3b - Méthode X13-ARIMA: le modèle Reg-ARIMA

Anna Smyk et Tanguy Barthélémy
Division Recueil et Traitement de l'Information
Département des Méthodes Statistiques

Objectifs de cette séquence

Connaître la partie pré-ajustement (modélisation Reg-Arima) de la méthode X13-ARIMA.

Après cette séquence, vous :

- connaîtrez les objectifs de l'utilisation d'un modèle Reg-ARIMA et sa structure
- pourrez identifier les modèles utilisés par JDemetra+
- pourrez modifier les spécifications du modèle

X13-ARIMA

Deux modules :

- X11 : phase de décomposition (T S I)
- Reg-ARIMA : phase de pré-ajustement
  - Régression linéaire pour correction préalable des « non-linéarités »
  - Modélisation ARIMA pour faire des prévisions

Sommaire

Linéariser la série

On décrit ici ce que fait l'algorithme AVANT la décomposition

Objectif : supprimer les effets déterministes par régression linéaire :

- outliers
- effets de calendrier (séquence spécifique)

$$Y_t = \sum \hat{\alpha}_i O_{it} + \sum \hat{\beta}_j C_{jt} + X_t$$

Série *linéarisée* : $X_t = Y_t - \sum \hat{\alpha}_i O_{it} - \sum \hat{\beta}_j C_{jt}$

le GROS résidu de la régression est modélisé par un modèle ARIMA

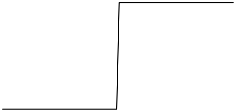
(Ici les $\hat{\alpha}_i$ et les $\hat{\beta}_j$ ne pourraient être estimés par moindres carrés ordinaires (MCO) car le résidu est autocorrélé dans les séries temporelles.
$\rightarrow$ le modèle ARIMA modélise justement cette autocorrélation et permet une estimation de type moindres carrés généralisés (MCG).)

La décomposition est réalisée sur la série linéarisée, prise en entrée de X-11.

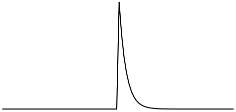
Les principaux types d'outliers

Choc ponctuel
Additive outlier (AO)
Alloué in fine, après la décomposition, à l'Irrégulier

Changement de niveau
Level Shift (LS)
Alloué in fine, après la décomposition, à la Tendance

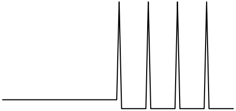


Changement de niveau transitoire
Transitory Change (TC)
Alloué in fine, après la décomposition, à l'Irrégulier



Seasonal Outlier

Rupture de profil saisonnier
Seasonal Outlier (SO)



Très rarement utilisé, non detecté automatiquement par défaut dans JD+
Affecte directement la Composante Saisonnière

La modélisation Reg-ARIMA

Le modèle Reg-ARIMA s'écrit comme suit :

$$\left(Y_t - \sum \alpha_i X_{it}\right) \sim ARIMA(p,d,q)(P,D,Q)$$

où les variables (regresseurs) X_i représentent les effets de calendrier et les outliers

Le modèle ARIMA décrit la structure d'autocorrélation du modèle de régression : il capture toute l'information temporelle. Le résidu du modèle Reg-ARIMA est un bruit blanc.

Le modèle ARIMA est utilisé pour :

- permettre l'estimation du modèle de régression (on sort des MCO car autocorrélations)
- prévoir la série linéarisée, afin que des moyennes mobiles symétriques puissent être utilisées par X-11 jusqu'au dernier point brut disponible.

Nous allons voir plus en détail la pertinence et la composition de ces modèles.

Sommaire

Stationnarité

La notion de **stationnarité** est ici importante car c'est une propriété nécessaire pour le type de modélisation que nous allons décrire.

La série est dite (faiblement) stationnaire lorsque :

- les moments d'ordre 1 et 2 ne dépendent pas du temps

$E(X_t) = m,\ V(X_t) = \sigma^2$

- la covariance entre t et $t - h$ ne dépend pas du temps, mais de la distance h : $cov(X_t, X_{t+h}) = \gamma(h)$

En pratique : regardée « par fenêtre mobile » la série a toujours « la même allure ».

