ell(\Gamma; c_1, \cdots, c_T) = \sum_{i=1}^{T} \left(\sum_{j=1}^{T} f_{ij} \log \gamma_{ij} \right) = \sum_{i=1}^{T} \ell_i$$

Les ℓ_i ne dépendent pas des mêmes paramètres et on peut donc maximiser ℓ en maximisant chacun des ℓ_i.
On remplace γ_{ii} par $1 - \sum_{k \neq i} \gamma_{ik}$. On calcule la dérivée de ℓ_i par rapport à γ_{ij} et on l'annule

$$0 = \frac{-n_{ii}}{1 - \sum_{k \neq i} \gamma_{ik}} + \frac{n_{ij}}{\gamma_{ij}} = \frac{-n_{ii}}{\gamma_{ii}} + \frac{n_{ij}}{\gamma_{ij}}$$

Ainsi, si aucun dénominateur n'est égal à zéro, on obtient $n_{ij}\gamma_{ii} = f_{ii}\gamma_{ij}$ et $\gamma_{ii} \sum_{j=1}^{m} n_{ij}$. Ceci implique que pour un maximum (local) de la vraisemblance

$$\gamma_{ii} = \frac{n_{ii}}{\sum_{j=1}^{m} n_{ij}} \text{ et } \gamma_{ij} = \frac{n_{ij}\gamma_{ii}}{n_{ii}} = \frac{n_{ij}}{\sum_{j=1}^{m} n_{ij}}$$

Nous n'avons pas utiliser de propriété de stationnarité pour écrire les estimateurs du maximum de vraisamblance conditionnés par la première observation. Si on peut supposer que la réalisation est issue d'une chaîne de Markov stationnaire, on peut alors utiliser la vraisemblance non conditionnée. Elle s'écrit comme la vraisemblance conditionnée multipliée par la distribution stationnaire δ. Dans certains cas particuliers, on peut obtenir des formes explicites des estimateurs du maximum de la vraisembalnce non conditionnée.

Dans le cas d'une chaîne de Markov irréductible et récurrente positive, la loi forte des grands nombres est en vigueur : la moyenne d'une fonction f sur les instances de la chaîne de Markov est égale à sa moyenne selon sa probabilité stationnaire. Plus précisément, sous l'hypothèse

$$\sum_{i \in E} |f(i)| delta_i \ < +\infty,$$

on a presque sûrement :

$$\lim_n \ \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \ = \ \sum_{i \in E} f(i)\, \delta_i \ = \ \delta f.$$

La moyenne de la valeur des instances est donc, sur le long terme, égale à l'espérance suivant la probabilité stationnaire. En particulier, cette équivalence sur les moyennes s'applique si f est la fonction indicatrice d'un sous-ensemble A de l'espace des états :

$$\lim_n \ \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(X_k) \ = \ \sum_{i \in A} \ \delta_i \ = \ \delta(A).$$

Cela permet d'approcher la probabilité stationnaire par la distribution empirique (qui est un histogramme construit à partir d'une séquence particulière).

Chapitre 3

Inférence dans les modèles à changement de régimes markovien

Rappelons qu'un modèle à chaîne de Markov cachée est un processus à temps discret $(X_t)_{t\in 1,\cdots,N}$ avec deux composantes $X_t = (S_t, Y_t)$. On supposera dans ce chapitre que S_t est à valeurs dans un ensemble fini $E = 1, ..., M$ (si $E \subset \mathbb{R}^d$, on parle de "modèles à espace d'état"). Y_t est à valeurs dans un espace F qui pourra être discret ou continu selon l'application.

Ce processus vérifie de plus les propriétés d'indépendance conditionnelles ci-dessous :

- $P(S_t|S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(S_t|S_{t-1} = s_{t-1})$
- $P(Y_t|S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(Y_t|S_t = s_t)$

En pratique, généralement seul le processus $\{Y_t\}$ sera observé. S_t est une variable "cachée" ou "latente" ou "manquante". On suppose de plus que la loi de Y_t conditionnellement à $S_t = s$ admet une densité $g(.; \theta^{(s)})$ pour tout t. On dit que $g(.; \theta^{(s)})$, $s \in E$ sont les probabilités d'émission.

Nous nous intéressons ici à l'inférence par maximum de vraisemblance. Une autre approche classique consiste à utiliser l'estimation bayésienne.

