Atelier 1

1 Rappels et généralités

Nous présentons dans cette partie comment modéliser une série, qui une fois la tendance et saisonnalité supprimées, est stationnaire. A noter que le seul fait de supprimer la tendance et la saisonnalité ne rend pas la série nécessairement stationnaire, puisque cela n'affecte pas la variance et l'autocovariance, qui doivent être constantes pour un processus stationnaire.

1.1 Autocorrélation partielle

Le coefficient de corrélation partielle entre les deux variables X_1 et X_n d'un processus stochastique $(X_t)_t$ est le coefficient de corrélation entre les deux variables auxquelles on a retranché leurs meilleures explications en terme de $X_2, \ldots, X_{n-1}$:

$$r(X_1, X_N) = corr(X_1 - P_{X2,\ldots,Xn-1}(X_1), X_n - P_{X2,\ldots,Xn-1}(X_n)),$$

où corr est le coefficient de corrélation classique et $P_{X2,\ldots,Xn-1}(X_1)$ est la projection de la variable X_1 dans l'espace vectoriel engendré par les variables $X_2, ..., X_{n-1}$. Ce coefficient exprime la dépendance entre les variables X_1 et X_n qui n'est pas due aux autres variables $X_2, ..., X_{n-1}$.
Dans le logiciel R, la fonction `pacf` permet ces estimations.

1.2 Les processus mixtes ARMA(p,q)

Cette classe générale de modèles définit des processus sous la forme d'une récurrence autorégressive avec un second membre de type moyenne mobile.
Un processus autorégressif moyenne mobile d'ordres p et q, ARMA(p,q), est de la forme :

$$X_t = \sum_{k=1}^{p} a_k X_{t-k} - \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t, \tag{1}$$

où $(\epsilon_t)_t$ est un bruit blanc de variance σ^2.
La stationnarité d'un ARMA(p,q) est assurée lorsque toutes les racines du polynôme

$$A(z) = 1 - a_1 z - \ldots - a_p z^p$$

sont de module différent de 1. Rappelons, que la causalité de ces processus est obtenue quant à elle quand les racines de $A(z)$ sont de module supérieur strictement à 1. Le polynôme $A(z)$ forme avec

$$B(z) = 1 - b_1 z - \ldots - b_q z^q$$

les deux polynômes caractéristiques du processus. On supposera également que les polynômes A et B n'ont pas de racine commune, afin de s'assurer qu'il n'y a pas de représentation plus courte.
Le processus s'écrit plus simplement avec l'opérateur retard L comme :

$$A(L)X_t = B(L)\epsilon_t,$$

où $LX_t = X_{t-1}$ et plus généralement $L^k X_t = X_{t-k}$, pour tout $k \in \mathbb{N}^*$.
Le tableau suivant récapitule les principales propriétés des processus MA(q), AR(p) et ARMA(p,q).

Table 1 – Propriétés des processus MA(q), AR(p) et ARMA(p,q)

Modèle	MA(q)	AR(p)	ARMA(p,q)
Covariance	$cov(h) = 0, \forall h > q$	$\lim_{h\to\infty} cov(h) = 0$	$\lim_{h\to\infty} cov(h) = 0$
Corrélation	$\rho(h) = 0, \forall h > q$	$\lim_{h\to\infty} \rho(h) = 0$	$\lim_{h\to\infty} \rho(h) = 0$
Corr partielle	$\lim_{h\to\infty} r(h) = 0$	$r(h) = 0, \forall h > p. \quad r(p) = a_p$	

2 Estimation, choix du modèle et prévision

Les principaux modèles de séries temporelles ont été définis. A partir d'une série observée, il faut maintenant choisir un modèle, éventuellement plusieurs, estimer ses paramètres et enfin faire des prévisions pour les réalisations futures. Dans le cas où l'on hésite entre plusieurs modèles, des critères de choix de modèles seront utilisés pour sélectionner le meilleur d'entre eux.

