Une Introduction aux Modèles à Composantes Inobservées et à la Proc UCM

Gilbert Colletaz

12 septembre 2019

Table des matières

Chapitre 1

Présentation générale

Dans cette modélisation rendue populaire par Harvey [1], la variable observée, y_t, est explicitement vue comme la somme d'un trend, d'une variable responsable de la présence d'un possible cycle, d'une partie saisonnière et d'un résidu orthogonal à toutes les précédentes et qualifié de composante irrégulière. On peut aussi prendre en compte l'influence de variables explicatives exogènes, $x_1, x_2, \ldots, x_m$, via un index linéaire à coefficients fixes usuel. L'écriture générale est donc de la forme :

$$y_t = T_t + S_t + C_t + r_t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{m} \beta_i x_{it} + \epsilon_t \tag{1.1}$$

où,

- T_t représente le trend,
- S_t décrit la saisonnalité,
- C_t est la partie cyclique,
- r_t un processus AR(1) pouvant représenter une composante cyclique particulière,
- $\sum_{i=1}^{p} \phi_i y_{t-i}$: une composante autorégressive
- ϵ_t est la composante irrégulière ou résidu.

En pratique, seule la série d'intérêt, y_t, est observée, d'où le nom donné à cette modélisation. L'intérêt de ce modèle vient de ce que, en ajoutant ou retirant telle ou telle composante ou/et en jouant sur la spécification des composantes retenues, il peut être utile pour représenter de manière satisfaisante un grand nombre de séries économiques. Par ailleurs il va permettre d'estimer

1. Andrew C. Harvey, *Forecasting structural time series models and the Kalman filter*, Cambridge University Press, 1989

les diverses composantes sous-jascentes à une trajectoire donnée, de faire des prévisions sur chacune et, par sommation de construire des prévisions sur la variable endogène y.

On voit donc que les deux approches, processus ARIMA et modèle à composantes inobservées peuvent remplir le même objectif à savoir la construction de prévisions. Elles ne s'opposent d'ailleurs pas : la forme réduite d'un modèle UCM est très souvent un processus ARIMA. L'avantage de l'UCM est de pouvoir construire un modèle intégrant les a priori que l'on peut avoir sur la série de travail, par exemple savoir qu'elle est soumis à un cycle sera aisément pris en compte dans l'UCM alors que cela est plus laborieux avec un ARIMA. L'autre avantage est l'interprétabilité de la modélisation : avec un UCM on a accès aux estimations des composantes trend, cycle, saisonnalité ce qui n'est généralement pas le cas dans la modèlisation ARIMA.

Sous SAS, les modèles à composantes inobservées sont estimés au moyen de la `Proc UCM`. Comme nous le verrons, cette procédure fournit des outils d'aide concernant les choix des composantes à retenir, leur spécification ainsi que des tests de validation de la modélisation retenue.

1.1 La modélisation du trend

Dans les modèles UCM, l.a composante T_t est souvent de type trend stochastique. Pour mémoire, dans l'écriture

$$T_t = T_{t-1} + \beta_t + \eta_t \sim i.i.d.\mathcal{N}(0, \sigma_\eta^2) \tag{1.2}$$

β_t est la pente du trend, ou *drift*.

La proc `UCM` va permettre de considérer deux sortes de trend :

1. Dans la première, qualifiée de *Local Level Model* (*LL*), on pose $\forall t,\ b_t = 0$ et

$$T_t = T_{t-1} + \eta_t,\ \eta_t \sim i.i.d.\mathcal{N}(0, \sigma_\eta^2) \tag{1.3}$$

   Le trend est alors une marche au hasard. En l'absence d'autres composantes autre que l'irrégulière, on peut noter deux cas particuliers du modèle LL :

   (a) Si $\sigma_\eta^2 = 0$, i.e. $\forall t\ , \eta_t = 0$ la composante trend est égale à une constante, T_1 et $y_t \sim i.i.d.\mathcal{N}(T1, \sigma_\epsilon^2)$.

   (b) Si $\sigma_\epsilon^2 = 0$ alors y obéit à une marche au hasard : $y_t = y_{t-1} + \eta_t$

   Afin d'illustrer les caractéristiques des modèles LL, nous allons considérer un cas simple où la série n'est constituée que de deux seules composantes :

un trend de type LL et une composante irrégulière de type bruit blanc, soit :

$$y_t = T_t + \epsilon_t, \text{ et}$$
$$T_t = T_{t-1} + \eta_t$$

Les séries Trend1 et Trend2 des graphes 1.1 et 1.2 sont simulées avec des écarts-types de la composante η_t respectivement égaux à 1.0 et 4.0, les évolutions de la première sont évidemment moins heurtées que celles de la seconde. Par ailleurs l'écart-type de l'innovation ϵ_t afférente aux variables $y1$ et $y2$ de la figure 1.1 est fixé à 3.0, alors que $\sigma_\epsilon = 9$ pour les séries $y3$ et $y4$ représentées dans la figure 1.2. Les fluctuations de ces deux dernières autour de leur trend sont ainsi plus prononcées que les déviations observées sur les deux premières.

2. Qualifiée de modèle à tendance localement linéaire (*LLT*) la seconde spécification correspond au système suivant :

$$T_t = T_{t-1} + \beta_{t-1} + \eta_t,\ \eta_t \sim i.i.d.\ \mathcal{N}(0, \sigma_\eta^2) \text{ et,} \tag{1.4}$$
$$\beta_t = \beta_{t-1} + \xi_t,\ \xi_t \sim i.i.d.\ \mathcal{N}(0, \sigma_\xi^2) \tag{1.5}$$

où η_t et ξ_t sont des gaussiennes indépendantes.

Alors que dans le modèle LL précédent un processus de marche au hasard gouvernait le niveau du trend, il gouverne maintenant la pente du trend β_t.
On peut à nouveau illustrer certaines caractéristiques du modèle LLT en considérant toujours le cas le plus simple : la variable y_t ne contient que deux composantes, un trend de type LLT et une composante irrégulière de type bruit blanc :

$$y_t = T_t + \epsilon_t,\ \epsilon_t \sim i.i.d \mathcal{N}(0, \sigma_\epsilon^2), \tag{1.6}$$
$$T_t = T_{t-1} + \beta_{t-1} + \eta_t,\ \eta_t \sim i.i.d\ \mathcal{N}(0, \sigma_\eta^2) \tag{1.7}$$
$$\beta_t = \beta_{t-1} + \xi_t,\ \xi_t \sim i.i.d\ \mathcal{N}(0, \sigma_\xi^2) \tag{1.8}$$

Plusieurs cas particuliers découlent du modèle LLT :

(a) Lorsque $\sigma_\xi = 0$, i.e. lorsque $\xi_t = 0.0$, alors $\Delta T_t = \beta_1 + \eta_t$: le trend est une marche aléatoire avec dérive, il incorpore un trend linéaire de pente β_1. Si, dans le même temps $\beta_1 = 0$, on retrouve le modèle LL.

(b) Lorsque $\sigma_\xi = 0$ et $\sigma_\eta = 0$, le trend est une droite de pente β_1 et d'ordonnée à l'origine égale à T_1. Dans ce cas, $y_t = \beta_1 t + T_1 + \epsilon_t$.

(c) Lorsque $\sigma_\xi > 0$ et $\sigma_\eta = 0$ alors la variation du trend est une marche au hasard, on qualifie ce type de trend de *marche au hasard intégrée*.

Les graphiques 1.3, 1.4 et 1.5 illustrent les précédentes remarques. Ces quelques illustrations montrent l'étendue des possibilités données par les modélisations LL ou LLT : en conséquences on peut penser qu'elles peuvent être utiles pour représenter les tendances observées dans un grand nombre de séries réelles.

1.1.1 Commandes relatives au trend dans proc `UCM`

Le choix entre le modèle LL et le modèle LLT est géré par les commandes `LEVEL` et `SLOPE`. La première demande la prise en compte d'un trend, la présence ou l'absence de la seconde en précise le type : LLT si elle apparaît, LL sinon.

- La commande LEVEL possède les options suivantes :
  - `CHECKBREAK` : recherche une modification permanente dans le niveau de la série à une date t_0, $1 < t_0 < T$. En itérant sur t_0, la procédure teste la significativité d'une variable indicatrice valant 0 avant cette date et 1 après. Si un changement permanent de niveau est détecté, la variable indicatrice en question peut être introduite comme explicative dans le modèle afin de prendre en compte ce choc structurel. L'exemple 34.7 de la documentation relatif au niveau des eaux du Nil illustre ce cas de figure.
  - `VARIANCE`= donne une valeur positive ou nulle à σ_η^2
  - `NOEST` si cette option est présente, la valeur donnée par `VARIANCE`= à σ_η^2 s'impose, i.e. il ne s'agit plus d'un paramètre à estimer. En son absence, la valeur indiquée est simplement une valeur initiale donnée à l'estimateur s_η^2 et transmise à l'algorithme d'estimation des paramètres.

