TP 3 & 4 : Décomposition d'une série chronologique

DESS d'Ingénierie Mathématique, Année 2003-2004
Module de Séries chronologiques
laurent.albera@fr.thalesgroup.com

I. Introduction

Ce TP s'effectuera sous MATLAB. En s'appuyant sur la méthodologie développée en cours, on programmera divers algorithmes de décomposition d'une série chronologique en sa partie déterministe et sa partie aléatoire. Ce prétraitement s'avère nécessaire, car une fois la partie déterministe de la série identifiée, on peut alors chercher à expliquer la fluctuation résiduelle, notamment en la modélisant par un processsus ARMA (ou autre) dont les paramètres sont à estimer. Une des motivations de ces traitements est la prédiction de données futures uniquement à partir des données passées.

II. Hypotheses et notations

Nous supposons dans la section V du TP que les observations x_i de la série chronologique considérée suivent un modèle additif, c'est-à-dire vérifiant :

$$x_i = g_i + s_i + w_i \tag{1}$$

où g_i, s_i et w_i sont respectivement la tendance, le facteur saisonnier et la fluctuation résiduelle à l'instant i. Le facteur saisonnier est supposé de moyenne nulle et de période p. Quant à la fluctuation résiduelle, elle est également à moyenne nulle.

Puis nous considérerons dans la section VI que les données vérifient le modèle suivant :

$$x_i = a \exp\{\mathrm{j}\,\omega\, i\} + w_i \tag{2}$$

où a est un nombre complexe et ω la pulsation dite *réduite*.

III. Representation graphique des donnees

A. Données

La première chose à faire consiste à représenter graphiquement les données afin de repérer sans prétraitement numérique les éventuels tendances et/ou facteurs saisonniers présents dans la série. Télécharger les fichiers *airline.dat*, *copper.dat*, *games.dat* et *oil.dat*. Tracer les courbes correspondantes et commenter.

B. Fonctions utiles

Utiliser sous MATLAB les fonctions suivantes : *clear*, *load('·')*, *figure(·)*, *plot(·)*, *title('·')* et *grid*. La commande *help plot* permet de connaître, à titre d'exemple, la fonction et la syntaxe de la commande *plot(·)*.

IV. Utilisation de la fonction d'autocorrelation

A. Outils

Soit $\rho_e(h)$ la fonction d'autocorrélation empirique associée à la série chronologique $\{x_i\}_{(1\leq i\leq n)}$, supposée stationnaire et ergodique :

$$\forall\ 0 \leq h \leq n-1, \quad \rho_e(h) = \frac{n^{-1}\sum_{i=1}^{n-h}(x_i - m_e)(x_{i+h} - m_e)}{n^{-1}\sum_{i=1}^{n}(x_i - m_e)^2} \tag{3}$$

où $m_e = n^{-1}\sum_{i=1}^{n} x_i$. Utilisons à présent, toujours sous les mêmes hypothèses, le périodogramme associé afin d'obtenir un estimateur empirique de la densité spectrale de probabilité :

$$\forall\ \omega_k = 2\pi k/n, \quad P_e(\omega_k) = (2\pi)^{-1}\sum_{|h|<n}\rho_e(h)\exp\{-\mathrm{j}h\omega_k\} = (2\pi n)^{-1}|\sum_{i=1}^{n} x_i \exp\{-\mathrm{j}i\omega_k\}|^2 \tag{4}$$

B. Principe

Il n'est pas toujours facile de distinguer à l'oeil nu d'éventuels tendances ou facteurs saisonniers cachés dans une série chronologique. Toutefois, nous disposons d'outils statistiques assez performants. Calculer et représenter graphiquement pour chacune des séries précédentes (III-A) la fonction d'autocorrélation et le périogramme empiriques correspondants. Commenter les résultats obtenus (les séries chronologiques étudiées contiennent-elles une partie déterministe?)

C. Fonctions utiles

Utiliser sous MATLAB les fonctions suivantes : *mean(·)*, *fft(·,·)* et *abs(·)*.

