Cours GIMAS8AD

Modélisation et prévision

Séries chronologiques - Séance 3
Compléments sur ARIMA, processus SARIMA

Frédéric Sur
École des Mines de Nancy

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

Séance 3

Modélisation des chroniques par (S)ARIMA

- soit la chronique est stationnaire
  (propriétés statistiques invariantes au cours du temps)

  → modélisation AR / MA / ARMA.

- soit elle ne l'est pas
  (tendance stochastique / marche aléatoire, ou tendance déterministe)

  → on commence par dériver pour stationnariser.

Méthode de Box-Jenkins

`proc arima`

1. **Transformation** (éventuelle) de la chronique (généralement log) pour stabiliser la variance.
2. **Identification** des paramètres p, d, q.
   → `identify` : identification des ordres p et q avec ACF et PACF de la chronique, éventuellement différentiée à l'ordre d au préalable.
3. **Estimation** des θ_j, ϕ_i, μ (ou `constant`) et σ.
   → `estimate` : estimation des paramètres.
4. **Validation** du modèle.
   → `estimate` : significativité, ACF, PACF et graphe des résidus, Portmanteau, AIC, SBC, σ.
5. **Prévision** du futur.
   → `forecast` : prévisions.

Prévision avec modèle ARIMA(p,d,q) (1)

Rappel : processus ARIMA

$$\Phi(B)\left((1-B)^d X_t - \mu\right) = \Theta(B)\varepsilon_t$$

ou :

$$\Psi(B)X_t = \theta_0 + \Theta(B)\varepsilon_t$$

c'est-à-dire *(pour simplifier, on suppose $\theta_0 = 0$)* :

$$X_t = \Psi_1 X_{t-1} + \ldots + \Psi_{p+d} X_{t-p-d} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}$$

Prévision : on connaît (X_t) jusque la date $t = T$,
on cherche une prévision $\widehat{X}_T(h)$ de X à un *horizon* de h après l'instant T.

Prévision avec modèle ARIMA(p,d,q) (2)

On écrit X_{T+h} sous la forme :

$$X_{T+h} = \sum_{i=1}^{p+d} \psi_i X_{T+h-i} + \varepsilon_{T+h} - \sum_{j=1}^{q} \theta_j \varepsilon_{T+h-j} \quad (*)$$

On définit $\forall h,\ \widehat{X}_T(h) = \mathbb{E}(X_{T+h}\,|\,(X_t)_{t\leqslant T})$.
(≪ meilleure ≫ approximation de X_{T+h} par comb. lin. des $(X_t)_{t\leqslant T}$)

On remarque : $\widehat{X}_T(h) = X_{T+h}$ si $h \leqslant 0$. *(heureusement...)*

Prévision avec modèle ARIMA(p,d,q) (3)

$$X_{T+h} = \sum_{i=1}^{p+d} \psi_i X_{T+h-i} + \varepsilon_{T+h} - \sum_{j=1}^{q} \theta_j \varepsilon_{T+h-j} \quad (*)$$

Donc :

$$\widehat{X}_T(h) = \sum_{i=1}^{p+d} \psi_i \widehat{X}_T(h-i) - \sum_{j=h}^{q} \theta_j \varepsilon_{T+h-j} \quad (**)$$

car $\mathbb{E}(\varepsilon_{T+h-j}|(X_t)_{t\leqslant T}) = \varepsilon_{T+h-j}$ si $j \geqslant h$ et $= 0$ sinon.
($\varepsilon_{T+h-j} \in \text{Vect}(X_t)_{t\leqslant T}$ si $j \geqslant h$ et (ε_t) non corrélés).

Conséquence 1 : avec $(*)$ et $(**)$, $\forall t,\ X_{t+1} - \widehat{X}_t(1) = \varepsilon_{t+1}$.