2.1 Estimation

L'estimation des paramètres des modèles ARMA est faite par maximum de vraisemblance. L'expression de la vraisemblance étant généralement trop complexe pour que l'on puisse obtenir un maximum explicite, des algorithmes numériques (type Newton) sont utilisés.

2.2 Choix de modèle

L'étude des autocovariances, autocorrélations et autocorrélations partielles peut conduire à certaines hypothèses sur la nature du modèle. Une fois quelques modèles choisis, et leurs paramètres estimés, des critères vont être utilisés pour choisir le modèle qui effectue le meilleur compromis entre :

— ajustement à la série de données,
— complexité du modèle.

Il est en effet très important de prendre en compte ce compromis, car si on ne s'intéressait qu'à coller au mieux aux données, on serait tenter de choisir un modèle ARMA avec un très grand nombre de paramètres. Or, plus il y a de paramètres, plus il faut de données pour les estimer. Et donc pour un nombre d'observations fixé de la série, plus le modèle sera complexe, moins bien seront estimés les paramètres.
Les critères de choix des modèles les plus courants sont :

1. Le critère AIC (Akaike Information Criterion), qui sera généralement préféré si l'objectif de l'étude est de faire de la prévision, et qui est défini par :

$$AIC = -2\log L(\theta) + 2v,$$

où L(.) est la vraisemblance du modèle, θ représente les paramètres du modèle et v le nombre de ces paramètres.

2. Le critère BIC (Bayesian Information Criterion) sera quant à lui généralement préféré si l'objectif de l'étude est de s'ajuster à la série observée, et est défini par :

$$BIC = -2\log L(\theta) + nv,$$

où n est le nombre d'observations de la série.

Les modèles ayant la plus petite valeur du critère devront être choisis. Ces deux critères conduisent donc à sélectionner des modèles dont la vraisemblance est grande, en la pénalisant par la complexité du modèle.

2.3 Prévision

L'objectif est de prévoir la valeur que va prendre la variable aléatoire X_{n+h}, h étant appelé l'horizon de la prévision, ayant observé la réalisation des variables aléatoires $X_1, ..., X_n$. Dans le cadre d'une modélisation ARMA, on choisit d'estimer X_{n+h} par une combinaison linéaire des $X_j, j = 1, \ldots, n$ précédents

$$X_{n,h} = c_{1,h}X_1 + ... + c_{n,h}X_n.$$

Les coefficients $c_{j,h}$ sont estimés de sorte qu'ils minimisent :

$$E[(X_{n+h} - c_{1,h}X_1 - ... - c_{n,h}X_n)^2].$$

L'estimateur ainsi défini n'est autre que la projection de X_{n+h} sur l'espace vectoriel engendré par les variables $X_1, ..., X_n$.

Quelques remarques sur la prévision :

i) L'erreur de prévision à horizon 1 pour un processus ARMA est le bruit d'innovation ϵ_{n+1}.

ii) La variance de l'erreur de prévision à horizon h dans un processus ARMA croît depuis la variance du bruit d'innovation (valeur prise pour h = 1) jusqu'à la variance du processus lui-même.

3 Mise en oeuvre sous R : processus ARMA, ARIMA.

— La fonction `arima.sim(modele,n)` permet de simuler un processus ARMA(p,q) défini par l'équation (1).
Les paramètres a_k et b_j du processus sont précisés dans le paramètre modèle de la fonction :
`modele<-list(ar=c(a1,..., ap),ma=c(b1,..., bq))`.
— Pour simuler un modèle $ARIMA(p,d,q)$ il faut ajouter le composant `order=c(p,d,q)` dans le paramètre modele de la fonction arima.sim.
— La fonction ar permet d'estimer les paramètres d'un processus AR(p) :
`out<-ar(data,aic=TRUE,order.max=NULL)`. L'ordre p du processus autorégressif est choisi (inférieur à order.max) à l'aide du critère AIC (si l'option aic est validée).
— La fonction `arima` permet d'estimer les paramètres :
d'un ARMA(p,q) : `out<-arima(serie,order=c(p,0,q))`
d'un ARIMA(p,d,q) : `out<-arima(serie,order=c(p,d,q))`
— Parmi les sorties de cette fonction, on peut obtenir :
  — out$coef : estimation des coefficients,
  — out$aic : valeur du critère AIC,
  — out$resid : estimation des résidus.
— La fonction `p=predict(out,h)` permet d'effectuer une prévision à l'horizon h. Parmi les sorties de cette fonction,
`p$pred`
contient les prévisions, et
`p$se`
contient l'écart-type de l'erreur de prévision.
— Il n'existe pas de fonction prédéfinie pour calculer un intervalle de confiance sur les prévisions, mais cela peut être fait manuellement grâce à ces deux sorties de la fonction predict.