3.1 Fonction de vraisemblance

On note ici $\theta \in \Theta$, l'ensemble des paramètres inconnus (ie la propabilité initiale π, la matrice de transition Q et les paramètres des probabilités d'émission $\theta^{(s)}$). On cherche à estimer θ à partir des observations disponibles $\{y_1, \cdots, y_T\}$ en cherchant la valeur des paramètres qui rend la fonction de vraisemblance $\mathcal{L}(\theta) = p(y_1, \cdots, y_T; \theta)$ maximale.

On a

$$p(y_1, \cdots, y_T; \theta) = \sum_{s_1,\cdots,s_T \in \{1,\cdots,M\}^T} p(y_1, \cdots, y_T, s_1, \cdots, s_T; \theta) \tag{3.1}$$

avec $p(y_1, \cdots, y_T, s_1, \cdots, s_T; \theta)$ la vraisemblance des données complètes qui vérifie, en utilisant

la formule de Bayes et les propriétés d'indépendance conditionnelle du modèle,

$$\begin{aligned}
\mathcal{L}_c(\theta) &= p(y_1,\cdots,y_T,s_1,\cdots,s_T;\theta) \\
&= p(s_1,y_1;\theta)\prod_{t=2}^{T} p(s_t,y_t|s_1,\cdots,s_{t-1},y_1,\cdots,y_{t-1};\theta) \\
&= p(s_1,y_1;\theta)\prod_{t=2}^{T} p(s_t,y_t|s_{t-1},y_{t-1};\theta) \\
&= p(s_1,y_1;\theta)\prod_{t=2}^{T} p(s_t|s_{t-1},y_{t-1};\theta)p(y_t|s_{t-1},y_{t-1},s_t;\theta) \\
&= p(s_1;\theta)p(y_1;s_1,\theta)\prod_{t=2}^{T} p(s_t|s_{t-1};\theta)p(y_t|s_t;\theta) \\
&= \pi_1 g(y_1;\theta^{(s_1)})\prod_{t=2}^{T} Q(s_{t-1},s_t)g(y_t;\theta^{(s_t)})
\end{aligned}$$

Cependant (3.1) est peu utile en pratique pour calculer numériquement la fonction de vraisemblance (et donc l'estimateur du maximum de vraisemblance) car elle fait intervenir la somme de M^T termes, chacun des termes étant un produit de T termes !

La méthode usuelle pour maximiser la vraisemblance consiste alors à utiliser un algorithme EM. Cet algorithme, comme dans le cas d'un modèle de mélange est un algortithme itératif dont chaque itération est constituée de deux étapes. Dans la première étape, on caractérise la chaîne cachée sachant les paramètres estimés à l'itération précédente et dans la seconde, on estime les paramètres du modèle sachant l'information obtenue sur la chaîne cachée. Dans la seconde étape on procède donc, en quelques sortes, comme si on connaissait les valeurs prises la chaîne de Markov.

3.2 Apprentissage supervisé : *étape M*

On fait ici comme si la chaîne de Markov était observée. Ceci peut-être le cas pour certaines applications. On parle alors d'apprentissage supervisé.

Supposons donc qu'on dispose des observations $\{s_1,y_1,\cdots,s_T,y_T\}$, la vraisemblance des données complètes s'écrit :

$$\begin{aligned}
\mathcal{L}_c(\theta) &= p(y_1,\cdots,y_T,s_1,\cdots,s_T;\theta) \\
&= \pi_1 g(y_1;\theta^{(s_1)})\prod_{t=2}^{T} Q(s_{t-1},s_t)\prod_{t=1}^{T} g(y_t;\theta^{(s_t)}) \\
&= \pi_1 g(y_1;\theta^{(s_1)})\prod_{s,s'\in\{1,\cdots,M\}} Q(s',s)^{N(s',s)}\prod_{s\in\{1,\cdots,M\}}\prod_{t\in\{1,\cdots,T\}|s=s_t} g(y_t;\theta^{(s)})
\end{aligned}$$

avec $N(s',s)=\text{card}(t\in\{2,\cdots,T\}|s_{t-1}=s',s_t=s)$ le nombre de transitions entre l'état s' et l'état s. On en déduit la log-vraisemblance des données complètes :

$$l_c = ln(\pi_1)+\sum_{s',s\in\{1,\cdots,M\}} N(s',s)ln(Q(s',s))+\sum_{s\in\{1,\cdots,M\}}\sum_{t\in\{1,\cdots,T\}|s_t=s} lng(y_t;\theta^{(s)})$$