Conséquence 2 : formules d'actualisation avec les ψ_i, θ_j, ε_t (estimés) :

$$\widehat{X}_T(1) = \sum_{i=1}^{p+d} \psi_i X_{T+1-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{T+1-j}$$

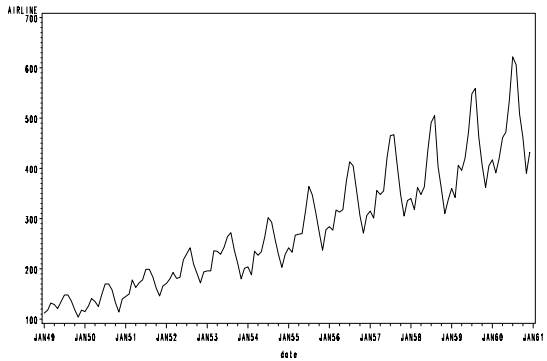
$$\widehat{X}_T(2) = \psi_1 \widehat{X}_T(1) + \sum_{i=2}^{p+d} \psi_i X_{T+2-i} - \sum_{j=2}^{q} \theta_j \varepsilon_{T+2-j}$$

...

SAS : I.C. pour $\widehat{X}_{T+h}$ sous hypothèse de normalité des ε_t.

Remarque : transformation de (X_t)...

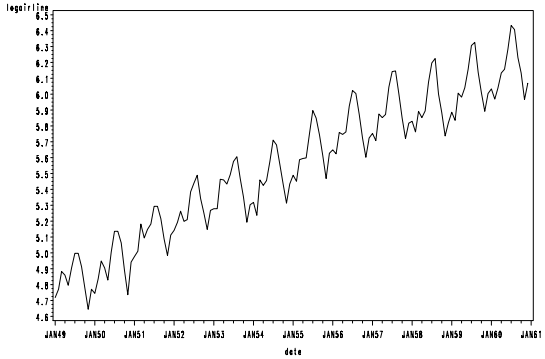
Exemple : chronique `airline`



→ passage au log...

Passage au log

Exemple : logarithme de la chronique `airline`



$\rightarrow$ modèle additif avec tendance, variance de la composante aléatoire stabilisée.

Question : *quid* de la prévision sur `airline` ?

Étude de $Y_t = \log(X_t)$

Hypothèse : $Y_\tau \sim \mathcal{N}(\widehat{Y_\tau}, \sigma_\tau^2) \quad (\tau > T)$

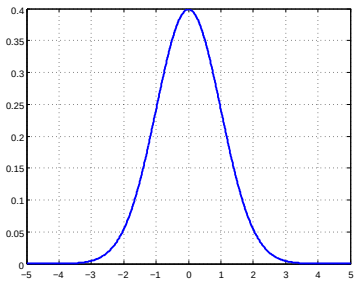
`forecast` : prévision $\widehat{Y}_\tau$ + int. de conf. $[L_\tau, U_\tau]$ à 95% (centré sur $\widehat{Y_\tau}$).

Question : intervalle et prévision pour $X_\tau = \exp(Y_\tau)$?

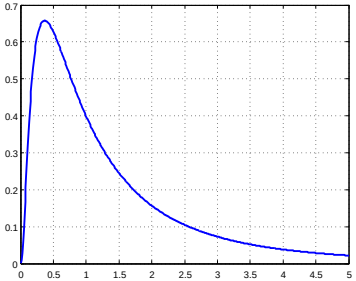
- Comme exp est croissante :
  $\Pr(\exp(Y_\tau) \in [\exp(L_\tau), \exp(U_\tau)]) \leqslant 95\%$.
  Donc ≪ intervalle de confiance ≫ à 95% :
  $[\exp(L_\tau), \exp(U_\tau)]$.
- Prévision ?
  Naïf : $\widehat{X}_\tau = \exp(\widehat{Y}_\tau) \quad (= \exp(\mathbb{E}(Y_\tau)) \neq \mathbb{E}(\exp(Y_\tau)))$
  Mieux : $\widehat{X}_\tau = \mathbb{E}(X_\tau) = \exp(\widehat{Y}_\tau + \sigma_\tau^2/2)$
  car X_τ suit une *loi log-normale*.
- *Remarque* : ≪ intervalle de confiance ≫ non centré sur $\mathbb{E}(X_\tau)\ldots$

Illustration : $Y_t = \log(X_t)$

$\widehat{Y}_\tau = 0, \quad \sigma_\tau = 1.$



loi de Y_τ (normale)



loi de X_τ (log-normale)