3.1 Un exemple ARMA(1,1)

1. Simuler 500 réalisations du processus suivant :

$$Y_t = 0.6Y_{t-1} + \epsilon_t + 0.5\epsilon_{t-1}$$

2. Représenter à chaque fois sur un même graphique, la série simulée, la fonction d'autocorrélation ainsi que la fonction d'autocorrélation partielle.
3. Ajuster un processus moyenne-mobile d'ordre 1 aux données.
4. Ajuster un processus auto-régressif d'ordre 1 aux données.
5. Ajuster un processus ARMA d'ordre (1,1).
6. Utiliser le critère d'Akaike pour décider du choix du modèle, on utilisera pour cela la fonction AIC.

On pensera à utiliser la fonction arima de R afin d'effectuer les estimations des questions 3 à 5.
Remarque : un processus ARIMA(p,d,q) est un processus tel que si on le différencie d fois on obtient un processus ARMA(p,q).

3.2 Simulation de processus ARMA

1. Donner la définition d'un processus ARMA(p,q). Rappeler les conditions sur les coefficients pour que ce processus soit stationnaire.
2. A l'aide de la fonction arima.sim, simuler plusieurs processus AR(p) et MA(q) (p et q pas trop grands). Avant toute simulation, écrire la définition mathématique du processus à simuler et veillez à ce que les conditions de stationnarité soient respectées.
3. Observer les autocorrélations empiriques (partielles ou non). Que constatez-vous ?
4. Simuler quelques ARMA(p,q), observer et interpréter les autocorrélations empiriques (partielles ou non).
5. Faire de même avec un modèle ARIMA(p,d,q), avec un d assez petit.
6. Représenter à chaque fois sur un même graphique, la série simulée, la fonction d'autocorrélation ainsi que la fonction d'autocorrélation partielle.

3.3 Identification d'un processus ARMA

1. Récupérer le fichier de données serie1.dat.
2. Ce processus vous semble-t-il modélisable par un processus ARMA ? Pourquoi ?
3. On travaille désormais avec la série obtenue en appliquant la fonction diff à la série. Quelle transformation a-t-on effectuée ? Pourquoi ?
4. En observant les autocorrélations empiriques et autocorrélations partielles empiriques, proposer des modèles AR(p) et MA(q) d'ordre faible pour modéliser cette série.
5. Estimer les paramètres des deux modèles sélectionnés.
6. Tester la blancheur des résidus.
7. Conclure pour choisir un modèle.

3.4 Prévision dans un processus ARMA

Soit le processus

$$X_t - X_{t-1} + 1/2X_{t-2} - 1/3X_{t-3} = \epsilon_t,$$

avec $(\epsilon_t)_t$ un bruit blanc gaussien centré réduit.

1. Après avoir identifié ce processus, simuler 50 réalisations de longueur 105.
2. Pour chaque simulation, extraire les 100 premières valeurs et estimer les paramètres d'un AR(3).
3. Pour chaque simulation, prédire les cinq valeurs suivantes.
4. Donner une estimation de l'erreur moyenne (biais) et de la variance de l'erreur de prévision à 1, 2, 3, 4 et 5 pas.
5. Recommencer en rallongeant la durée d'observation, et comparer aux résultats précédents.