V. Estimation et elimination de la tendance ainsi que du facteur saisonnier

Le modèle des observations x_i retenu dans cette section est celui défini par l'équation (1).

A. Estimation de la tendance

Méthode S_1 : recherche d'une tendance polynomiale g_i avec la méthode des moindres carrés

$$g_i = a_0 + a_1 * i + a_2 * i^2 + \cdots + a_d * i^d \tag{5}$$

Le problème de minimisation du critère des moindres carrés $\mathrm{S}_1(\boldsymbol{\theta}^\intercal) = \sum_i (x_i - g_i)^2$ admet une solution explicite:

$$\boldsymbol{x} \approx \boldsymbol{G} = \boldsymbol{I}\,\boldsymbol{\theta} \quad \Rightarrow \quad \boldsymbol{\theta}^{MC} = \boldsymbol{I}^\sharp \boldsymbol{x} = (\boldsymbol{I}^\intercal \boldsymbol{I})^{-1}\boldsymbol{I}^\intercal \boldsymbol{x} = \boldsymbol{\theta}_e \tag{6}$$

où:

$$\boldsymbol{I} = \begin{pmatrix} 1 & i_1 & i_1^2 & \cdots & i_1^d \\ 1 & i_2 & i_2^2 & \cdots & i_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & i_N & i_N^2 & \cdots & i_N^d \end{pmatrix} \quad \boldsymbol{x} = \begin{pmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_N} \end{pmatrix} \quad \boldsymbol{G} = \begin{pmatrix} g_{i_1} \\ g_{i_2} \\ \vdots \\ g_{i_N} \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} \tag{7}$$

et où $\sharp$ désigne l'opérateur pseudo-inverse.

NB : La méthode des moindres carrés próposée dans ce TP est l'extension de celle présentée en cours (c.f. tendances et facteurs saisonniers, section 3.VI) à un ordre $d \geq 2$.

Méthode S_2 : Lissage par moyenne mobile

Le filtrage employé doit permettre de restituer la tendance en éliminant dans un premier temps le facteur

saissonnier de période p. Si p est pair ($p = 2r$), on estime la tendance $\{g_i\}_{(r \leq i \leq N-r)}$ de la manière suivante:

$$(g_i)_e = (0.5x_{i-r} + x_{i-r+1} + \cdots + x_{i+r-1} + 0.5x_{i+r})/p \tag{8}$$

Au contraire, si p est impair ($p = 2r + 1$), on estime plutôt la tendance $\{g_i\}_{(r+1 \leq i \leq N-r)}$ de la manière suivante:

$$(g_i)_e = (2r+1)^{-1} \sum_{m=-r}^{r} x_{i+m} \tag{9}$$

Méthode S_3 : Différenciation à l'ordre p

Afin de traiter le cas d'une série chronologique avec facteur saisonnier de période p, la méthode de différenciation doit être considérée à l'ordre p. Rappelons la signification de l'opérateur gradient ∇_p:

$$\nabla_p x_i = x_i - x_{i-p} = (1 - B^p)x_i \tag{10}$$

L'opérateur gradient ∇_p ne doit pas être confondu avec l'opérateur de différentiation ∇^p défini par $\nabla^p = (1 - B)^p$. En appliquant l'opérateur ∇_p sur les données, on obtient:

$$\nabla_p x_i = g_i - g_{i-p} + w_i - w_{i-p} \tag{11}$$

Contrairement aux méthodes S_1 et S_2, la tendance g_i n'est pas ici explicitement calculée, seule la différence $g_i - g_{i-p}$ l'est. Toutefois, ce n'est pas important dans la mesure où à terme seule l'estimation de la fluctuation résiduelle compte. La section (V-C) nous montrera comment l'obtenir à partir de l'équation (11).