Ceci s'écrit encore sous une forme qui sera utile pour la suite :

$$\begin{aligned} l_c &= \sum_{s\in\{1,\cdots,M\}} u_1(s)ln(\pi_1) + \sum_{s',s\in\{1,\cdots,M\}} ln(Q(s',s)) \sum_{t\in\{1,\cdots,T\}} v_t(s',s) \\ &+ \sum_{s\in\{1,\cdots,M\}} \sum_{t\in\{1,\cdots,T\}|s_t=s} u_t(s)lng(y_t;\theta^{(s)}) \end{aligned}$$

avec $u(s) = \delta_{\{s\}}(s_t)$ et $v_t(s',s) = \delta_{\{s',s\}}(s_{t-1},s_t)$.

La fonction de log-vraisemblance complète s'écrit donc comme une somme de $M+2$ termes qui peuvent s'optimiser séparément. En ce qui concerne les paramètres qui décrivent l'évolution de la chaîne cachée, on retrouve les estimateurs du maximum de vraisemblance classiques pour une chaîne de Markov, soit :

$$\hat{\pi}(s) = u_1(s)$$

$$\begin{aligned} \hat{Q}(s',s) &= \frac{\sum_{t=2}^{T} v_t(s',s)}{\sum_{t=1}^{T}\sum_{s=1}^{M} v_t(s',s)} \\ &= \frac{N(s',s)}{N(s')} \end{aligned}$$

avec $N(s') = \sum_{s=1}^{M} N(s',s) = \text{card}(t \in \{2,\cdots,T\}|s_t = s)$.

Il existe par ailleurs des expressions analytiques pour les paramètres des probabilités d'émission pour certaines lois usuelles. Par exemple pour les émissions gaussiennes,

$$g(y_t;\theta^{(s)}) = \frac{1}{\sigma^{(s)}\sqrt{2\pi}} \exp\left(-\frac{(x-\mu^{(s)})^2}{2(\sigma^{(s)})^2}\right)$$

et on peut montrer que les estimateurs qui maximisent la log-vraisemblance sont :

$$\hat{\mu}^{(s)} = \frac{\sum_{t=1}^{T} u_t(s)y_t}{\sum_{t=1}^{T} u_t(s)}$$

$$\hat{\sigma}^{(s)} = \frac{\sum_{t=1}^{T} u_t(s)(y_t - \hat{\mu}^{(s)})}{\sum_{t=1}^{T} u_t(s)}$$

En l'absence d'expression analytique on utilise généralement des algorithmes numériques.

Une fois les paramètres estimés, on peut utiliser le modèle pour faire de la classification.

3.3 Classification : *étape E*

Supposons que l'on dispose de nouvelles observations du processus $\{Y_t\}$, notées abusivement $\{y_1,\cdots,y_T\}$ et qu'on cherche à retrouver la séquence d'états $\{s_1,\cdots,s_T\}$ qui correspond à ces observations. Il existe deux méthodes classiques : l'*algorithme de Viterbi* et l'*algorithme Forward-Barckward*.

3.3.1 Algorithme de Viterbi

L'algorithme de Viterbi (Andrew Viterbi, 1967) permet de calculer la séquence la plus vraisemblable de la chaîne cachée sachant les observations, c'est à dire

$$arg\max p(s_1,\cdots,s_T|y_1,\cdots,y_T;\theta)$$

3.3.2 Algorithme Forward-Backward (FB)

L'algorithme FB permet de calculer les quantités

$$p(s_t|y_1, \cdots, y_T)$$

appelées probabilités de lissage. Cet algorithme consiste à parcourir les observations dans les deux sens ("forward" puis "backward") de manière itérative.