- La commande `SLOPE` qui ne peut pas être utilisée sans la présence de la commande `TREND` possède les options suivantes :
  - `VARIANCE`= donne une valeur positive ou nulle à σ_ξ^2.
  - `NOEST` si cette option est présente, la valeur donnée par `VARIANCE`= à σ_ξ^2 s'impose et ce paramètre est exclu de la liste des coefficients à estimer, sinon, il s'agit de la valeur initiale à partir de laquelle doit s'effectuer la recherche des estimations.

Exemples :

- Marche aléatoire avec dérive, i.e. trend localement linéaire de pente constante :

FIGURE 1.1 – Local Level Model, $\sigma_\eta = 1.0$ ou 4.0, $\sigma_\epsilon = 3.0$

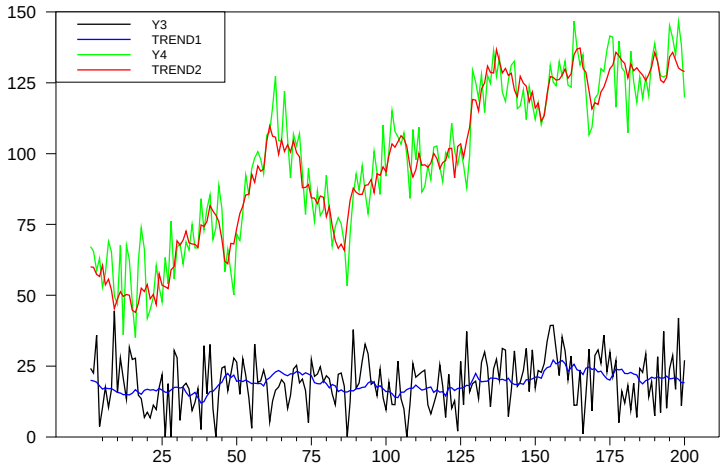

FIGURE 1.2 – Local Level Model
$\sigma_\eta = 1.0$ ou 4.0, $\sigma_\epsilon = 9.0$

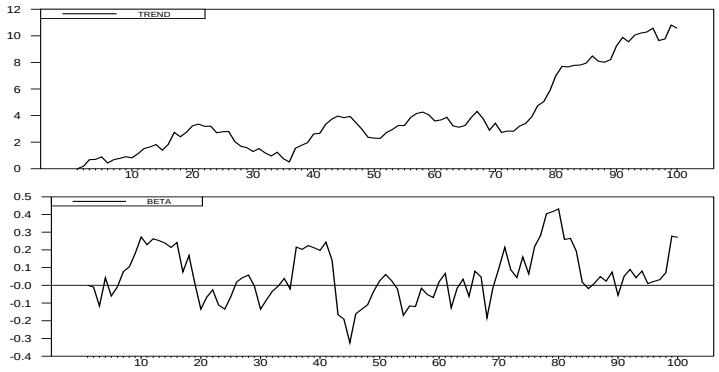

FIGURE 1.3 – Local Linear Trend Model : exemple de trend et de pente

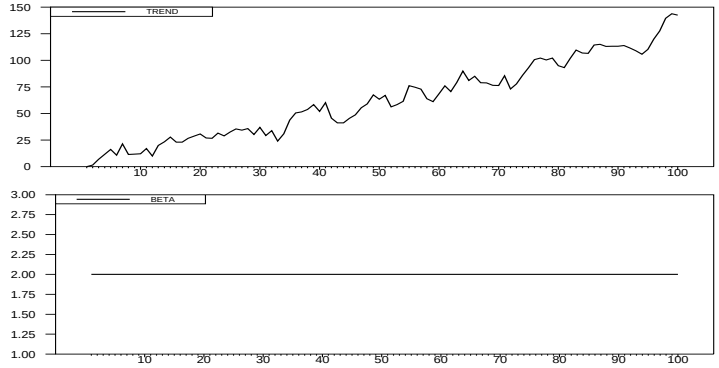

FIGURE 1.4 – Local Linear Trend Model : exemple de trend avec pente constante, *i.e.* $\sigma_\xi = 0.0$

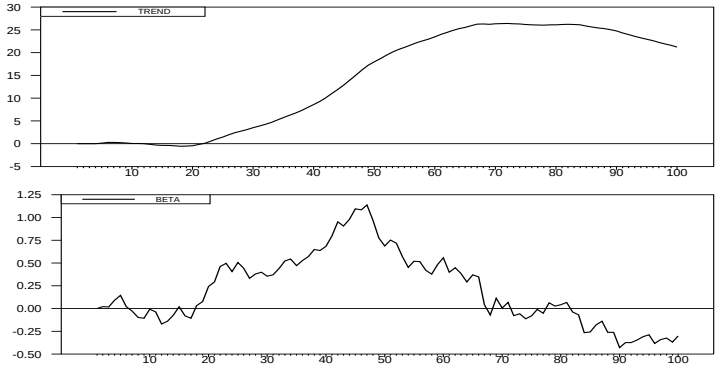

FIGURE 1.5 – Local Linear Trend Model : integrated random walk case, *i.e.* $\sigma_\eta = 0.0$, $\sigma_\xi > 0.0$

```
level;
slope variance=0 noest;
```

- Trend obéissant à un integrated random walk :

```
level variance=0 noest;
slope;
```

- Le trend se résume à la présence d'une constante :

```
level variance=0 noest;
```

- Le trend est linéaire déterministe :

```
level variance=0 noest;
slope variance=0 noest;
```

1.2 La modélisation des cycles

Traditionnellement on distingue quatre phases qui se succèdent pour constituer un cycle économique : une période d'expansion, suivie d'un ralentissement ou récession, puis d'une dépression, période la plus défavorable qui précède normalement la reprise. Ce simple énoncé fait évidemment penser que des fonctions périodiques sont adaptées à la modélisation de ces cycles : on sait que presque toutes les fonctions périodiques continues peuvent être écrites comme somme de sinusoïdes. Ce sont d'ailleurs de telles fonctions qui sont mobilisées au sein de la `proc UCM`. La prise en compte des cycles est bâtie sur des fonctions déterministes qui ont été enrichies par des éléments stochastiques afin d'apporter plus de souplesse et/ou de parcimonie dans les outils d'analyse. Le principe est de pouvoir relativement simplement mettre en place un système capable, via l'estimation d'un nombre assez réduit de paramètres, de reproduire les évolutions d'un cycle qui lui peut être complexe.

Avant d'exposer les possibilités de modélisation des cycles qu'offre la proc `UCM`, et afin de faciliter la compréhension de celles-ci, une brève introduction à l'analyse spectrale s'impose.

1.2.1 Quelques éléments d'analyse spectrale

Une série temporelle est une succession d'observations représentant l'évolution d'une variable dans le temps. Il est dès lors naturel de l'étudier dans le domaine des temps au moyen d'outils adaptés à ce domaine comme la fonction d'autocovariance. Pour autant, on peut également l'étudier dans le domaine des fréquences : les composantes de la série (tendancielle, saisonnière, cyclique) sont perçues comme des composantes périodiques de fréquences et

d'amplitudes différentes. Dans cette approche, l'objectif sera d'identifier les composantes en question en repérant les fréquences principales qui composent la partie stationnaire du processus au moyen de sa densité spectrale.

Une variable périodique stationnaire va se répéter à l'identique à des intervalles de temps égaux. Elle sera caractérisée en particulier par sa période et son amplitude.

- On appelle période la plus petite durée de ces intervalles de temps et l'inverse de la période définit la fréquence, *i.e.* $T = 1/f$. Cette fréquence, f, est le nombre de fois où le signal se répète au cours d'un intervalle de temps donné [2]. On va considérer que les signaux de fréquence nulle ont une période infinie, i.e. ne sont pas périodiques. On définit aussi la fréquence angulaire ou vitesse de rotation $\omega = 2\pi f = 2\pi/T$ qui se mesure en radians [3] .
- L'amplitude est l'écart maximal entre la série et son espérance et l'amplitude crête à crête est l'écart entre les valeurs min et max du signal.

Les graphiques 1.6 et 1.7 illustrent ces définitions. Le signal sinusoïdal de la première figure 1.6 a une amplitude égale à l'unité et une période à 2π. La seconde montre l'impact d'un déphasage 180 degrés, ou π radians, entre deux signaux.

Dans cette introduction il est hors de question de faire une présentation rigoureuse des outils de l'analyse spectrale. On va simplement essayer de faire comprendre l'intérêt de celle-ci lorsque l'on cherche à extraire les principales composantes (saisonnalités, cycles) d'une série stationnaire.