Ici : $\widehat{Y}_\tau = 0, \quad$ I.C. 95% : $[-1.96, 1.96]$.
Prévision sur X_τ : I.C. 95% : $[0.14, 7.1]$
$\exp(\widehat{Y}_\tau) = 1$
$\widehat{X}_\tau = \exp(\widehat{Y}_\tau + \sigma_\tau^2/2) = 1.6$

Remarque : bien sûr, correction négligeable si $\sigma_\tau^2/2 << \widehat{Y}_\tau$

Séance 3

Cas des chroniques périodiques (période π)

Remarque 1 : X_t peut ne pas être stationnaire à cause d'un comportement du type :

$$X_t = S_t + u_t \qquad (S_t \text{ déterministe } \pi\text{-périodique})$$

ou

$$X_t = X_{t-\pi} + u_t \qquad (\text{cf marche aléatoire})$$

$\rightarrow$ on peut stationnariser en étudiant $(1 - B^\pi)X_t$.

Remarque 2 : des corrélations saisonnières (période π) peuvent être présentes dans la chronique X_t
(corrélations / corrélations partielles aux décalages de π, 2π, $3\pi\ldots$)

$\rightarrow$ ARMA *saisonnier* :

$$\phi(B^\pi)X_t = \theta(B^\pi)\varepsilon_t$$

Les processus SARIMA

Définition : processus SARIMA$(p, d, q)(P, D, Q)_\pi$

Ce sont les processus (X_t) du type :

$$\Phi_p(B)\Phi_P(B^\pi)(1-B)^d(1-B^\pi)^D X_t = \theta_0 + \Theta_q(B)\Theta_Q(B^\pi)\varepsilon_t$$

où $p, d, q, P, D, Q \geqslant 0$, π est la période de la saisonnalité et (ε_t) est un bruit blanc gaussien.

Intérêt : traiter les chroniques non-stationnaires, avec tendance et saisonnalité ou comportement style ≪ marche aléatoire ≫.

Remarque : SARIMA = ARIMA particulier, mais la factorisation limite le nombre de coefficients à estimer.
(cf *parcimonie*, *rasoir d'Ockham*)

Exemples

1. processus SARIMA$(1, 0, 2)(1, 1, 0)_4$:
(chronique trimestrielle, période annuelle)

$$(1-\phi_1 B)(1-\phi'_1 B^4)(1-B^4)X_t = \theta_0+(1-\theta_1 B-\theta_2 B^2)\varepsilon_t$$

ou

$$(1-\phi_1 B)(1-\phi'_1 B^4)\left((1-B^4)X_t - \mu\right) = (1-\theta_1 B-\theta_2 B^2)\varepsilon_t$$

2. processus SARIMA$(0, 1, 1)(1, 0, 0)_{12}$:
(chronique mensuelle, période annuelle)

$$(1-\phi'_1 B^{12})(1-B)X_t = \theta_0 + (1-\theta_1 B)\varepsilon_t$$

ou

$$(1-\phi'_1 B^{12})\left((1-B)X_t - \mu\right) = (1-\theta_1 B)\varepsilon_t$$

Identification des modèles

TP précédent : identification des ordres des processus ARIMA selon ACF et PACF.

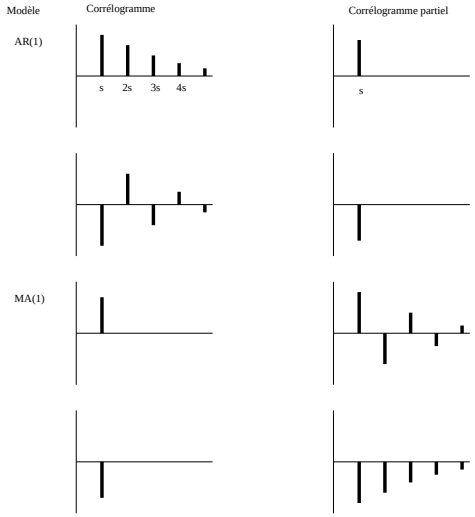
TP aujourd'hui : identification des ordres des processus SARIMA : regarder aussi les pics 12 et 24 de l'ACF et du PACF (pour saisonnalité de période $\pi = 12$).

Important : on cherche des modèles *simples*...