3.5 Processus non stationnaires

Considérons le processus $(X_t)_{t\in\mathbb{Z}}$ défini, pour tout $t \in \mathbb{Z}$, par :

$$X_t = 1 + t^2 + \frac{1}{3}X_{t-1} + \epsilon_t,$$

où $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc gaussien de variance égale à 1.

1. Créer la fonction *SimProcess* prenant en argument un entier n et renvoyant une trajectoire de taille n du processus $(X_t)_{t\in\mathbb{Z}}$.
2. Utiliser la fonction *SimProcess* pour simuler une trajectoire de taille 1000 du processus $(X_t)_{t\in\mathbb{Z}}$.
3. Représenter sur un même graphique la série simulée, sa fonction d'autocorrélation empirique et sa fonction d'autocorrélation partielle empirique. Commenter.
4. Étudier la stationnarité des processus $(X_t)_{t\in\mathbb{Z}}$, $(\Delta X_t)_{t\in\mathbb{Z}}$ et $(\Delta^2 X_t)_{t\in\mathbb{Z}}$ (Δ représente l'opérateur de différenciation, *i.e.* $\Delta = 1 - L$ où L est l'opérateur retard).
5. Ajuster un AR(1), un MA(2), un ARMA(1,2) et un ARMA(2,2) à la série des différenciations secondes de la série simulée dans la question 2. Utiliser le critère d'Akaike pour décider du choix du modèle.
6. Étudier l'adéquation de la structure des résidus du modèle retenu avec celle d'un bruit blanc. Commenter.
7. Montrer théoriquement que le processus $(\Delta^2 X_t)_{t\in\mathbb{Z}}$ est un processus ARMA(1,2).

3.6 AR faibles

Nous proposons dans cet exercice d'étudier numériquement le comportement asymptotique de l'estimateur des moindres carrés ordinaires du paramètre autorégressif du modèle suivant :

$$X_t = aX_{t-1} + \epsilon_t, \tag{2}$$

où, pour tout $t \in \mathbb{Z}$,

$$\epsilon_t = \eta_t^2 \eta_{t-1}, \tag{3}$$

avec $(\eta_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires gaussiennes indépendantes centrées et de variance égale à 1.

1. Rappeler la condition sur le paramètre a sous laquelle le processus $(X_t)_{t\in\mathbb{Z}}$ est stationnaire au second ordre.
2. Créer la fonction *Bruit* qui prend en argument un entier n et qui renvoie une trajectoire de taille n du processus $(\epsilon_t)_{t\in\mathbb{Z}}$ défini dans (3).
3. Créer la fonction *SimulProcess* prenant en argument un entier n et un réel a et renvoyant une trajectoire de taille n du processus $(X_t)_{t\in\mathbb{Z}}$ défini dans (2) (penser à utiliser la fonction *Bruit* implémentée juste avant).
4. Prenons $a = 0.3$. Simuler une trajectoire de taille 5000 du processus $(X_t)_{t\in\mathbb{Z}}$. Tracer sur une même fenêtre cette trajectoire, ses autocorrélations empiriques et ses autocorrélations partielles empiriques. Commenter.
5. L'estimateur des moindres carrés ordinaires de a est donné par :

$$\hat{a}_n = \frac{\sum_{t=1}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2}. \tag{4}$$

   Créer la fonction *Estim* qui prend en argument une série X et qui renvoie l'estimateur des moindres carrés ordinaires du paramètre autorégressif dans la modélisation de X par un AR(1) comme dans (2).
6. Prenons toujours $a = 0.3$. Générer 1000 trajectoires chacune de taille $n = 5000$ de $(X_t)_{t\in\mathbb{Z}}$ et calculer pour chaque trajectoire l'estimateur des moindres carrés ordinaires défini dans (4). Stocker ces estimations dans un vecteur de taille 1000.
7. Représenter graphiquement l'histogramme de la distribution des estimateurs calculés dans la question précédente. Commenter.