La densité spectrale

Le premier outil à considérer est la densité spectrale. Si y_t est un processus stationnaire ayant l'écriture de Wold $y_t = \sum_{i=0}^{\infty} \psi_i u_{t-i}$, où donc u est un processus en bruits blancs et $\sum_{i=0}^{\infty} |\psi_i| < \infty$, alors la densité spectrale de y est la fonction f définie par :

$$\forall \omega \in [-\pi, \pi], f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_y(h) e^{i\omega h} \tag{1.9}$$

avec $\gamma_y(h)$ la fonction d'autocovariance de y, $e^{i\omega h} = \cos(\omega h) + i \sin(\omega h)$ et $i^2 = -1$.

Par ailleurs, pour tout entier h,

$$\gamma_y(h) = \int_{-\pi}^{\pi} e^{-i\omega h} f(\omega) d\omega \tag{1.10}$$

2. Supposez qu'un manège effectue 10 tours pendant pendant 3 minutes, c'est la fréquence, alors sa période est de 1/10$^{\text{ème}}$ de 3 minutes : vous pouvez faire coucou à un enfant toutes les 18 secondes, c'est la période.

3. il y a 2π radians dans un tour complet de 360 degrés et donc un radian vaut $360^o/2\pi = 57.3$ degrés.

avec $\gamma_y(h)$ la fonction d'autocovariance de y et $e^{i\omega h} = \cos(\omega h) + i\sin(\omega h)$.

Les deux équations précédentes montre qu'un processus caractérisé par sa fonction d'autocovariance peut tout aussi bien l'être par sa fonction de densité : avec (1.9) on passe des autocovariances à la densité spectrale, avec (1.10) on reconstruit les autocovariances à partir de la densité spectrale.

On peut montrer que f est une fonction paire : $f(\omega) = f(-\omega)$, et qu'elle est périodique de période 2π. Il suffit donc de l'étudier sur le support $[0, \pi]$.

Le théorème de représentation spectrale

Ce théorème permet en quelque sorte la reconstruction d'un signal à partir de sa densité spectrale. Formellement, il affirme que

$$y_t = \int_{-\pi}^{\pi} e^{-i\omega h} d\xi(\omega), \tag{1.11}$$

où $\xi(\omega)$ est une mesure stochastique associée à y, ce que nous n'expliquerons pas ici, qui vérifie :

$$\text{cov}(d\xi(\omega_1), d\xi(\omega_2)) = 0 \tag{1.12}$$

L'importance de ce théorème vient de ce qu'il affirme que chaque observation y_t est une somme pondérée des termes $e^{-i\omega h}$ avec des poids dépendant de la fréquence ω et orthogonaux entre eux [4].

Enfin, un autre résultat important montre que chacun de ces poids est proportionnel à $f(\omega)$. En conséquence, si $f(\omega_1) > f(\omega_2)$ alors la composante de fréquence ω_1 est plus importante que celle de fréquence ω_2 dans la décomposion de y_t.

Ainsi, l'observation de pics à certaines fréquences angulaires permet donc d'identifier les composantes importantes d'une série. Ainsi,

- Pour des données trimestrielles, la période est $T = 4$, ce qui correspond à $\omega = \pi/2$. Un pic dans la densité spectrale à cette fréquence traduit donc une saisonnalité trimestrielle.
- Pour des observations mensuelles, la période est $T = 12$: un pic à $\omega = \pi/6$ révèle une saisonnalité mensuelle.
- un pic à $\omega = \pi$ signale une composante stationnaire ayant une période de 2 unités de temps, i.e. une composante dont les observations successives se situent de part et d'autre de sa moyenne.

4. Vous pouvez d'ailleurs maintenant comprendre que l'analyse spectrale peut être utilisée pour désaisonnaliser une série : dans la reconstruction de y_t, il suffit de masquer les fréquences proches et égale à la saisonnalité que l'on veut éliminer.

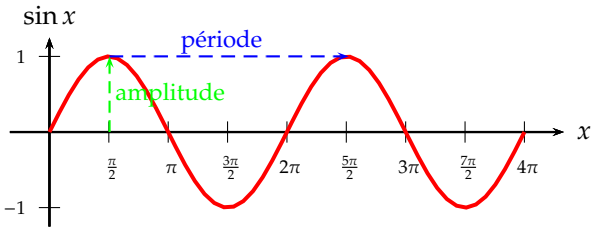

FIGURE 1.6 – Exemple de signal périodique de période 2π, d'amplitude unitaire et de phase à l'origine nulle.

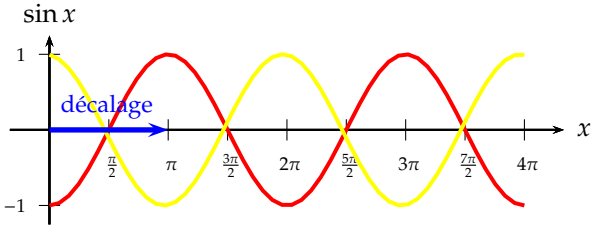

FIGURE 1.7 – Exemple de signaux en opposition de phase, *i.e.* déphasés de 180^o.

La densité spectrale d'un processus ARMA(p,q) stationnaire d'écriture $\phi(L)x_t = \theta(L)u_t$ s'écrit encore :

$$f(\omega) = \frac{\sigma_X^2}{2\pi} \frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})}, \text{ avec } z = e^{i\omega} \tag{1.13}$$

Quelques exemples :

- La densité spectrale d'un bruit blanc : $f(\omega) = \frac{\sigma_X^2}{2\pi}$. C'est une constante et l'absence de pics signale l'absence de composantes périodiques.
- La densité spectrale d'un AR(1) [5] :

$$\begin{aligned} f(\omega) &= \frac{\sigma_X^2}{2\pi}(1 - \phi e^{i\omega})^{-1}(1 - \phi e^{-i\omega})^{-1} \\ &= \frac{\sigma_X^2}{2\pi}\left(1 - \phi(e^{i\omega} + e^{-i\omega}) + \phi^2\right)^{-1} \\ &= \frac{\sigma_X^2}{2\pi}(1 - 2\phi\cos\omega + \phi^2)^{-1} \end{aligned}$$

5. Rappels : $\cos\omega = \cos(-\omega)$, $\sin\omega = -\sin(-\omega)$, $e^{i\omega} = \cos\omega + i\sin\omega$.

et

$$f'(\omega) = \frac{\sigma_X^2}{2\pi} \frac{-2\phi \sin \omega}{((1 - 2\phi \cos \omega + \phi^2)^2}$$

Ainsi, sur $[0, \pi]$

$$f'(\omega) \begin{cases} < 0 \text{ si } \phi > 0, \\ > 0 \text{ si } \phi < 0 \end{cases}$$

La densité spectrale d'un AR(1) est donc décroissante pour $\phi > 0$ et croissante lorsque $\phi < 0$. Ainsi, pour des autocorrélations positives, $\phi > 0$, la série est dominée par des composantes de fréquences nulles et proches de zéro, elle tend à se comporter comme une série non périodique. En revanche, avec $\phi < 0$, elle sera dominée par des composantes de fréquences élevées et aura donc un comportement périodique marqué. On peut montrer que selon la valeur de ϕ deux pics apparaîtront : un en zéro ($\phi > 0$), l'autre en π, correspondant respectivement à des séries de période infinie pour le premier et de période égale à 2 pour le second. Ces résultats permettent de comprendre la modélisation via un AR(1) de ces deux cas particuliers de composantes cycliques que va offrir la proc `UCM` avec la commande `AUTOREG` [6].

1.2.2 Cycles déterministes et stochastiques dans la proc `UCM`

Un cycle déterministe ψ_t de fréquence ω, d'amplitude $\sqrt{\alpha^2 + \beta^2}$ peut s'écrire comme :

$$\psi_t = \alpha \cos(\omega t) + \beta \sin(\omega t),\ 0 < \omega < \pi, \tag{1.14}$$

et sa période, i.e. le temps qu'il lui faut pour se reproduire à l'identique, est égale à $T = 2\pi/\omega$. Un exemple d'un tel cycle construit avec $\alpha = \beta = 1$, donc d'amplitude égale à $\sqrt{2}$ et de période égale à 4 est donné dans la figure 1.8 [7]

Naturellement dans les séries économiques il est rare de trouver une telle régularité des cycles, avec période et amplitude invariantes dans le temps. Nous pouvons alors imaginer d'additionner des cycles déterministes de paramètres différents afin d'approximer raisonnablement bien les cycles observés. Cette possibilité est offerte par la proc `UCM`, cependant d'une part nous ne connaissons pas a priori le nombre de cycles déterministes qu'il faudrait spécifier, et d'autre part ce nombre pourrait être élevé, aboutissant à une modélisation "lourde" de cette composante cyclique. Pour ces raisons de simplification et de parcimonie, la procédure autorise la construction de cycles stochastiques dont les caractéristiques vont se modifier en fonction d'un nombre restreint de paramètres, permettant ainsi d'approcher des cycles irréguliers.