B. Elimination de la tendance et estimation du facteur saisonnier : méthodes S_1 et S_2 uniquement

Après avoir estimé et soustrait la tendance $(g_i)_e$ aux données x_i pour les différentes valeurs de i, il reste à estimer le facteur saisonnier s_i de la manière suivante:

$$\forall\ 1 \leq m \leq p, \quad \exists\ K = \mathrm{Int}(n/p), \quad (s_m)_e = \frac{1}{K} \sum_{k=1}^{K} \left[(x_{kp+m} - (g_{kp+m})_e) - \frac{1}{p} \sum_{m=1}^{p} (x_{kp+m} - (g_{kp+m})_e) \right] \tag{12}$$

où $\mathrm{Int}(\cdot)$ est l'opérateur partie entière.

C. Estimation de la fluctuation résiduelle

Concernant les méthodes S_1 et S_2, après avoir estimé et soustrait le facteur saisonnier $(s_i)_e$ aux données $x_i - (g_i)_e$ pour les différentes valeurs de i, on obtient alors une estimée de la fluctuation résiduelle $w_i = x_i - (g_i)_e - (s_i)_e$ à chaque instant i. A noter qu'une amélioration de la méthode S_2 consiste, après avoir estimé le facteur saisonnier s_i, à réestimer la tendance par la méthode des moindres carrés ou à nouveau par un filtrage à moyenne mobile, à partir des observations corrigées $x_i - (s_i)_e$. Enfin, la méthode S_3 n'impose pas d'estimer le facteur saisonnier pour obtenir la fluctuation résiduelle. En effet, il suffit d'appliquer l'opérateur ∇^k à l'équation (11), où k est une certaine puissance à définir.

D. Mise en oeuvre des méthodes

Programmer les trois méthodes, les appliquer à la série chronologique issue du fichier ***airline.dat*** et les comparer en terme d'erreur résiduelle notée e_r et définie par $e_r = \sum_i (x_i - (g_i)_e - (s_i)_e)^2$.

E. Fonctions utiles

La fonction *pinv(·)* peut être utilisée, toutefois il est préférable de construire soi-même l'opérateur de pseudo-inversion et de vérifier qu'il respecte bien les propriétés attendues. D'autre part, noter que sous MATLAB l'opérateur "′" effectue bien une transposition mais conjugue également les éléments, aussi vaut-il mieux lui préférer l'opérateur ".′" lorsqu'une conjugaison complexe ne s'impose pas.

VI. Elimination de composantes spectrales particulierès

Dans cette partie, nous considérons que la partie déterministe de la série observée n'est autre qu'un signal sinusoïdal. Les données suivent alors le modèle décrit par l'équation (2).

A. Principe du filtrage

Méthode S_4 : construire le processus $\{y_i\}$ défini par $y_i - \alpha \exp\{\mathrm{j}\omega\} y_{i-1} = \alpha(x_i - \exp\{\mathrm{j}\omega\} x_{i-1})$ avec $0 < \alpha < 1$.

B. Mise en oeuvre de la méthode

Soit le processus AR(2) $\{w_i\}$ défini par $\phi(\mathrm{B}) w_i = \theta(\mathrm{B}) z_i$ où $\phi(\mathrm{B}) = -(\mathrm{B}+2)(\mathrm{B}-7)/14$, $\theta(\mathrm{B}) = 1$ et $\sigma^2 = 1$. Sous MATLAB, en utilisant la fonction *filter*, générer le processus $\{w_i\}$ à partir de $N = 500$ échantillons ($\{z_i\}$ sera supposé gaussien). D'autre part, construire le processus $\{x_i = \sin(0.3 * i * \pi) + w_i\}$.

Puis, réutiliser la fonction *filter* pour générer le processus $\{y_i\}$ à partir de $\{x_i\}$ en prenant $\alpha = 0.99$ et $\omega = 0.3$. Représenter alors sous MATLAB les deux processus, $\{x_i\}$ et $\{y_i\}$ tout d'abord dans le domaine temporel, puis dans le domaine fréquentiel. Commenter, que nous apporte la représentation spectrale? Quel est l'effet du filtrage exercé sur $\{x_i\}$.

VII. Conclusion

L'estimation des composantes déterministes constitue une étape essentielle dans l'étude d'une série chronologique. Elle devance et conditionne l'étape de modélisation de la fluctuation résiduelle.