Exemple fondamental : un bruit blanc, noté usuellement ε_t

- espérance nulle $E(\varepsilon_t) = 0$
- variance non nulle et constante $\forall t, V(\varepsilon_t) = \sigma^2$
- covariance entre t et t' nulle : $cov(\varepsilon_t, \varepsilon_{t'}) = 0$

Sommaire

Modèles Autorégressifs (AR)

Un modèle *ARMA* (on verra le sens du *I* enlevé) comporte une partie Auto Regressive *AR* et une partie Moving Average (*MA*)

Pour écrire ces modèles, on introduit un opérateur retard souvent noté B (Backwards) ou L (Lag) :

B opérateur retard : $B(X_t) = X_{t-1}$, et $B^p(X_t) = X_{t-p}$

Modèle *autorégressif* d'ordre p, $AR(p)$:

$$
\begin{aligned}
& X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \\
\Longleftrightarrow\ & (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t = \varepsilon_t \\
\Longleftrightarrow\ & \Phi(B) X_t = \varepsilon_t
\end{aligned}
$$

ε_t *innovation* du processus (bruit blanc indépendant du passé de X)

ε_t n'est pas corrélé aux $X_1, X_2......., X_{t-1}$

Un $AR(p)$ modélise l'influence de p réalisations passées sur la réalisation courante : effet mémoire.

Modèles « Moving Average » (MA)

Modèle *moyenne mobile* d'ordre q, $MA(q)$:

$$\begin{aligned} & X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \\ \iff & X_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \varepsilon_t \\ \iff & X_t = \Theta(B) \varepsilon_t \end{aligned}$$

Processus MA toujours stationnaire

Résulte d'une accumulation non persistante de “q” chocs indépendants.

On décrit les phénomènes qui fluctuent autour d'une moyenne par un MA(1) avec une constante.

Modèles ARMA

Modèles $ARMA(p,q)$: combine $AR(p)$ et $MA(q)$, sans ou avec constante

$$\Phi(B)X_t = \Theta(B)\varepsilon_t$$

$$\Phi(B)X_t = \mu + \Theta(B)\varepsilon_t$$

Un processus $ARMA$ résulte d'un effet "mémoire" et d'une accumulation non persistante de chocs aléatoires indépendants.

Stationnarité et modélisation ARMA

Les séries stationnaires peuvent, en théorie, être modélisées par un modèle *ARMA*.

Théorème de Wold :

Sous hypothèse de stationnarité (faible)

- il existe un « modèle ARMA » qui approche la série.
- les erreurs de prévision se comportent comme le résidu du modèle (bruit blanc, non corrélé avec le passé des observations)

Modèles SARMA : écriture adaptée aux séries saisonnières des modèles ARMA (1/2)

Pour modéliser des séries saisonnières, on écrit des modèles **S**ARMA qui permettent de mettre en évidence les retards d'ordre égal à la périodicité s.

Aucun ajout conceptuel par rapport aux modèles *ARMA* mais plus grande clarté d'écriture.

Modèle $SARMA(P, Q)$: *ARMA* avec polynôme d'ordre s (4 pour les séries trimestrielles, 12 pour les séries mensuelles) :

$$\Phi(B^s)X_t = \Theta(B^s)\varepsilon_t \text{ ou } \Phi_s(B)X_t = \Theta_s(B)\varepsilon_t$$

Intérêt :

- mettre en relief les autocorrélations d'ordre s
- simplifier l'écriture par factorisation

Modèles SARMA : écriture adaptée aux séries saisonnières des modèles ARMA (2/2)

Un $ARMA(p,q)(P,Q)$ combine parties régulière et saisonnière : $ARMA(p,q) \times SARMA(P,Q)$.

Identique à un $ARMA(p+P*s, q+Q*s)$

Exemple série mensuelle : $ARMA(1,1)(1,1) = ARMA(13,13)$

Mais l'écriture $ARMA(13,13)$ est plus ambiguë que $ARMA(1,1)(1,1)$ à moins de préciser que les retards de 2 à 11 ont des coefficients nuls.

Traitement des processus non stationnaires : Modèles dits **I**ntégrés

Seul un processus stationnaire est modélisable avec un modèle ARMA. Or les séries temporelles couramment rencontrées, notamment les indicateurs économiques, sont rarement sationnaires. On va identifier l'origine de la non-stationnarité et voir comment **stationnariser le processus**.

Soit X un processus « tendance linéaire » :

$$X_t = \alpha + \beta t + \varepsilon_t$$

Pour voir si une série est stationnaire, on calcule l'espérance puis si nécessaire la variance et les autocovariances. (NB. La variance est l'autocovariance avec un retard $k = 0$.)