– L'**étape Forward** est basée sur la formule ci-dessous, valable pour $t \in \{1, \cdots, T\}$,

$$\begin{aligned} p(s_t|y_1, \cdots, y_{t-1}; \theta) &= \sum_{s_{t-1} \in \{1, \cdots, M\}} p(s_{t-1}, s_t|y_1, \cdots, y_{t-1}; \theta) \text{ par Bayes/ptés de la chaîne de Marko} \\ &= \sum_{s_{t-1} \in \{1, \cdots, M\}} p(s_t|s_{t-1}, y_1, \cdots, y_{t-1}; \theta) p(s_{t-1}|y_1, \cdots, y_{t-1}; \theta) \text{ par Bayes} \\ &= \sum_{s_{t-1} \in \{1, \cdots, M\}} p(s_t|s_{t-1}; \theta) \alpha_{t-1} \text{ par les ptés d'ind. cond.} \end{aligned}$$

qui permet d'exprimer les *probabilités de prédiction* $p(s_t|y_1, \cdots, y_{t-1}; \theta)$ en fonction des *probabilités de filtrage* $\alpha_{t-1} = p(s_{t-1}|y_1, \cdots, y_{t-1}; \theta)$. Puis

$$p(s_t|y_1, \cdots, y_t; \theta) = \frac{p(s_t, y_t|y_1, \cdots, y_{t-1}; \theta)}{p(y_t|y_1, \cdots, y_{t-1}; \theta)} \text{ par Bayes}$$

avec le numérateur donné par

$$\begin{aligned} p(s_t, y_t|y_1, \cdots, y_{t-1}; \theta) &= p(s_t|y_1, \cdots, y_{t-1}; \theta) p(y_t|s_t, y_1, \cdots, y_{t-1}; \theta) \\ &= \sum_{s_{t-1} \in \{1, \cdots, M\}} p(s_t|s_{t-1}) p(s_{t-1}|y_1, \cdots, y_{t-1}; \theta) p(y_t|s_t; \theta) \\ &= \sum_{s_{t-1} \in \{1, \cdots, M\}} Q(s_{t-1}, s_t) \alpha_{t-1} p(y_t|s_t; \theta) \end{aligned}$$

et le dénominateur par

$$p(y_t|y_1, \cdots, y_{t-1}; \theta) = \sum_{s_t \in 1, \cdots, T\}} p(s_t, y_t|y_1, \cdots, y_{t-1}; \theta)$$

Ces formules permettent donc d'estimer simplement $\alpha_t = p(s_t|y_1, \cdots, y_t; \theta)$ en fonction de $\alpha_{t-1} = p(s_{t-1}|y_1, \cdots, y_{t-1}; \theta)$. On peut ainsi calculer itérativement $\alpha_t = p(s_t|y_1, \cdots, y_t; \theta)$ pour $t \in \{1, \cdots, T\}$ en partant, pour $t = 1$, de

$$p(s_1|y_1; \theta) = \frac{p(y_1, s_1; \theta)}{p(y_1; \theta)}$$

avec $p(y_1, s_1; \theta) = p(y_1|s_1; \theta) p(s_1; \theta)$ et $p(y_1; \theta) = \sum_{s_1 \in \{1, \cdots, M\}} p(y_1, s_1; \theta)$.

– L'**étape Backward** est basée sur la formule, valable pour $t \in \{1, \cdots, T-1\}$,

$$p(s_t|y_1, \cdots, y_T; \theta) = p(s_t|y_1, \cdots, y_t; \theta) \sum_{s_t \in \{1, \cdots, M\}} \frac{p(s_{t+1}|s_t; \theta}{p(s_{t+1}|y_1, \cdots, y_t; \theta)} p(s_{t+1}|y_1, \cdots, y_T)$$

qui permet de calculer itérativvement les probabilité de lissage $p(s_t|y_1, \cdots, y_T; \theta)$ à partir de $p(s_{t+1}|y_1, \cdots, y_T)$. On remarque que pour $t = T$, $p(s_T|y_1, \cdots, y_T; \theta)$ est donné par l'étape forward.