6. Voir infra la section 1.2.4.

7. L'aspect sinusoïdal n'apparaît pas et la série représentée n'atteint pas son amplitude car nous l'évaluons en une succession de points discrets et non pas sur $\mathcal{R}$.

Pour cela on considère au départ un système récursif de calcul de cycles déterministes :

$$\begin{pmatrix}\psi_{1t}\\ \psi_{2t}\end{pmatrix} = \begin{bmatrix}\cos\omega & \sin\omega\\ -\sin\omega & \cos\omega\end{bmatrix}\begin{pmatrix}\psi_{1,t-1}\\ \psi^*_{2,t-1}\end{pmatrix} \tag{1.15}$$

initialisé à $\psi_0 = \alpha$ et $\psi^*_0 = \beta$.

On donne dans le graphique 1.9 un exemple de cycles générés par ce système avec les paramètres $\alpha = \beta = 1$ et une période égale à 12. On passe alors à des trends stochastiques simplement en ajoutant aux équations précédentes des innovations bruits blancs gaussiens indépendants de même variance σ^2_v :

$$\begin{pmatrix}\psi_{1t}\\ \psi_{2t}\end{pmatrix} = \begin{bmatrix}\cos\omega & \sin\omega\\ -\sin\omega & \cos\omega\end{bmatrix}\begin{pmatrix}\psi_{1,t-1}\\ \psi^*_{2,t-1}\end{pmatrix} + \begin{pmatrix}v_{1t}\\ v_{2t}\end{pmatrix} \tag{1.16}$$

À la sortie on récupère des cycles qui ont toujours la même période mais des amplitudes et des phases variant avec t. Un exemple est donné par le graphique 1.10 qui utilise les mêmes paramètres α, β et la même période que ceux utilisés pour le graphique 1.9 et où on a ajouté des réalisations de gaussiennes indépendantes conformément à l'équation (1.16) [8].

Enfin, toujours afin d'accroître la flexibilité du modèle et sa capacité à s'adapter à diverses configurations réelles, il est possible de donner une forme autorégressive aux systèmes de cycles stochastiques précédents selon l'écriture :

$$\begin{pmatrix}\psi_{1t}\\ \psi_{2t}\end{pmatrix} = \rho\begin{bmatrix}\cos\omega & \sin\omega\\ -\sin\omega & \cos\omega\end{bmatrix}\begin{pmatrix}\psi_{1,t-1}\\ \psi^*_{2,t-1}\end{pmatrix} + \begin{pmatrix}v_{1t}\\ v_{2t}\end{pmatrix} \tag{1.17}$$

Comme on le sait, ces cycles seront stationnaires si $|\rho| < 1$ et non stationnaires à période constante si $\rho = 1$

1.2.3 Commandes relatives au cycle dans proc UCM

Des cycles seront incorporés dans le modèle selon que la commande `CYCLE` est présente ou non. Par ailleurs, le nombre de fois où cette commande apparaît détermine le nombre de cycles qui seront pris en compte.
Les options suivantes sont disponibles :

- `PERIOD=`, `VARIANCE=`, `RHO=`, permettent d'attribuer des valeurs à la période du cycle, à σ^2_v, à ρ, sachant que $0 \leq \rho < 1$ et $\sigma^2_v \geq 0$. Ainsi, contraindre ρ à zéro élimine l'écriture autorégressive sur le cycle stochastique. Lui donner une valeur unitaire impose un trend non stationnaire. Contraindre $\sigma^2_v = 0$ impose des cycles déterministes.

8. Dans cet exemple nous avons pris des réalisations de gaussiennes centrées-réduites.

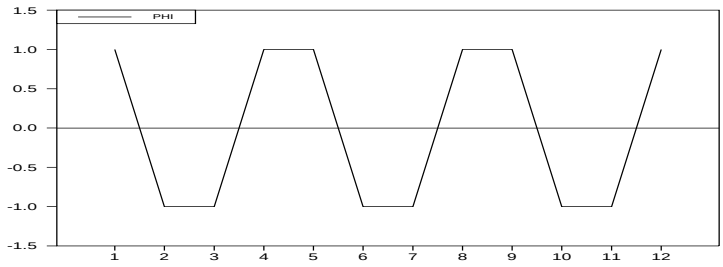

FIGURE 1.8 – Exemple de cycle déterministe d'amplitude 1.414 et de période 4

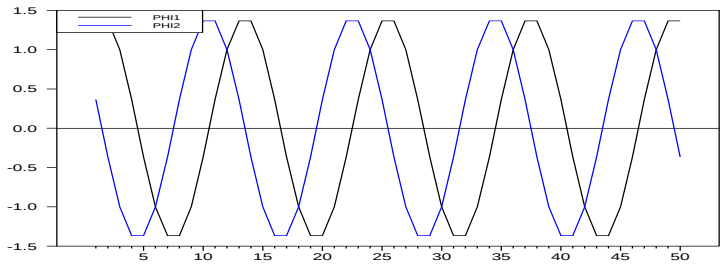

FIGURE 1.9 – Exemple de cycles déterministes générés par le système (1.15)

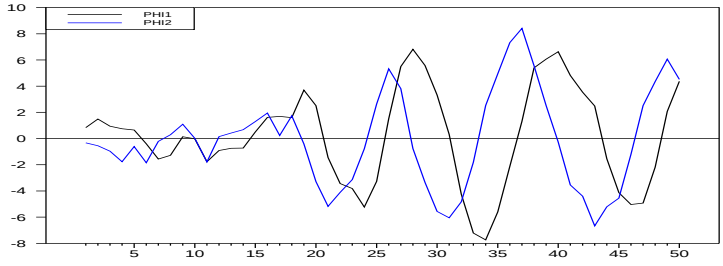

FIGURE 1.10 – Exemple de cycles stochastiques générés par le système (1.16)

- `NOEST=PERIOD, NOEST=VARIANCE, NOEST=RHO` excluent le paramètre concerné de la liste des coefficients à estimer et leur imposent la valeur spécifiée dans l'option précédemment décrite. En l'absence de cette option `NOEST=`, la valeur en question est simplement une valeur initiale transmise à l'algorithme d'estimation.

Exemples :

- deux cycles stochastiques seront présents dont tous les paramètres seront estimés avec

```
cycle;
cycle;
```

- incorporation d'un cycle stochastique non stationnaire de période 4 avec

```
cycle period=4 rho=1 noest=(period rho);
```

1.2.4 Cas particulier : la composante AR(1) non observée

On sait que la densité spectrale d'un processus AR(1) d'écriture

$$r_t = \rho r_{t-1} + v_t \tag{1.18}$$

a un pic maximal pour une fréquence angulaire nulle lorsque $\rho > 0$ et égale à π lorsque $\rho < 0$. Le premier cas correspond à une période de durée infinie, ce qui signifie l'absence de cyclicité dans le processus. Dans le second, la période est de longueur 2 : le cycle dominant implique un passage de la variable de part et d'autre de sa moyenne à chaque incrément du temps. Afin de pouvoir traiter ces deux cas particuliers, on peut incorporer une composante autorégressive telle que donnée par l'équation (1.18), avec $v_t \sim \ i.i.d. \mathcal{N}(0, \sigma_v^2)$.
On sait également que pour $|\rho| < 1$ le processus est stationnaire. Il est non stationnaire pour $\rho = \pm 1$ et explosif si $|\rho| > 1$.

1.2.5 Commandes relatives à la composante autorégressive

L'introduction d'une composante autorégressive d'ordre 1 décrite par l'équation (1.18) dans le modèle est réclamée par la commande `AUTOREG`. Les options disponibles sont :

- `RHO=` : permet de spécifier une valeur pour le coefficient ρ. Les valeurs autorisées sont telles que $\rho \in (-1, 1($. Notez que $\rho = 1$ est exclue car alors cette composante autorégressive n'est rien d'autre qu'un trend de type LL déjà géré par les commandes `TREND` et `SLOPE`.

- `VARIANCE=` : donne une valeur positive ou nulle à σ_v^2.

- `NOEST=RHO` : avec cette option $\hat{\rho}$ prend la valeur précisée par `RHO=`. En son absence, cette valeur est une valeur initiale transmise à l'algorithme d'estimation.
- `NOEST=VARIANCE` : s_v^2 prend la valeur précisée par `VARIANCE=`. En son absence, la valeur en question est simplement une valeur initiale transmise à l'algorithme d'estimation.
- `NOEST=(RHO VARIANCE)` : $\hat{\rho}$ et s_v^2 prennent les valeurs données par `RHO=` et `VARIANCE=`

Exemple :

- Modèle de la forme $y_t = \ldots + r_t + \ldots$ et $r_t = \rho r_{t-1} + v_t, v_t \sim$ *i.i.d.* $\mathcal{N}(0, \sigma_v^2)$:

```
autoreg;
```

- Le même en imposant $\rho = 0.5$:

```
autoreg rho=0.5 noest=rho;
```

1.3 La modélisation de la saisonnalité

La saisonnalité est responsable de déviations des observations autour de la somme trend+cycle, déviations qui ont tendance à se répéter sur des périodes qui se succèdent. Une particularité de ces composantes saisonnières est que la durée de ces périodes, ou span, est connue a priori. Par exemple, des déviations identiques observées sur les observations des mêmes mois et d'années différentes sur des données mensuelles, ou bien à chaque même trimestre sur des données trimestrielles, etc... Une particularité de cette composante est que l'on peut avoir à prendre en compte plusieurs saisonnalités au sein d'une série donnée. Ainsi, sur des observations collectées toutes les heures, nous pourrions avoir une saisonnalité horaire, le même type de déviation est observé tous les jours à 8h, à midi, etc..., ainsi qu'une saisonnalité journalière : les données du weekend décalant de celles des jours ouvrés, celles du lundi ayant aussi un profil particulier, etc...

