Cours GIMAS8AD

Modélisation et prévision

Séries chronologiques - Séance 2
Les processus ARIMA

Frédéric Sur
École des Mines de Nancy

https://members.loria.fr/FSur/enseignement/modprev/

Motivation / objectif

Box et Jenkins 1970 (à la suite de Yule et Slutsky 1927) :

de nombreuses phénomènes temporels (ou leur dérivée*), dans de nombreux domaines, peuvent être représentés par des* **processus stationnaires**.

Rappel :
les processus stationnaires considérés sont de la forme

$$X_t = m + \sum_{j=0}^{+\infty} a_j \varepsilon_{t-j}$$

avec $\sum_{j=0}^{+\infty} |a_j| < +\infty$ et (ε_t) **bruit blanc** (faible)

En fait (ε_t) sera même généralement un bruit blanc *gaussien*.
$\rightarrow$ on parle de *processus linéaire gaussien*.

Séance 2

Séance 2

Les processus MA *(=moyenne mobile)*

Définition : processus MA(q) (Slutsky 1927)

Ce sont les processus (X_t) du type :

$$\forall t,\ X_t = m + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

où $q \geqslant 1$, $(\theta_i)_{1\leqslant i\leqslant q}$ réels, et (ε_t) est un bruit blanc (**gaussien** dans notre cours).

Convention d'écriture

$$X_t - m = \Theta(B)\varepsilon_t$$

où $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$
et B est l'opérateur retard : $B^k \varepsilon_t = \varepsilon_{t-k}$.

Remarque : les processus MA(q) sont **stationnaires**.

Caractérisation de l'ordre d'un MA(q)

$$\forall h \geqslant 0,\ \rho(h) = \frac{\mathrm{Cov}(X_t, X_{t+h})}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t+h})}} = \frac{\gamma(h)}{\gamma(0)}.$$

Pour un processus MA(q) : (calcul algébrique simple)

$$\gamma(h) = \begin{cases} \left(1 + \theta_1^2 + \cdots + \theta_q^2\right)\sigma^2 & \text{si } h = 0 \\ \left(\theta_h + \theta_{h+1}\theta_1 + \cdots + \theta_q\theta_{q-h}\right)\sigma^2 & \text{si } 1 \leqslant h \leqslant q \\ 0 & \text{si } h > q \end{cases}$$

Propriété : fonction d'autocorrélation (ACF) d'un MA(q)

$$\rho(h) = \begin{cases} 1 & \text{si } h = 0 \\ \frac{\theta_h + \theta_{h+1}\theta_1 + \cdots + \theta_q\theta_{q-h}}{\theta_1^2 + \cdots + \theta_q^2} & \text{si } 1 \leqslant h \leqslant q \\ 0 & \text{si } h > q \end{cases}$$

Intérêt : détermination de q.

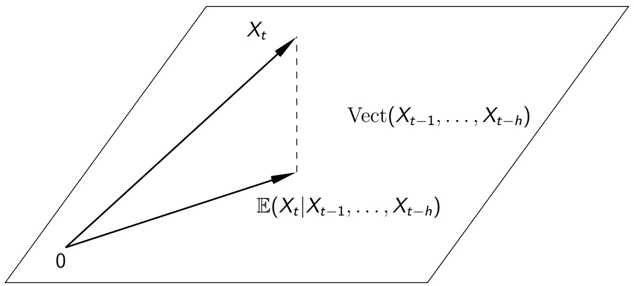
Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Espérance conditionnelle si (X_t) gaussien

Définition ("rappel") : espérance conditionnelle

$$\mathbb{E}(X_t | X_{t-1}, \dots, X_{t-h}) = a_1(h)X_{t-1} + \cdots + a_h(h)X_{t-h}$$

est la « meilleure » approximation de X_t par une combinaison linéaire de $(X_{t-1}, \dots, X_{t-h})$ dans L^2 muni du produit scalaire $< X, Y >= \mathbb{E}(XY)$.
(cf régression)



Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Fonction d'autocorrélation partielle (PACF)

(X_t) stationnaire donc dans $(L^2, \mathbb{E}(XY))$.

$$\mathbb{E}(X_t | X_{t-1}, \dots, X_{t-h}) = a_1(h)X_{t-1} + \cdots + a_h(h)X_{t-h}$$

Définition : PACF de (X_t)

Le terme $r(h) = a_h(h)$ est la *corrélation partielle* d'ordre h.