3.4 Algorithme EM

On se place maintenant dans le cas plus courant ou on ne dispose pas d'information sur la chaîne $\{S_t\$(apprentissagenon-supervis).L'algorithmeEMestunalgorithmeitratifpartantd'unevaleurini$ des paramètres. Notons $\hat{\theta}_k$ la valeur des paramètres après k itérations. Rappelons que la log-vraisemblance des données complètes peut se récrire :

$$\begin{aligned} l_c &= \sum_{s\in\{1,\cdots,M\}} u_1(s)ln(\pi_1) + \sum_{s',s\in\{1,\cdots,M\}} ln(Q(s',s)) \sum_{t\in\{1,\cdots,T\}} v_t(s',s) \\ &+ \sum_{s\in\{1,\cdots,M\}} \sum_{t\in\{1,\cdots,T\}|s_t=s} u_t(s)lng(y_t;\theta^{(s)}) \end{aligned}$$

avec $u(s) = \delta_{\{s\}}(s_t)$ et $v_t(s',s) = \delta_{\{s',s\}}(s_{t-1},s_t)$. La chaîne $\{s_t\}$ n'étant pas observée, ces quantités sont inconnues et on les remplace par leurs "meilleurs" aprximations disponibles, c'est à dire par leurs espérances conditionnelles connaissants les observations $\{y_t\}$ et a valeur courante des paramètres $\hat{\theta}_k$:

- $u(s) = \delta_{\{s\}}(s_t) = \mathbb{I}_{\{s\}}(s_t)$

$$\hat{u}(s;\hat{\theta}_k) = E\left[\mathbb{I}_{\{s\}}(S_t)|y_1,\cdots,y_T;\hat{\theta}_k\right] = P(S_t = s|y_1,\cdots,y_T;\hat{\theta}_k)$$

  On retrouve les probabilités de lissage qu'on peut calculer par l'algorithme forward-backward.
- $v_t(s',s) = \delta_{\{s',s\}}(s_{t-1},s_t) = \mathbb{I}_{\{s',s\}}(s_{t-1},s_t)$

$$\hat{v}(s,s';\hat{\theta}_k) = E\left[\mathbb{I}_{\{s,s'\}}(S_{t-1},S_t)|y_1,\cdots,y_T;\hat{\theta}_k\right] = P(S_{t-1} = s, S_t = s'|y_1,\cdots,y_T;\hat{\theta}_k)$$

Ces quantités peuvent être calculée par l'algorithme forward-backward. On défini ensuite la fonction intermédiaire :

$$\begin{aligned} Q(\theta;\hat{\theta}_k) &= \sum_{s\in\{1,\cdots,M\}} \hat{u}_1(s;\hat{\theta}_k)ln(\pi_1) + \sum_{s',s\in\{1,\cdots,M\}} ln(Q(s',s)) \sum_{t\in\{1,\cdots,T\}} \hat{v}_t(s',s;\hat{\theta}_k) \\ &+ \sum_{s\in\{1,\cdots,M\}} \sum_{t\in\{1,\cdots,T\}|s_t=s} \hat{u}_t(s;\hat{\theta}_k)lng(y_t;\theta^{(s)}) \end{aligned}$$

puis

$$\hat{\theta}_{k+1} = arg\max_{\theta\in\Theta}(Q(\theta;\hat{\theta}_k))$$

Comme dans le cas des données complètes, on peut trouver des expressions analytiques pour les paramètres qui dérivent l'évlution de la chaîne cachée

$$\hat{\pi}_{k+1} = \hat{u}_1(s;\hat{\theta}_k)$$

$$q_{k+1}(s,s') = \frac{\sum_{t=2}^{T}\hat{v}_t(s',s;\hat{\theta}_k)}{\sum_{t=1}^{T}\sum_{s=1}^{M}\hat{v}_t(s',s;\hat{\theta}_k)}$$

et pour les paramètres des lois d'émissions dans certains familles de lois. Par exemple pour les émissions gaussiennes,

$$\hat{\mu}_{k+1}^{(s)} = \frac{\sum_{t=1}^{T}\hat{u}_t(s;\hat{\theta}_k)y_t}{\sum_{t=1}^{T}\hat{u}_t(s;\hat{\theta}_k)}$$

$$\hat{\sigma}_{k+1}^{(s)} = \frac{\sum_{t=1}^{T}\hat{u}_t(s;\hat{\theta}_k)(y_t-\hat{\mu}^{(s)})}{\sum_{t=1}^{T}\hat{u}_t(s;\hat{\theta}_k)}$$

L'algorithme EM pour les modèles à chaîne de Markov cachée admet le même type de propriétés que pour les modèles de mélange : convergence quadratique vers la solution, problèmes des extrêmas locaux.