La proc `UCM` offre au moins deux possibilités de représentation de ces saisonnalités :

1. via des indicatrices : Une prise en compte bien connue de la saisonnalité est d'introduire des constantes spécifiques à chacune des observations contenues dans un span donné. Ainsi, pour un span de longueur s, nous pourrions introduire s constantes $\gamma_1, \gamma_2, \ldots, \gamma_s$. Par ailleurs, afin de distinguer saisonnalité, trend et cycle, on va contraindre à l'absence d'effet de la saisonnalité sur le niveau moyen de la variable au cours d'un span,

ce qui va imposer qu'à chaque date t, $\sum_{i=0}^{s-1} \gamma_{t-i} = 0$, ce qui implique que seules $s-1$ de ces constantes sont libres.

La proc `UCM` va autoriser une généralisation de cette saisonnalité déterministe en postulant que les termes γ_{it} ne sont pas des constantes mais des aléatoires vérifiant :

$$\sum_{i=0}^{s-1} \gamma_{it-s} = \omega_t \text{ avec } \omega_t \sim i.i.d.\, \mathcal{N}(0, \sigma_\omega^2) \tag{1.19}$$

En d'autres termes, la somme des effets saisonniers au cours d'un span n'est plus nulle mais seulement nulle en moyenne. Ceci va autoriser des variations au cours du temps dans les réalisations des γ_{it}, les amplitudes de ces variations dépendant notamment de σ_ω^2. On voit d'ailleurs aisément qu'en imposant la nullité de σ_ω^2 cette modélisation par variables aléatoires redevient une modélisation de la saisonnalité par variables déterministes de type dummy variables.

2. via des représentations fréquentielles : γ_t est construit comme une somme de cycles de fréquences différentes :

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt}, \tag{1.20}$$

où $[s/2] = s/2$ si s est pair et $(s-1)/2$ si s est impair, les γ_{jt} étant des cycles de fréquences $\omega_j = 2\pi j/s$. Ces cycles stochastiques de base sont obtenus par récurrence selon

$$\begin{pmatrix} \psi_{1t} \\ \psi_{2t} \end{pmatrix} = \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \begin{pmatrix} \psi_{1,t-1} \\ \psi^*_{2,t-1} \end{pmatrix} + \begin{pmatrix} \omega_{1t} \\ \omega_{2t} \end{pmatrix} \tag{1.21}$$

c'est-à-dire au moyen d'un système semblable à celui employé pour la modélisation des cycles [9].

Afin d'illustrer l'intérêt des représentations fréquentielles de la saisonnalité, on peut considérer l'exemple des deux séries représentées, avec leurs autocorrélations partielles, dans le graphe 1.11. Selon ses auteurs, Irma Hindrayanto, John A.D. Aston, Siem Jan Koopman et Marius Ooms [10], la première série ne requiert pas une représentation fréquentielle alors qu'elle serait adaptée sur la seconde série en raison de la multiplicité des périodicités apparentes dans le corrélogramme.

9. Cf. l'équation (1.16).
10. Exemple repris dans "Modeling Trigonometric Seasonal Components for Monthly Economic Time Series", *Tinbergen Institute Discussion Paper*, TI 2010-018/4.

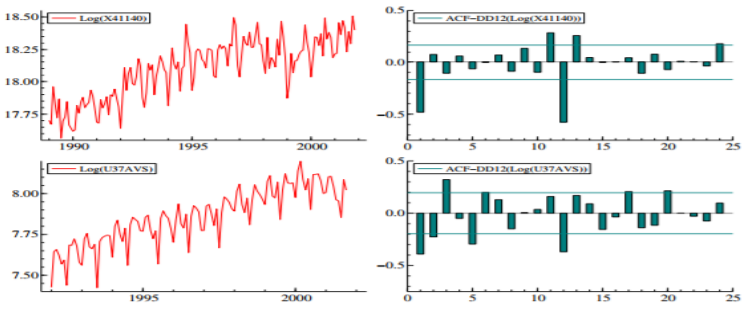

FIGURE 1.11 – Exemple

Avec les deux méthodes précédentes, variables muettes ou représentation fréquentielle, nous savons que pour un span de longueur s, le modèle va incorporer $s-1$ paramètres libres. En conséquence, il peut parfois être utile, notamment pour des valeurs de span très élevées d'introduire des contraintes qui vont permettre de diminuer ce nombre de paramètres libres. Par exemple, dans le cas de la modélisation trigonométrique, pour définir γ_t, au lieu de la somme complète $\sum_{j=1}^{[s/2]} \gamma_{jt}$, il sera possible d'éliminer certains cycles de base, souvent ceux de hautes fréquences dont la contribution est généralement faible ou, de manière équivalente, de spécifier une sous-liste de cycles de base à conserver.
On notera enfin que si on impose une saisonnalité déterministe, i.e. $\sigma^2\omega = 0$ alors les deux méthodes sont équivalentes et donneront les mêmes résultats.

1.3.1 Commandes relatives à la saisonnalité dans proc `UCM`

La commande `SEASON` permet de préciser les choix de modélisation de la saisonnalité. En son absence, cette composante saisonnière n'existe pas dans le modèle ajusté. Les options disponibles sont notamment les suivantes :

- `TYPE=DUMMY|TRIG`. Indique le type de modélisation à prendre : variables indicatrices si `TYPE=DUMMY`, représentation trigonométrique si `TYPE=TRIG`. Par défaut `TYPE=DUMMY` est sélectionné.

- `LENGHT=` : cette option sert à donner la longueur du span afférent à une saisonnalité que l'on veut prendre en considération. C'est une option qui doit être obligatoirement renseignée dans chacune des commandes SEASON utilisées.

- `VARIANCE=` précise éventuellement une valeur, qui peut être positive ou nulle pour le terme σ_ω^2 des équations (1.19) ou (1.21). Son statut est défini par la présence ou l'absence de l'option `NOEST`.

- `NOEST` : si cette option est présente, σ_ω^2 prend la valeur spécifiée par `VARIANCE=` et n'est plus un paramètre à estimer. Pour mémoire, si cette

valeur est nulle, alors la modélisation de la saisonnalité est déterministe. En son absence, la valeur en question est simplement la valeur initiale utilisée par l'algorithme d'estimation.

- `DROPHARMONICS=` : donne la liste des cycles de base à éliminer dans la construction des γ_t définie en (1.20). Cette liste doit contenir des entiers compris entre 1 et $[s/2]$.
- `KEEPHARMONICS=` : à l'opposé de l'option précédente, on va ici donner la liste des cycles de base à conserver dans la construction des γ_t définie en (1.20). Cette liste doit contenir des entiers compris entre 1 et $[s/2]$.

1.4 Prise en compte de variables explicatives

L'équation (1.1) donne l'écriture générale d'un modèle à composantes inobservées dans la proc `UCM` et montre que celle-ci autorise une dépendance de la variable d'intérêt y_t à un certain nombre d'autres variables, $x_{1t}, x_{2t}, \ldots, x_{mt}$ via l'index linéaire habituel dans un modèle de régression linéaire $\sum_{i=1}^{m} \beta_i x_{it}$ où les β_i sont des constantes inconnues. Elle va également autoriser la présence d'explicatives à coefficients variables dans ce modèle linéaire. Il est aussi possible d'ajuster une relation non linéaire non constante dans le temps via une commande `SPLINEREG` qui ne sera pas vue dans ce descriptif. Remarquons déjà que la présence de telles variables complique la construction des prévisions sur la variable expliquée : puisque les valeurs prises par les explicatives impactent le niveau de l'expliquée, on ne pourra évaluer que des prévisions conditionnelles à ces valeurs. En conséquence, pour obtenir des prévisions sur l'endogène, il faut être en mesure de spécifier des valeurs pour les explicatives sur la période de prévision.