Remarque : on commence par regarder ACF/PACF pour ≪ petits décalages ≫ ($h \leqslant 6$), puis pour $h = 12, 24, 36$.
(pour se débarrasser de l'influence des corrélations ≪ court termes ≫ sur la composante saisonnière)

En effet : si par exemple $X_t = (1 - \theta_1 B)(1 - \theta'_1 B^{12})\varepsilon_t$
alors : $X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta'_1\varepsilon_{t-12} + \theta_1\theta'_1\varepsilon_{t-13}$
(influence des corrélations ≪ court termes ≫ sur le ≪ long terme ≫)

ACF et PACF pour (S)AR(1) et (S)MA(1)



Séance 3

Quelques ≪ règles ≫ complémentaires : différentiation (1)

- ordre de différentiation saisonnière : 0 ou 1.
- ordre de différentiation totale (saisonnière & non saisonnière) : $d + D \leqslant 2$.
- si la décroissance de l'ACF est lente, penser à différentier plutôt qu'introduire un AR.
  (cf chronique `magnesium`)
- si l'ACF est périodique (effet ≪ pont suspendu ≫), alors différentiation saisonnière.

Quelques ≪ règles ≫ complémentaires : différentiation (2)

- si ACF pour décalage 1 est négatif (≪ ≤ -0.5 ≫), la chronique est trop différentiée.

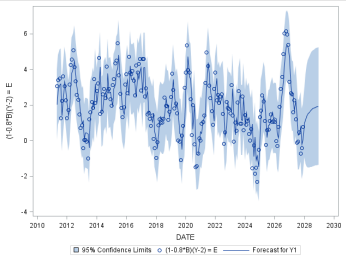
  $\rightarrow$ enlever un ordre de dérivation plutôt qu'introduire un MA.

  ≪ *Justification* ≫ :
  si X_t déjà stationnaire et $Y_t = (1 - B)X_t$, alors :
  $\gamma_Y(1) = 2\gamma_X(1) - \gamma_X(0) - \gamma_X(2)$
  $\gamma_Y(0) = 2\gamma_X(0) - 2\gamma_X(1)$
  donc $\rho_Y(1) = -\frac{1}{2}\left(\frac{1-2\rho_X(1)+\rho_X(2)}{1-\rho_X(1)}\right)$

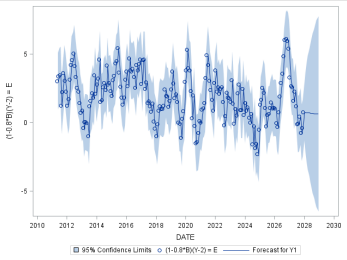
Pourquoi éviter de trop différentier ?

Cf intervalles de confiance de la prévision...

Exemple : chronique Y_1 (exercice 1 séance 2)



$(1 - 0.805B)Y_t = 0.3939 + \varepsilon_t$
$(\sigma = 1.002)$



$(1 - B)Y_t = -0.011 + \varepsilon_t$
$(\sigma = 1.052)$

Quelques ≪ règles ≫ complémentaires : la constante

$$\Phi_p(B)\Phi_P(B^\pi)(1-B)^d(1-B^\pi)^D X_t = \underline{\theta_0} + \Theta_q(B)\Theta_Q(B^\pi)\varepsilon_t$$

- chronique différentiée à l'ordre 1 :
  constante = pente de la tendance.
  On peut avoir une constante nulle
  (ex : marche aléatoire)
  ou pas (ex : $ax + b$)
- chronique différentiée à l'ordre 2 (la pente ≪ varie ≫) :
  constante = coef du terme quadratique.
  (tendance quadratique rare, donc *constante nulle*)

Quelques ≪ règles ≫ complémentaires : divers

- éviter de mélanger SAR et SMA.
- les termes en AR et MA peuvent se compenser.

  Ex : si ARIMA(2,d,1) identifié,
  on peut essayer ARIMA(1,d,0)
  (cas où les racines de AR et MA se ≪ compensent ≫).

Séance 3

Tous les modèles sont faux . . .
mais certains sont utiles !

"Remember that all models are wrong ; the practical question is how wrong do they have to be to not be useful."

George E. P. Box, Norman R. Draper
Empirical Model-Building and Response Surface,
Wiley, 1987.

Exemple : chronique SNCF

Exemple du polycopié, sous SAS. . .