Propriété (cf poly : géométrie dans un espace de Hilbert)

$$\forall h \geqslant 1,\ a_h(h) = \mathrm{corr}\left(\widetilde{X}_t, \widetilde{X}_{t-h}\right)$$

où : $\widetilde{X}_t = X_t - \mathbb{E}(X_t | X_{t-1}, \dots, X_{t-h+1})$
et : $\widetilde{X}_{t-h} = X_{t-h} - \mathbb{E}(X_{t-h} | X_{t-1}, \dots, X_{t-h+1})$

Interprétation : corrélation entre X_t et X_{t-h} lorsque l'influence des variables $X_{t-1}, \dots, X_{t-h+1}$ a été retirée.

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Retour aux processus MA(q)

L'expression générale du PACF d'un processus MA(q) est compliquée.

Cas particulier : MA(1)

$$\forall t,\ X_t = m + \varepsilon_t + \theta \varepsilon_{t-1}.$$

alors :

$$r(h) = a_h(h) = \frac{\theta^h(\theta^2 - 1)}{1 - \theta^{2h+2}}.$$

$\rightarrow$ $(|r(h)|)_{h \geqslant 1}$ décroît de manière exponentielle (si $\theta \neq 1$).

Résumé : les processus MA *(moyennes mobiles)*

Définition : processus MA(q)

$$X_t = m + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

$$X_t - m = \Theta(B)\varepsilon_t$$

avec $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$.

Propriété

Un processus MA(q) est stationnaire.

Corrélogrammes :

- ACF : $\rho(h) = 0$ si $h > q$,
- PACF : $r(h)$ tend vers 0. ($q = 1$ « vitesse exponentielle »)

$\longrightarrow$ permet l'identification d'un MA(q).

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Les processus AR *(=autorégressif)*

Définition **: processus AR(p)** (Yule 1927)

Ce sont les processus (X_t) du type :

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

où $p \geqslant 1$, $(\phi_i)_{1 \leqslant i \leqslant p}$ réels, et (ε_t) est un bruit blanc (gaussien).

Convention d'écriture : (constante c vs. moyenne m)

$$\Phi(B)X_t = c + \varepsilon_t \quad \text{ou :}\ \Phi(B)(X_t - m) = \varepsilon_t$$

avec $\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ (B : opérateur retard)
et : $(1 - \phi_1 - \dots - \phi_p)m = c$

Question : un processus AR(p) est-il stationnaire ?

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Exemple : processus AR(1)

$$\forall t,\ X_t = \phi X_{t-1} + \varepsilon_t \quad (\text{ ou } (1 - \phi B)X_t = \varepsilon_t\ )$$

Exemple-type, car :

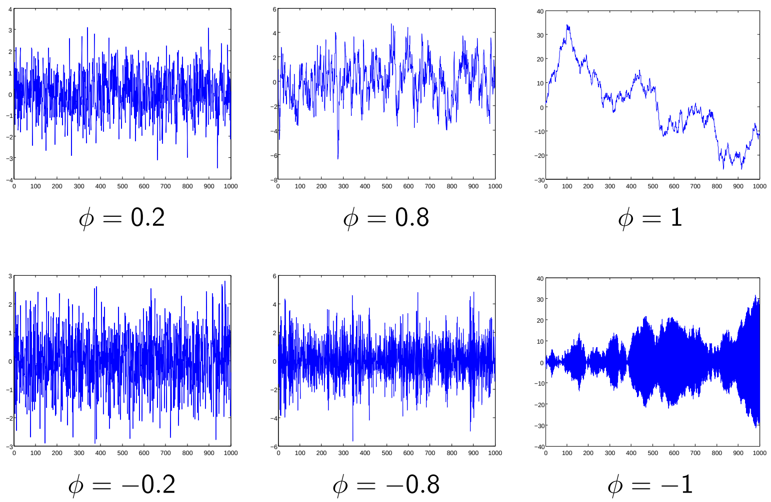
$$\Phi(B)X_t = \varepsilon_t \quad \Leftrightarrow \quad \prod_{i=1}^{p} (1 - \phi_i' B)\, X_t = \varepsilon_t$$

- si $|\phi| < 1$ alors
  $X_t = \left(1 + \phi B + \phi^2 B^2 + \dots\right)\varepsilon_t$
  $\rightarrow$ (X_t) **stationnaire**, « **équivalent** » **à MA(∞)**.
- si $|\phi| > 1$ alors
  $X_t = -\left(\phi^{-1} B^{-1} + \phi^{-2} B^{-2} + \dots\right)\varepsilon_t$
  $\rightarrow$ (X_t) **stationnaire** (mais **non causal** sous cette forme).
- si $\phi = 1$ (marche aléatoire) : $X_t = X_{t-1} + \varepsilon_t$.
  On a bien $\forall t,\ \mathbb{E}(X_t) = \mathbb{E}(X_{t-1})$.
  Mais var(X_t) dépend de t (non constante).
  $\rightarrow$ (X_t) **non-stationnaire**.