Chapitre 4

Validation de modèles

4.1 Introduction

Nous pouvons distinguer des outils de différentes natures :
- Adéquation : les tests d'adéquation[1] permettent de décider si un modèle est 'bon'
- Comparaison : les critères bayésiens sont généralement utilisés pour sélectionner le meilleur modèle parmis plusieurs possibilités

D'autre part, nous verrons que la plupart des critères statistiques sont basés sur des résultats asymptotiques (quand le nombre d'observations tend vers l'infini). Quand on veut raisonner à nombre d'observations fini, il est souvent difficile d'obtenir des critères théoriques. On a alors recours à la simulation pour valider les modèles : on utilise des méthodes de Monte Carlo.

4.2 Tests d'adéquation

4.2.1 Observations continues

Soit X une variable aléatoire de fonction de répartition F et $x_1, \cdots, x_n$ n réalisations indépendantes de X. Les hypothèses d'un test d'adéquation à une distribution de fonction de répartition H s'écrivent

$$H_0 : F = H \text{ contre } H_1 : F \neq H$$

- Test de Kolmogorov
  La statistique du test de Kolmogorov est basé sur l'écart entre la fonction de répartition empirique de l'achantillon et la fonciton de répartition théorique.

  $$D_n = \max_{x \in \mathbb{R}} |F_n(x) - H(x)|$$

  avec F_n la fonction de répartition de l'échantillon $x_1, \cdots, x_n$.
- Test de Cramer-von Mises
  La statistique du test de Cramer-von Mises est basé sur la norme L^2 de l'écart entre la fonctions de répartition théoriques et empiriques.

  $$S = \int_{\mathbb{R}} [F_n(x) - H(x)]^2 dF_n(x)$$

  On démontre que

  $$S = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - H(x_{(i)}) \right]^2$$

[1] En anglais : goodness-of-fit tests

où les $x_{(i)}$ sont les statistiques d'ordre de l'échantillon.
L'indicateur d'écart du test de Cramer-Von Mises prend mieux en compte l'ensemble des données en ce sens que la somme des écarts intervient. Le test de Kolmogorov est donc beaucoup plus sensible à l'existence de points abérents dans un échantillon que le test de Cramer-Von Mises. On pense généralement que ce dernier test est plus puissant, mais cela n'a pas été démontré théoriquement.

– Test d'Anserson-Darling
La statistique du test d'Anderson-Darling est données par $A = -n - S$ où

$$S =_s um_{i=1}^{n} \frac{2i-1}{n} \left(\ln F(x_{(i)} + \ln(1 - F(x_{(n+1-i)})\right)$$

Ce test donne un poids plus important aux queues de la loi que les tests de Kolmogorov ou Cramer-von Mises. Mais, les valeurs critiques du test d'Anderson-Darling dépendent de la distribution testée.

Remarque - Ces test sont essentiellement utilisés dans des problèmes de modélisation de distribution univariée. On peut généraliser à des problèmes comportant un petit nombre de variables.

4.2.2 Observation discrètes

Quand les observations sont discrètes, on utilise un test du chi2 pour comparer la probabilité des différentes réponses observées avec celles prédites par le modèle. On remarque qu'on peut considérer la statistique de test du chi 2 comme une distance entre le modèle théorique et les observations.

Ce test sera utilisé, par exemple pour valider des modèles à classes latentes.

Les principaux inconvénients du test du chi 2 sont les suivants :

– Le test du chi 2 est raisonnable dans les situations où la taille de l'échantillon est importante et le nombre de variables faible. S'il trop de cellules sont d'effectif faible, le test n'est pas valide.
– On remarque que le nombre de réponses possibles dans un modèle à classes latentes est 2^p avec p le nombre total de variables. Par exemple, si on a 4 variables, il y a 16 réponses possibles. Si on a 20 variables, le nombre de réponses possibles s'élève à 1,048,576. Imaginez le nombre d'observation dont on a besoin pour avoir au moins une observations par réponse possible... Si les réponses sont fortement dépendantes, il faut encore plus d'observations...

Alternatives

Utiliser des mesures marginales. Exemple moyennes marginales : on peut alors calculer une erreur en moyenne quadratique entre les moyennes. Ce critère est peut discriminant.