Pour une tendance linéraire (déterministe) + bruit blanc (aléatoire)

$$E(X_t) = \alpha + \beta t + E(\varepsilon_t) = \alpha + \beta t$$

L'espérance dépend du temps, la série n'est pas stationnaire.

Tendance linéaire et différenciation

Calculons la variance.

$$V(X_t) = V(\varepsilon_t) = \sigma^2$$

La variance, elle, ne dépend pas du temps.

On va **différencier** cette série

Différence d'ordre 1 :

$$(I - B)X_t = X_t - X_{t-1} = \alpha + \beta t + \varepsilon_t - \alpha - \beta(t-1) - \varepsilon_{t-1}$$

soit

$$(I - B)X_t = \beta + \varepsilon_t - \varepsilon_{t-1}$$

Stationnarisation par différenciation

Calculons l'espérance, la variance et les covariances de la série différenciée.

$$E((I-B)X_t) = \beta$$

$\rightarrow$ l'espérance ne dépend pas du temps

$$V((I-B)X_t) = V(\varepsilon_t) + V(\varepsilon_{t-1}) = 2\sigma^2$$

car les ε_t et ε_{t-1} ne sont pas corrélés

Calculons la covariance entre t et $t+h$, quand $h>0$:
$cov((I-B)X_t, (I-B)X_{t+h}) = cov(\beta + \varepsilon_t - \varepsilon_{t-1}, \beta + \varepsilon_{t+h} - \varepsilon_{t+h-1}) = 0$
car $\forall t \neq t' \ : \ cov(\varepsilon_t, \varepsilon_{t'}) = 0$

$\rightarrow$ la différenciation (ici d'ordre 1) a permis de stationnariser la série.

Ordres de différenciation

Si X est un processus « tendance polynomiale d'ordre 2 », différencier deux fois va stationnariser la série : $(I-B)^2X_t$ est stationnaire.

De manière générale $(I-B)^d$ stationnarisera un polynôme d'ordre d.

$\rightarrow$ d est appelé ordre de la différenciation.

Un modèle AR**I**MA aura donc 3 ordres : (p, d, q).

L'ordre et le choix des lettres sont conventionnels : AR d'ordre p et MA d'ordre q.

Stationnarisation d'un processus saisonnier stable

Soit X, processus « saisonnier stable » :

$$X_t = S_t + \varepsilon_t \quad \text{avec} \quad \forall t,\, S_t = S_{t+s} \quad \text{et} \quad \varepsilon_t \text{ un bruit blanc}$$

Le processus X_t est-il stationnaire ?

Calculons son espérance : $E(X_t) = S_t$.

On aura $E(X_t) = E(X_{t+12})$ mais $E(X_t) \neq E(X_{t+11})$, donc le processus n'est pas stationnaire.

$V(X_t) = E((X_t - E(X_t))^2) = E((S_t + \varepsilon_t - S_t - 0)^2) = E(\varepsilon_t^2) = \sigma^2$

Différenciation saisonnière

Pour stationnariser X_t on va appliquer une **différenciation saisonnière** d'ordre 1 :

$$(I - B^s)X_t = S_t + \varepsilon_t - S_{t-12} - \varepsilon_{t-12} = \varepsilon_t - \varepsilon_{t-12}$$

On obtient une combinaison linéaire de bruits blancs, qui est donc stationnaire.

Si X comportait en plus une tendance linéaire $X_t = S_t + a \times t + b + \varepsilon_t$:

$$(I-B^s)X_t = S_t+a\times t+b+\varepsilon_t-S_{t-12}-a\times(t-12)-b-\varepsilon_{t-12} = 12\times a+\varepsilon_t-\varepsilon_{t-12}$$

On voit que même avec l'ajout d'une tendance linéaire, la différenciation saisonnière stationnariserait la série.

On remarque que la différenciation simple est comprise dans la différenciation saisonnière :

$$(I - B^s)X_t = (I - B)(I + B + ... + B^{s-1})X_t$$

Ordre de différenciation et tendances polynomiales

De manière générale, une différenciation « simple » d'ordre d supprime les tendances polynomiales d'ordre d :

$$(I - B)^d X_t$$

Une différenciation « saisonnière » supprime aussi les tendances linéaires :

$$(I - B^s) X_t$$

Dans JD+, $D \leq 1$ et $d \leq 2$ dans X-13-ARIMA.