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Simulation de processus AR(1)

(ε_t) i.i.d. $\sim \mathcal{N}(0,1)$



Caractérisation de l'ordre d'un AR(p) stationnaire

Propriété : PACF d'un processus AR(p)

Si $r(h)$ est la fonction d'autocorrélation partielle :

$$\forall h > p, \ r(h) = 0$$

Preuve : $X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=p+1}^{h} 0 \cdot X_{t-i} + \varepsilon_t$, puis projection sur $(X_{t-1}, \dots, X_{t-h})$ sachant ε_t décorrélé de $X_{t-1}, \dots, X_{t-h}$.

Intérêt : détermination de p.

Propriété : ACF d'un processus AR(p)

$(|\rho(h)|)_{h\geqslant}$ décroît vers 0 de manière exponentielle.

Preuve : cf équations de Yule-Walker, qui permettent de retrouver les ϕ_i à partir des $\rho(h)$ (poly).

Résumé : les processus AR *(autorégressifs)*

Définition : processus AR(p)

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

$$\Phi(B)(X_t - m) = \varepsilon_t \qquad \Phi(B)(X_t) = c + \varepsilon_t$$

avec $\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$.

Propriété

Un processus AR(p) est stationnaire si Φ n'a pas de racine unité.

Corrélogrammes :

- ACF : $\rho(h)$ décroît de « manière exponentielle » vers 0
- PACF : $\forall h > p, \ r(h) = 0$

$\longrightarrow$ permet l'identification d'un AR(p).

Les processus ARMA *(=autorégressif moyenne mobile)*

Définition : processus ARMA(p,q)

Ce sont les processus (X_t) du type :

$$\forall t, \ X_t = \theta_0 + \sum_{i=1}^{p} \phi_i X_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

où $p \geqslant 1$, $(\phi_i)_{1 \leqslant i \leqslant p}$ et $(\theta_j)_{0 \leqslant j \leqslant q}$ réels, et (ε_t) b.b. gaussien.

Notation symbolique : $\Phi(B)X_t = \theta_0 + \Theta(B)\varepsilon_t$.
ou : $\Phi(B)(X_t - \mu) = \Theta(B)\varepsilon_t$
avec : $\theta_0 = \Phi(B)\mu$ $\quad$ (θ_0 joue le rôle de c dans proc. AR)

Remarque : si Φ a ses racines à l'extérieur du cercle unité, alors on peut écrire :

$$X_t = \sum_{j=0}^{+\infty} a_j \varepsilon_{t-j} \quad \text{(sous forme MA}(+\infty)\text{) + relations entre } a_j)$$

$\rightarrow$ *cf. principe de parcimonie.*

Séance 2

Les processus non stationnaires

Forme générale d'une chronique (modèle linéaire) :

$$X_t = T_t + S_t + u_t.$$

avec T_t tendance, S_t saisonnalité, u_t perturbation aléatoire.

$\rightarrow$ (X_t) n'est pas stationnaire.

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Les processus ARIMA *(=ARMA intégré)*

Opérateur de différentiation :
$\nabla X_t = X_t - X_{t-1} = (1-B)X_t$.

Définition **: processus ARIMA(p,d,q)**

Ce sont les processus (X_t) du type :

$$\Phi(B)(1-B)^d X_t = \theta_0 + \Theta(B)\varepsilon_t$$

où Φ sans racine unitaire, $p, d, q \geqslant 0$,
et (ε_t) est un bruit blanc gaussien (écart-type σ).

Remarque : processus ARIMA non stationnaires,
mais d fixé tel que $(1-B)^d X_t$ stationnaire.

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Dérivation des processus

ARIMA : $\Phi(B)(1-B)^d X_t = \theta_0 + \Theta(B)\varepsilon_t$

Intérêt 1 : permet de traiter les chroniques avec tendance.

- $X_t = at + b + u_t$:
  alors $(1-B)X_t = a + (1-B)u_t$ est stationnaire.
- $X_t = at^2 + bt + c + u_t$:
  $(1-B)^2 X_t$ est stationnaire.
- etc.