1.4.1 Explicatives à coefficients constants

Celle-ci s'effectue selon la commande usuelle `MODEL` et sa syntaxe bien connue :

```
MODEL nom de l'expliquée = liste des noms des explicatives;
```

1.4.2 Explicatives à coefficients variables

Il est possible d'introduire des variables explicatives dont les coefficients obéissent à une marche au hasard :

$$\beta_{it} = \beta_{it-1} + \eta_t \quad \texttt{avec} \quad \eta_t \sim i.i.d.\, \mathcal{N}(0, \sigma_\eta^2) \tag{1.22}$$

Les variables explicatives concernées sont listées dans la commande `RANDOMREG` selon

```
RANDOMREG liste des noms des explicatives concernées;
```

Les options suivantes sont disponibles :

- `VARIANCE=` instruction qui permet de donner une valeur à σ_η^2
- `NOEST` , option qui, si elle est présente fige la valeur de σ_η^2 à celle indiquée dans l'option précédente. En son absence, la valeur en question est simplement une valeur initiale transmise à l'algorithme d'estimation.

Exemple :

- Modèle à trois explicatives de la forme $y_t = \ldots + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_{3,t} x_{3t}$ et $\beta_{3,t} = \beta_{3,t-1} + \eta_t$, η_t tel que défini en (1.22) :

```
model y = x1 x2 ;
randomreg x3 ;
```

1.5 Prise en compte d'une autorégression sur l'endogène

La commande `DEPLAG` permet d'intégrer des valeurs retardées de l'expliquée dans l'équation estimée. Elle possède les options suivantes :

- `LAG=`p, donne l'ordre de l'AR, ou encore
- `LAG=`$(l_1, l_2, \ldots, l_p)(L_1, L_2, \ldots, L_P)$, précise les retards des composantes non saisonnières et saisonnières de l'autorégression,
- `PHI=` liste de valeurs qui seront attribuées aux coefficients de l'autorégression dans l'ordre correspondant à celui indiqué par `LAG=`
- `NOEST` , en sa présence les valeurs précédentes s'imposent et les coefficients ϕ_i ne sont pas estimés. En son absence, les valeurs en question sont seulement des valeurs initiales transmises à l'algorithme d'estimation.

Exemple :

- partie autorégressive d'ordre 2 du type $\phi_1 y_{t-1} + \phi_2 y_{t-2}$:

```
deplag lag=2;
```

- modèle autorégressif AR(1)× SAR(1) de span 12 : $(1 - \phi_1 L)(1 - \phi_2 L^{12}) y_t$

```
deplag lag=(1)(12);
```

1.6 La composante résiduelle

Il s'agit ici de préciser les caractéristiques de l'aléatoire ϵ_t de l'équation (1.1). La présence de cette composante résiduelle dans le modèle est imposée par la présence de la commande `IRREGULAR`. Généralement cette présence doit toujours être imposée, sachant aussi qu'il ne peut y avoir qu'une composante résiduelle et une seule.
Afin de prendre en compte de possibles autocorrélations résiduelles non nulles, la proc `UCM` permet de spécifier que le terme ϵ_t est un ARMA(p,q)×SARMA(P,Q).

Les options disponibles sont :

- `VARIANCE=` , donne une valeur à σ_ϵ^2.
- `NOEST` ,en présence de cette option, la valeur de σ_ϵ^2 est figée à celle indiquée dans la précédente option.
- + un ensemble d'options permettant de décrire le modèle ARMA éventuel, la syntaxe rappelant celle de la proc `ARIMA` :
  - `AR =` : liste de valeurs initiales pour la partie AR non saisonnière,
  - `MA =` : liste de valeurs initiales pour la partie MA non saisonnière,
  - `SAR =` : liste de valeurs initiales pour la partie AR saisonnière,
  - `SMA =` : liste de valeurs initiales pour la partie MA saisonnière,
  - `P=` : entier positif ou nul indiquant l'ordre de la partie AR,
  - `Q=` : entier positif ou nul indiquant l'ordre de la partie MA,
  - `SP=` : entier positif ou nul donnant l'ordre de la partie AR saisonnière,
  - `SQ=` : entier positif ou nul donnant l'ordre de la partie MA saisonnière,
  - `S=` : entier positif ou nul indiquant la période de la saisonnalité,
  - `NOEST=(<VARIANCE> <AR> <MA> <SAR> <SMA>)` : selon le mot-clef utilisé, fige les valeurs des coefficients concernés à celles indiquées dans les options `VARIANCE=, AR=, MA=, SAR=, SMA=`.

Exemples

- MA(1)×SMA(1) sur données mensuelles :

```
irregular q=1 sq=1 s=12 ;
```

- SMA(1) de coefficient imposé égal à 0.8 sur données trimestrielles :

```
irregular sq=1 S=4 sma=0.8 noest=(sma);
```

1.7 La recherche d'outliers

La proc `UCM` permet la recherche d'observations atypiques, au sens où leurs valeurs décalent significativement de celles projetées par le modèle estimé. La démarche est la suivante : soit $\hat{y}_{t_0}$ la valeur expliquée de l'observation de rang t_0 par un modèle donné. On ajoute à ce modèle une nouvelle explicative I_t valant 1 si $t = t_0$ et 0 sinon. Soit β_I son coefficient, alors par construction, $y_t = \hat{y}_{t_0} + \beta_I$. Seront alors qualifiées qualifiées d'*additive outliers* les observations pour lesquelles β_I sera significativement non nul, *i.e.* celles dont la valeur observée décale significativement de la valeur expliquée.
En complément, on sait que des ruptures de pente peuvent être recherchées en activant l'option `checkbreak` dans la commande `level`.
Généralement la recherche d'observations atypiques est une étape qui précède une estimation finale en permettant d'apporter d'éventuelles corrections aux observations ainsi repérées. En sortie, la procédure liste les observations pour lesquelles les probabilités critiques associées à H0 :$\beta_I = 0$ sont les plus faibles. Cette recherche est activée par la commande `OUTLIERS` qui dispose des options suivantes :

- `ALPHA=` , va préciser le seuil de risque permettant de juger si β_I est significatif. Par défaut, il est égal à 5%.
- `MAXNUM=` , limite le nombre d'observations atypiques à rechercher. Par défaut cette option est fixée à 5.
- `MAXPCT=` , même rôle que la précédente, mais ici le nombre est indiqué en pourcentage de la longueur de la série traitée.
- `PRINT=SHORT|DETAIL` , gouverne le volume d'informations affiché sur les observations repérées comme atypiques. Par défaut, `PRINT=SHORT` : on récupère la liste de ces observations atypiques en question avec leurs probabilités critiques. Avec `DETAIL`, en plus des informations précédentes, la procédure affiche les valeurs critiques sur l'ensemble des observations.

Exemple

- recherche de 5 observations atypiques au maximum, au seuil de 10% :

```
outliers alpha=0.10;
```

1.8 L'estimation

L'estimation des paramètres d'un modèle UCM s'effectue en maximisant une log-vraisemblance. L'algorithme utilisé est le filtre de Kalman dont nous donnons ici une brève introduction.

1.8.1 Le filtre de Kalman

Typiquement le filtre de Kalman est utile lorsqu'on est intéressé à l'estimation de la dynamique d'un système, Z_t, qui est seulement observé avec une marge d'erreur. On parle alors de modèle espace-états ou encore de modèle état-mesure. Les deux équations fondamentales sont les suivantes :

1. une équation d'état, décrivant la dynamique du système d'intérêt

$$Z_t = A_t Z_{t-1} + \epsilon_t \quad t = 1, 2, \ldots \tag{1.23}$$

2. une équation de mesure, ou d'observation, reliant les observations d'une variable à l'état du système

$$Y_t = C_t Z_t + \eta_t, \quad t = 1, 2, \ldots \tag{1.24}$$

où Z_t et Y_t sont des vecteurs de dimensions n_Z et n_Y, A_t et C_t des matrices de coefficients déterministes pouvant dépendre du temps, ϵ un vecteur d'innovations constitué de bruits blancs gaussiens, η_t un vecteur d'erreurs de mesure également formé de bruits blancs gaussiens , et où les conditions initiales, Z_0, sont des gaussiennes indépendantes des innovations et des erreurs de mesure.

Dans ce système on se pose les problèmes suivants :

a) Un problème de filtrage : quelle est la meilleure approximation, au sens de l'erreur quadratique moyenne, de Z_t connaissant $Y_0, Y_1, \ldots, Y_t$?

b) Un problème de lissage : quelle est la meilleure approximation, au sens de l'erreur quadratique moyenne, de Z_t connaissant $Y_0, Y_1, \ldots, Y_s$ avec s>t ?

c) Un problème de prévision : quelle est la meilleure approximation, au sens de l'erreur quadratique moyenne, de Z_t connaissant $Y_0, Y_1, \ldots, Y_s$ avec s<t ?