Modèles ARIMA

$ARIMA(p, d, q)$ modélise les séries non saisonnières avec tendance :

$$\Phi(B)(I - B)^d X_t = \Theta(B)\varepsilon_t$$

En mettant en évidence les polynômes en B^s $ARIMA(p, d, q)(P, D, Q)$ modélise les séries avec tendance et saisonnalité :

$$\Phi(B)\Phi_s(B)(I - B)^d (I - B^s)^D X_t = \Theta(B)\Theta_s(B)\varepsilon_t$$

Modèles ARIMA et saisonnalité (1/3)

Considérons la partie saisonnière d'un ARIMA :

1 - Une série avec modèle $(p, d, q)(0, 0, 0)$ est-elle saisonnière ?

2 - que dire de $(p, d, q)(0, 0, Q)$?

3 - $(p, d, q)(0, 1, 0)$?

Modèles ARIMA et saisonnalité (2/3)

Réponses :

1 - Non, aucune autocorrélation d'ordre s.

2 - Non, un MA reflète des fluctuations non persistantes, la saisonnalité persiste dans le temps.

3 - Oui, une saisonnalité stable.

Modèles ARIMA et saisonnalité (3/3)

Deux cas fréquents :

- $(p, d, q)(0, 1, 1)$ saisonnalité stable en moyenne, avec des fluctuations ponctuelles du niveau de θ_s (plus θ_s est grand, plus ça fluctue)
- $(p, d, q)(1, 1, 1)$ saisonnalité évolutive avec dérive + fluctuations ponctuelles de niveau θ_s

Détermination du modèle ARIMA dans X13

- Première estimation complète (coefficients de la régression et coefficients du modèle Arima) avec le modèle ARIMA le plus fréquent (0,1,1)(0,1,1), appelé *Airline* car il a historiquement servi à modéliser des flux de passagers dans le transport aérien
- Recherche des modèles concurrents à comparer à ce modèle par défaut (avec des critères d'information type AIC, BIC)

Sommaire

Les hypothèses sur les résidus

Les estimateurs des coefficients du modèle regARIMA sont sans biais, convergents et de variance minimale, si les résidus ε_t sont :

- *de moyenne nulle* (souvent par construction des estimateurs)
- *décorrélés* : $\forall t \neq t' \;:\; cov(\varepsilon_t, \varepsilon_{t'}) = 0$ (*autocorrélés* sinon)
- *homoscédastiques* : $\forall t, t' \;:\; \mathbb{V}\left[\varepsilon_t\right] = \mathbb{V}\left[\varepsilon_{t'}\right]$ (*hétéroscédastiques* sinon)

On dit que les résidus sont un bruit blanc : il ne contiennent plus d'information, ce qui est l'objectif recherché.

S'ils sont distribués selon une *loi normale*, on peut faire des tests sur la validité des paramètres et calculer des intervalles de confiance.

Causes possibles de la violation des hypothèses

L'autocorrélation des résidus peut provenir de :

- variable omise : il manque une variable explicative importante (ici regresseurs de calendrier, outliers)
- interpolation, lissage
- erreurs de mesure

L'autocorrélation peut souvent être corrigée en :

- modifiant les regresseurs, notamment les points atypiques
- augmentant l'ordre de AR (attention à la parcimonie et à la stabilité du modèle)

L'hétéroscedasticité est liée à la nature des variables explicatives il faut également modifier les regresseurs et les points atypiques.

Dans ces deux cas les estimateurs sont sans biais, peuvent être convergents mais ne sont plus de variance minimale.

Traiter en priorité l'autocorrélation.

Sommaire

Les essentiels

Les séries économiques ne sont pas stationnaires : ni leur niveau, ni leurs fluctuations ne sont constants dans le temps

Différencier un processus permet de le stationnariser.

Un MA capte les fluctuations non persistantes autour d'un niveau constant : c'est donc un processus stationnaire.

Un AR met en évidence l'influence des réalisations passées sur la réalisation courante.

Un ARIMA reflète la structure des autocorrélations de la série, ainsi que son degré de variabilité dans le temps.

L'examen des résidus permet de valider les modèles. Le choix entre plusieurs modèles valides se fait grâce aux critères d'information.

La modélisation ARIMA du module de pré-ajustement permet de linéariser et de prolonger la série. Elle ne permet pas d'isoler la composante saisonnière.

Exercice

Ecrire les modèles Reg-ARIMA de vos séries à partir des éléments donnés par JDemetra+.