Intérêt 2 : permet de stabiliser la variance.

- $X_t = X_{t-1} + u_t$ (marche aléatoire),
  alors $(1-B)X_t = u_t$ est stationnaire.

Remarque : dans les deux cas, *racine unité* dans la partie AR.

Modélisation et prévision
F. Sur - ENSMN
Processus stationnaires
Processus MA
Processus AR
Processus ARMA
Processus non stationnaires
Processus ARIMA
Processus SARIMA
La méthode de Box-Jenkins
Identification
Estimation
Validation
Prévision
Sous SAS
Conclusion

Les processus SARIMA

Permettent de traiter le cas de la saisonnalité.

$\rightarrow$ cf prochaine séance.

Séance 2

Modélisation des chroniques par (S)ARIMA

Comment modéliser une chronique X_t par un ARIMA(p,d,q) ?

$\rightarrow$ méthode de Box et Jenkins (1970)

1. **Identification** des paramètres p, d, q.
2. **Estimation** des θ_j,ϕ_i, et σ
3. **Validation** du modèle.
4. **Prévision** du futur.

Identification du modèle ARIMA(p,d,q)

0. Transformation de la chronique (log, exp, $\sqrt{\cdot}$, Box-Cox) pour stabiliser la variance, passage au ≪ modèle additif ≫.

1. Identification de d.

Indices de non-stationnarité :

- ≪ ça se voit ≫ sur le graphique (tendance),
- l'ACF ne décroît pas assez vite.

$\rightarrow$ dériver.

Attention à ne pas *trop* dériver...
Si X_t déjà stationnaire, alors $Y_t = X_t - X_{t-1}$ a tendance à présenter un $\rho(1)$ (ACF) ≪ proche ≫ de -1.
(introduction d'autocorrélations négatives)
$\rightarrow$ indication de d trop grand...

Identification du modèle ARIMA(p,d,q)

2. Identification de p et q.

Utilisation des corrélogrammes.

Rappel :

	ACF	PACF
AR(p)	$\rho(h)$: décroissance exponentielle vers 0	$r(h)=0$ si $h>p$
MA(q)	$\rho(h)=0$ si $h>q$	$r(h)$ décroissance $\sim$ exponentielle vers 0

Remarque 1 : on dispose des ACF et PACF **empiriques**. SAS donne des intervalles de confiance pour les $\widehat{\rho}(h)$ et $\widehat{r}(h)$, on vérifie que « les pics ne sont pas significatifs ».

Remarque 2 : si « vrai » ARMA(p,q), comme on se limitera à $p,q\leqslant 2$, on teste les différents modèles...
(sinon méthodes plus sophistiquées, cf *méthode du coin*.)

Estimation des paramètres θ_j, ϕ_i, σ (1)

Définition **: vraisemblance de la réalisation d'un v.a. X**

Soit f_θ la densité de X, et x une réalisation de X :

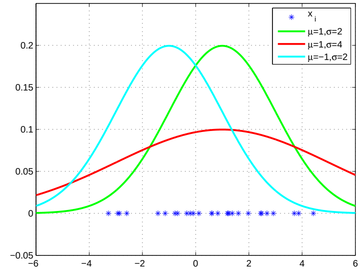
$$\mathcal{L}(x;\theta)=f_\theta(x)$$

(θ : paramètres de la densité d'un vecteur aléatoire)

Méthode du **maximum de vraisemblance** :
estimer θ qui maximise $\mathcal{L}(x;\theta)$, connaissant x.
(*idée* : plus la vraisemblance est grande, plus il est probable que le modèle pour les (X_t) soit bon)

Exemple : $x_1,x_2,\dots,x_n$ réalisations i.i.d. gaussiennes :

$$\mathcal{L}(x_1,x_2,\dots,x_n;\mu,\sigma)=\frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$



Estimation des paramètres θ_j, ϕ_i, σ (2)

$$\forall t,\ X_t=\theta_0+\sum_{i=1}^{p}\phi_i X_{t-i}+\varepsilon_t-\sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$$

Hypothèse : (ε_t) i.i.d. $\sim\mathcal{N}(0,\sigma^2)$.
Donc $(X_t)_{0\leqslant t\leqslant T}$ est un vecteur gaussien.
(car combinaison linéaire des (ε_t))

Vraisemblance des T observations X_t :

$$\mathcal{L}((X_t)|(\phi_i),(\theta_j),\sigma)=\frac{1}{(2\pi)^{T/2}\sqrt{\det\Omega}}\exp\left(-\frac{1}{2}X'\Omega^{-1}X\right)$$

où Ω dépend des ϕ_i, θ_j, σ ; $X=(X_t)_{1\leqslant t\leqslant T}$

$\rightarrow$ Estimation des paramètres $(\phi_i),(\theta_j),\sigma$ par la méthode (numérique) du *maximum de vraisemblance*.