On sait maintenant que les réponse à ces interrogations dans un univers gaussien sont données par les espérances conditionnelles. Il faudra donc évaluer pour chacun des trois problèmes :

a) ${}_tZ_t = E[Z_t|Y_0, Y_1, \ldots, Y_t]$,

b) ${}_{t^+}Z_t = E[Z_t|Y_0, Y_1, \ldots, Y_s]$ avec $s > t$,

c) ${}_{t^-}Z_t = E[Z_t|Y_0, Y_1, \ldots, Y_s]$ avec $s < t$,

On est comme toujours intéressé également par la précision de l'approximation et on devra donc distinguer les différentes MSE

a) ${}_t\Sigma_t = E[(Z_t - {}_t\hat{Z}_t)(Z_t - {}_t\hat{Z}_t)^\top]$,

b) ${}_t\Sigma_t = E[(Z_t - {}_{t^+}\hat{Z}_t)(Z_t - {}_{t^+}\hat{Z}_t)^\top]$,

c) ${}_t\Sigma_t = E[(Z_t - {}_{t^-}\hat{Z}_t)(Z_t - {}_{t^-}\hat{Z}_t)^\top]$,

L'algorithme adapté au calcul de ces approximations est le filtre de Kalman. Il s'agit d'un ensemble d'équations récursives qui va donner les expressions des approximations optimales et de leurs MSE en fonction des observations de la variable de mesure Y. Un des intérêts de la procédure est d'intégrer rapidement toute nouvelle information : chaque nouvelle observation de Y donne lieu à une nouvelle évaluation des approximations optimales. Ces équations de mise à jour sont les suivantes :

- Pour le filtrage :

  a) d'une part

$$ {}_t\hat{Z}_t = {}_{t-1}\hat{Z}_t + K_t[Y_t - C_t\, {}_{t-1}\hat{Z}_t], \text{ avec} \tag{1.25} $$

$$ K_t = {}_{t-1}\Sigma_t\, C_t^\top [C_t\, {}_{t-1}\Sigma_t\, C^\top + \Sigma_\eta]^{-1} \tag{1.26} $$

  a') d'autre part

$$ {}_t\Sigma_t = [I - K_t C_t]\, {}_{t-1}\Sigma_t \tag{1.27} $$

- Pour les prévisions à l'horizon 1 :

  c) pour le système Z

$$ {}_t\hat{Z}_{t+1} = A_t\, {}_t\hat{Z}_t \tag{1.28} $$

  c') pour la MSE associée

$$ {}_t\Sigma_{t+1} = A_t\, {}_t\Sigma_t\, A_t^\top + \Sigma_\epsilon \tag{1.29} $$

- Pour le lissage :

  b) pour le système Z :

$$ {}_T\hat{Z}_t = {}_t\hat{Z}_t + F_t({}_T\hat{Z}_{t+1} - {}_t\hat{Z}_{t+1}) \tag{1.30} $$

avec $F_t = {}_t\Sigma_t\, A_t^\top\, {}_t\Sigma_{t+1}^{-1}$,

  b') et pour la MSE

$$ {}_T\Sigma_t = {}_t\Sigma_t + F_t({}_T\Sigma_{t+1} - {}_t\Sigma_{t+1})F_t^\top \tag{1.31} $$

Alors que pour le filtrage et les prévisions le temps est pris dans son ordre naturel, *i.e.* $t = 1, 2, \ldots, T$, pour le lissage les équations (1.30) et (1.31) sont utilisées en remontant le temps, $t = T, T-1, T-2, \ldots 2, 1$. La démarche est la suivante : avec les équations de filtrage et de prévision on a calculé $\left({}_t\hat{Z}_t,\, {}_t\Sigma_t\right)$ d'une part, $\left({}_{t-1}\hat{Z}_t,\, {}_{t-1}\Sigma_t\right)$ d'autre part pour $t = 1, \ldots, T$. Connaissant $\left({}_T\hat{Z}_T,\, {}_T\Sigma_T\right)$ et $\left({}_{T-1}\hat{Z}_T,\, {}_{T-1}\Sigma_T\right)$, on peut évaluer les équations (1.30) et (1.31) au temps $t = T-1$ pour obtenir respectivement :

$$ {}_T\hat{Z}_{T-1} = {}_{T-1}\hat{Z}_{T-1} + F_{T-1}({}_T\hat{Z}_T - {}_{T-1}\hat{Z}_T), \text{ et} $$
$$ {}_T\Sigma_{T-1} = {}_{T-1}\Sigma_{T-1} + F_{T-1}({}_T\Sigma_T - {}_{T-1}\Sigma_T)F_t^\top $$

Ces deux valeurs remises dans les mêmes équations avec $t = T - 2$ vont permettre de calculer $\left({}_T\hat{Z}_{T-2,T}\, \Sigma_{T-2}\right)$. Il suffit de répéter ces itérations pour finalement connaître l'ensemble des estimations des valeurs lissées $\left({}_T\hat{Z}_{t,T}\, \Sigma_t\right), t = T, T-1, T-2, \ldots, 2, 1$.

On ne va donner ici que la justification [11] de certaines des équations précédentes, les autres sont laissées à titre d'exercice.
Soit tout d'abord $e_{z_t} = Y_t - {}_{t-1}\hat{Y}_t = Y_t - C_{t\,t-1}\hat{Z}_t$ l'erreur de prévision à l'horizon d'une période commise sur la variable de mesure, *i.e.* l'écart entre l'état du système en t et sa valeur anticipée en $t-1$. Il faut se rappeler que dans un univers gaussien, l'espérance conditionnelle de la variable d'intérêt est donnée par la projection de la variable en question sur les variables de conditionnement. Ainsi, comme ${}_{t-1}\hat{Y}_t = E[Y_t|Y_0, Y_1, \ldots, Y_{t-1}]$, cette erreur e_{z_t} est orthogonale à $Y_0, Y_1, \ldots, Y_{t-1}$.

- preuve [12] de la proposition a)

$$\begin{aligned}
{}_t\hat{Z}_t &= E[Z_t|Y_0, Y_1, \ldots, Y_t] \\
&= E[Z_t|Y_0, Y_1, \ldots, Y_{t-1}, e_{z_t}] \\
&= E[Z_t|Y_0, Y_1, \ldots, Y_{t-1}] + E[Z_t|e_{z_t}] - E[Z_t]
\end{aligned}$$

Comme $E[Z_t|e_{z_t}]$ est donnée par le modèle multivarié de régression de Z_t sur l'erreur, on a aussi :

$$\begin{aligned}
E[Z_t|e_{z_t}] &= \frac{Cov(Z_t, e_{z_t})}{Var(e_{z_t})} e_{z_t} + E[Z_t] \\
&= \frac{Cov(Z_t, Y_t - C_{t\,t-1}\hat{Z}_t)}{Var(Y_t - C_{t\,t-1}\hat{Z}_t)} (Y_t - C_{t\,t-1}\hat{Z}_t) + E[Z_t] \\
&= \frac{Cov\left(Z_t, C_t(Z_t - {}_{t-1}\hat{Z}_t) + \eta_t\right)}{Var\left(C_t(Z_t - {}_{t-1}\hat{Z}_t) + \eta_t\right)} (Y_t - C_{t\,t-1}\hat{Z}_t) + E[Z_t] \\
&= {}_{t-1}\Sigma_t C_t^\top \left(C_{t\,t-1}\Sigma_t C_t^\top + \Sigma_\eta\right)^{-1} (Y_t - C_{t\,t-1}\hat{Z}_t) + E[Z_t] \\
&= K_t(Y_t - C_{t\,t-1}\hat{Z}_t) + E[Z_t]
\end{aligned}$$

Si on remplace dans l'expression précédente de ${}_t\hat{Z}_t$ le terme $E[Z_t|e_{z_t}]$ par

11. On suit ici la démonstration donnée dans Gourieroux et Montfort, *Séries Temporelles et modèles dynamiques*, Economica, 1995 ; pp. 569 et suivantes.