Mesures bivariées : on construit des tables de contingences (observées et prédites) pour 2 variables. Puis on utilise une mesure d'association pour tester l'adéquation des 2 tables (corrélation, ...).

4.3 Critères basés sur la vraisemblance

$$\text{AIC} = -2 \ln L + 2 \ln k$$

$$\text{BIC} = -2 \ln L + k \ln n$$

avec L la vraisemblance, k le nombre de paramètres et n le nombre d'observations.

On choisit généralement le modèle qui conduit à la valeur la plus faible du critère BIC (resp. AIC).

1. It is independent of the prior.
2. It can measure the efficiency of the parameterized model in terms of predicting the data.
3. It penalizes the complexity of the model where complexity refers to the number of parameters in model.
4. It is exactly equal to the minimum description length criterion but with negative sign.
5. It can be used to choose the number of clusters according to the intrinsic complexity present in a particular dataset.

Il faut garder en tête que ce type de critères permet de comparer des modèles : ce ne sont pas des mesures d'adéquation !

4.3.1 Entropy

Quand les variables latentes sont cachées, on peut aussi considérer l'entropie.

$$E = -\sum_{i=1}^{n}\sum_{k=1}^{M} \alpha_{ik} \ln \alpha_{ik} \in [0, +\infty[$$

avec $\alpha_{ik} = P(i \in \text{groupe } k)$

on préfère généralement l'entropie relative

$$1 - \frac{E}{n \ln M} \in [0, 1]$$

L'entropie est proche de proche de 1 si la classification est bonne, elle est proche, proche de 0 sinon.

4.4 Méthodes de Monte Carlo

The most widespread method for model validation consists in comparing certain statistics calculated from the observations with those corresponding to the considered model. In general, several criteria are used, such as the matching of the mean and the variance of the marginal distributions, or more generally its cdf. When the temporal dependence is important for the applications, other features are also considered, like the autocorrelation functions or the distribution of the time duration of sojourns below or above given levels.

Meanwhile, the authors often perform only visual comparisons. Such approach remains not entirely satisfactory because it does not make it possible to decide whether the observed differences are significant or not. A more formal method, based on Monte Carlo tests, is proposed below. For the sake of simplicity, it is presented in the simple case of comparing means, but its generalization to other statistics is straightforward.

Let $\{y_t\}_{t=1,\cdots,T}$ be an observed sequence of a real valued process $\{Y_t\}$ with mean m and $\{Z_t\}$ a process corresponding to the model which has to be validated. The mean of the marginal distribution of $\{Z_t\}$ is denoted m_0. We want to test

$$H_0 : m = m_0 \quad \text{versus} \quad H_1 : m \neq m_0 \tag{4.1}$$

The considered test statistic is the empirical mean $\overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$, and the associated decision rule is given by

$$H_0 \text{ is rejected if } \overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \in R(\alpha)$$

where α is the level of the test. The critical region $R(\alpha)$ is such that $P_{H_0}(\overline{Y} \in R(\alpha)) = \alpha$.

In order to compute $R(\alpha)$, we need to know the distribution of the test statistic $\overline{Y}$ when H_0 is true. When the model is complex, it is not always possible to derive the exact distribution of the test statistic. In this case, we can use the Monte Carlo method described hereafter to approximate this distribution :

1. Simulate B time series of length T with the model :

$$\begin{matrix} \{z_1^{(1)}, & \cdots & , z_T^{(1)}\} \\ & \vdots & \\ \{z_1^{(B)}, & \cdots & , z_T^{(B)}\} \end{matrix}$$

2. Compute the empirical mean $\overline{z}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} z_t^{(i)}$ for each simulated sample $i = 1, \cdots, B$
3. Approximate the distribution of $\overline{Y}$ under H_0 by the empirical distribution of $\{\overline{y}^{(1)}, \cdots, \overline{y}^{(B)}\}$. This allows to compute an approximation of $P_{H_0}(\overline{Y} \in R)$ for any region $R \subset \mathbb{R}$ or equivalently deduce an approximative critical region $\tilde{R}(\alpha)$ such that $\frac{1}{B} card(\{i \in \{1, \cdots, B\} | \overline{z}^{(i)} \in \tilde{R}(\alpha)\}) = \alpha$.

Finally, H_0 will be accepted if and only if $\overline{y} \in \tilde{R}(\alpha)$.