Remarque : il existe d'autres méthodes d'estimation.
$\rightarrow$ *à utiliser lorsque SAS avertit d'un problème d'optimisation numérique.*

Examen de la validité du modèle

1. Significativité des paramètres estimés (Student)

2. Résidus = **bruit blanc gaussien** ?
Estimation e_t des résidus : écart entre observation X_t et prévision $\widehat{X}_{t-1}(1)$ (cf *prévision*)
$\rightarrow$ Vérifier que les résidus (e_t) sont bien des « réalisations » d'un bruit blanc gaussien (ε_t) :

- vérification graphique : moyenne nulle, variance constante, pas de « structure ».
- ACF et PACF **des résidus** ne doivent pas avoir de pics significatifs.
- *Portmanteau Test* (Box-Pierce) **sur les résidus** : soit $\rho(h)$ l'ACF des résidus.
  Alors : $(T-K)\sum_{h=1}^{K}\rho(h)^2$ suit un χ^2 à $K-p-q$ d.d.l., sous hypothèse (ε_t) bruit blanc gaussien.
  Mieux : Ljung-Box.
  (sous SAS, généralement $K=6, K=12\dots$)

Sélection de modèle

Question : comment choisir entre différents modèles ARIMA(p,d,q) *valides* ?

Plusieurs possibilités :

- Minimisation de la variance des résidus σ^2.
- Minimisation de *critères d'information* :
  $AIC = -2\log(\mathcal{L}) + 2(p+q)$
  ou $BIC = SBC = -2\log(\mathcal{L}) + (p+q)\log(T)$.
  (compromis entre vraisemblance et nombre de paramètres)

Remarque 1 : « le » choix entre différents modèles se fait selon **un** critère.

Remarque 2 : on préfère les modèles simples (principe du *rasoir d'Ockham* ou de *parcimonie*).

Prévision avec modèle ARIMA(p,d,q)

→ la semaine prochaine.

Séance 2

La procédure `ARIMA`

Identification d'un modèle ARIMA(p,d,q) :

```
proc arima data=serie;
identify var=variable;
run;
identify var=variable(1);
run;
estimate p=1 plot;
run;
estimate p=1 q=1 plot;
run;
forecast out=sortie back=10 lead=10;
run;
```

→ `identify` : identification des ordres p et q avec ACF et PACF de la chronique, éventuellement différentiée (ordre d).

→ `estimate` : estimation des paramètres et significativité, ACF et PACF des résidus, portmanteau.

→ `forecast` : prévisions.

Écriture des différents modèles : `estimate`

`estimate p=3;`

$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t$
ARIMA$(3, d, 0)$

`estimate p=3 noconstant;`

$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t$
sous-modèle de ARIMA$(3, d, 0)$

`estimate p=1 q=2;`

$Y_t = \theta_0 + \phi_1 Y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$
ARIMA$(1, d, 2)$

`estimate p=2 q=(2);`

$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t - \theta_2 \varepsilon_{t-2}$
sous-modèle de ARIMA$(2, d, 2)$

`estimate q=(1 3);`

$Y_t = \theta_0 - \theta_1 \varepsilon_{t-1} - \theta_3 \varepsilon_{t-3}$
sous-modèle de ARIMA$(0, d, 3)$

`estimate p=(12) q=2;`

$Y_t = \theta_0 + \phi_{12} Y_{t-12} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$
SARIMA$(0, d, 2)(1, d', 0)_{12}$

Séance 2

Conclusion

- Processus stationnaire ?
  - AR « pur » (sans racine unité),
  - MA « pur »,
  - ARMA.

  $\rightarrow$ identifications des ordres p / q : ACF et PACF.

- Processus non-stationnaire ?
  $\rightarrow$ on stationnarise d'abord en transformant la variable et / ou dérivant (ARIMA / SARIMA)

- Il faut aussi valider le modèle identifié
  (significativité des coefficients, résidus b.b.g., etc).

$\longrightarrow$ à pratiquer en TD !