12. on va utiliser le résultat suivant : Soit A, B, C des gaussiennes telles que $A \perp B$, alors

$$\begin{aligned}
E[C|A, B] &= \beta_A(A - E[A]) + \beta_B(B - E[B]) + E[C], \\
E[C|A] &= \beta_A(A - E[A]) + E[C], \\
E[C|B] &= \beta_B(B - E[B]) + E[C] \text{ et donc,} \\
E[C|A, B &= E[C|A] + E[C|B] - E[C]
\end{aligned}$$

ce qui vient d'être trouvé, on retrouve bien l'équation (1.25) :

$$ {}_t\hat{Z}_t = {}_{t-1}\hat{Z}_t + K_t(Y_t - C_{t\,t-1}\hat{Z}_t) $$

- preuve de la proposition a') Il s'agit de l'équation qui met à jour la MSE de l'approximation du système en t, connaissant toutes les mesures présentes et passées, soit ${}_t\Sigma_t = E[(Z_t - {}_t\hat{Z}_t)(Z_t - {}_t\hat{Z}_t)^\top]$. On a, d'après le résultat précédent :

$$ \begin{aligned} Z_t - {}_t\hat{Z}_t &= (Z_t - {}_{t-1}\hat{Z}_t - K_t(Y_t - C_{t\,t-1}\hat{Z}_t) \\ &= Z_t - {}_{t-1}\hat{Z}_t - K_t\, e_{z_t}, \text{ soit encore} \\ Z_t - {}_{t-1}\hat{Z}_t &= Z_t - {}_t\hat{Z}_t + K_t\, e_{z_t} \end{aligned} $$

Comme ${}_t\hat{Z}_t = E[Z_t|Y_0, Y_1, \ldots, Y_t]$, l'écart $Z_t - {}_t\hat{Z}_t$ est orthogonal à $Y_0, Y_1, \ldots, Y_t$ et en conséquence orthogonal à l'erreur de prévision e_{z_t}, il vient :

$$ \begin{aligned} {}_t\Sigma_t &= Var(Z_t - {}_t\hat{Z}_t) \\ &= Var(Z_t - {}_{t-1}\hat{Z}_t) - Var(K_t\, e_{z_t}) \\ &= {}_{t-1}\Sigma_t - K_t\, Var(e_{z_t})\, K_t^\top \\ &= {}_{t-1}\Sigma_t - K_t\, Var(C_t(Z_t - {}_{t-1}\hat{Z}_t) + \eta_t)\, K_t^\top \\ &= {}_{t-1}\Sigma_t - K_t\, (C_{t\,t-1}\Sigma_t\, C_t^\top + \Sigma\eta)\, K_t^\top, \text{ soit, avec (1.26)} \\ &= {}_{t-1}\Sigma_t - {}_{t-1}\Sigma_t\, C_t^\top\, [C_{t\,t-1}\Sigma_t\, C^\top + \Sigma_\eta]^{-1}\, C_{t\,t-1}\Sigma_t \\ &= \left[I - K_t\, C_t^\top\right]_{t-1}\Sigma_t \end{aligned} $$

justifiant ainsi l'équation (1.27). Remarquez encore que, ${}_{t-1}\Sigma_t$ et Σ_η étant définies-positives, l'antépénultième équation de la dernière suite d'égalités précédentes d'écriture

$$ {}_t\Sigma_t = {}_{t-1}\Sigma_t - K_t\, (C_{t\,t-1}\Sigma_t\, C_t^\top + \Sigma\eta)\, K_t^\top $$

fait clairement apparaître que la connaissance d'une nouvelle observation de la variable de mesure entraîne une réduction de la MSE filtrée relativement à la MSE anticipée, réduction dont l'ampleur est déterminée par les valeurs de K_t. Ceci explique que cette matrice K_t est appelée gain du filtre à la date t, ou matrice de gain.

1.8.2 La maximisation de la vraisemblance

Dans les développements précédents, les diverses matrices $A_t, C_t, \Sigma_\eta, \Sigma_\epsilon$ sont supposées connues ce qui, en pratique n'est évidemment pas le cas. Soit θ l'ensemble des paramètres inconnus qu'il faut estimer. Pour cela, on va maximiser une vraisemblance de la forme :

$$ L(y;\theta) = \prod_{t=1}^{T} f(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \theta) \tag{1.32} $$

Lorsque les innovations, les erreurs de mesure et les valeurs initiales sont des gaussiennes, alors la distribution conditionnelle de y_t est elle aussi gaussienne avec des moyennes et des covariances données par le filtre de Kalman. Plus précisément les prévisions, à un horizon de une période, de la variable de mesure et de sa matrice de variance-covariance en t sont données par :

$$ {}_{t-1}\hat{Y}_t = C_t \, {}_{t-1}\hat{Z}_t \tag{1.33} $$

$$ {}_{t-1}V(Y_t) = C_t \, {}_{t-1}\Sigma_t \, C_t^\top + \Sigma_\eta \tag{1.34} $$

Si on indice par 0 les valeurs initiales [13], la log-vraisemblance à considérer est alors :

$$ \begin{aligned} l(\theta; Y_0, Y_1, \ldots, Y_T) = & \frac{(T+1)n_y}{2} \log 2\pi - \frac{1}{2} \sum_{t=0}^{T} \log |{}_{t-1}V(Y_t)|^{-1} \\ & - \frac{1}{2} \sum_{t=0}^{T} (Y_t - {}_{t-1}\hat{Y}_t)^\top \, [{}_{t-1}V(Y_t)]^{-1} \, (Y_t - {}_{t-1}\hat{Y}_t) \end{aligned} \tag{1.35} $$

L'estimation est alors effectué au moyen d'un algorithme dénommé EM, pour **E**spérance-**M**aximisation signifiant que l'on commence par calculer des espérances pour des valeurs données des paramètres, espérances qui sont entrées dans la fonction de vraisemblance. Celle-ci est alors maximisée, ce qui donne de nouvelles valeurs aux paramètres. On calcule les nouvelles espérances et le processus est itéré sur la totalité des observations. Le schéma de la figure 1.12 décrit cette procédure EM dans le cas qui nous intéresse ici.

Lorsque l'optimum a été trouvé, la procédure affiche les valeurs estimées des paramètres ainsi que leurs écarts-types asymptotiques et des tests de nullité qui d'après l'aide de SAS doivent être regardés avec prudence et cela tout particulièrement lorsque l'hypothèse nulle porte sur une borne de l'espace de définition du paramètre en question [14]. Enfin, si le modèle ajusté est satisfaisant sa composante résiduelle doit posséder les caractéristiques d'un bruit blanc, et on peut retrouver en sortie les outils usuels permettant d'étudier cet aspect : histogramme, ACF, PACF, Q-Q plot et test de Ljung-Box.

Par ailleurs, les équations du filtre de Kalman sont à nouveau utilisées pour calculer

- les valeurs filtrées de Z correspondant à ${}_t\hat{Z}_t = E[Z_t | Y_0, Y_1, \ldots, Y_t; \hat{\theta}]$, pour $t = 1, 2, \ldots, T$ et celles de ses composantes (trend, cycle,...)
- les valeurs lissées de Z données par ${}_T\hat{Z}_t = E[Z_t | Y_0, Y_1, \ldots, Y_T; \hat{\theta}]$, pour $t = 1, 2, \ldots, T$ ainsi que celles de ses composantes. Ces valeurs qui à

13. On ne détaille pas ici l'étape d'initialisation qui maximise une vraisemblance dite *diffuse* sur les observations du début de l'échantillon, par exemple les n_0 premières, et l'étape de post-initialisation qui maximise la vraisemblance usuelle sur les $T - n_0$ observations suivantes.

14. C'est par exemple le cas lorsque le test porte sur la longueur d'un cycle.

chaque date t utilisent la totalité des observations de la variable de mesure sont effectivement toujours plus "lisses" que les précédentes qui elles sont fondées à chaque date t uniquement sur les observations de Y connues jusqu'en t.

Pour chaque composante, le graphe de ces valeurs est obtenu en utilisant les options `plot=filter`, `plot=smooth` ou encore `plot=(filter smooth)` dans la commande de création de la composante en question. On pourra par exemple utiliser

```
cycle plot=smooth;
ou
level plot=(filter smooth);
```

1.8.3 Les options de la commande `Estimate`

- `BACK`=n ou `SKIPLAST`=n : cette option joue le même rôle que dans la proc ARIMA. L'estimation des paramètres s'effectue en ignorant les n dernières observations. Elle est souvent utilisée en conjonction avec les options back=n lead=n de la commande `forecast` pour juger de la capacité du modèle retenu à prévoir les n dernières observations connues. Par défaut $n = 0$: toutes les observations sont utilisées dans la phase d'estimation.

- `PLOT`= : l'utilisateur peut spécifier un ou plusieurs graphiques, par exemple, `plot=(pacf acf histogram QQ residual model WN)` dont les noms vous sont connus. Une version intéressante est `plot=panel` qui renvoie l'ensemble des graphes histogram, QQ acf et pacf afférents à la série des résidus empiriques.

- `OUTEST`= : permet de donner le nom d'une table qui contiendra les valeurs des paramètres et leur écart-types estimés, les t-statistiques associées ainsi que d'autres informations (paramètre libre ou non, convergence de l'algorithme d'estimation satisfaite ou non,...)

1.9 Les prévisions

Le modèle estimé peut ensuite être utilisé pour construire des prévisions, les grapher ou/et les sauvegarder. On mobilise pour cela la commande `FORECAST` et ses diverses options parmi lesquelles :

- `BACK`=n_1 ou SKIPLAST=n_1 : les n_1 dernières observations ne sont pas utilisées pour la construction des prévisions. Cette option permet, combinée avec `LEAD`=, d'obtenir en sortie une table comparant les n_1 dernières valeurs avec leurs prévisions qui sont alors dynamiques ou multistep, *i.e.* construites avec une information s'arrêtant en $T - n_1$. Par défaut, $n_1 = 0$.

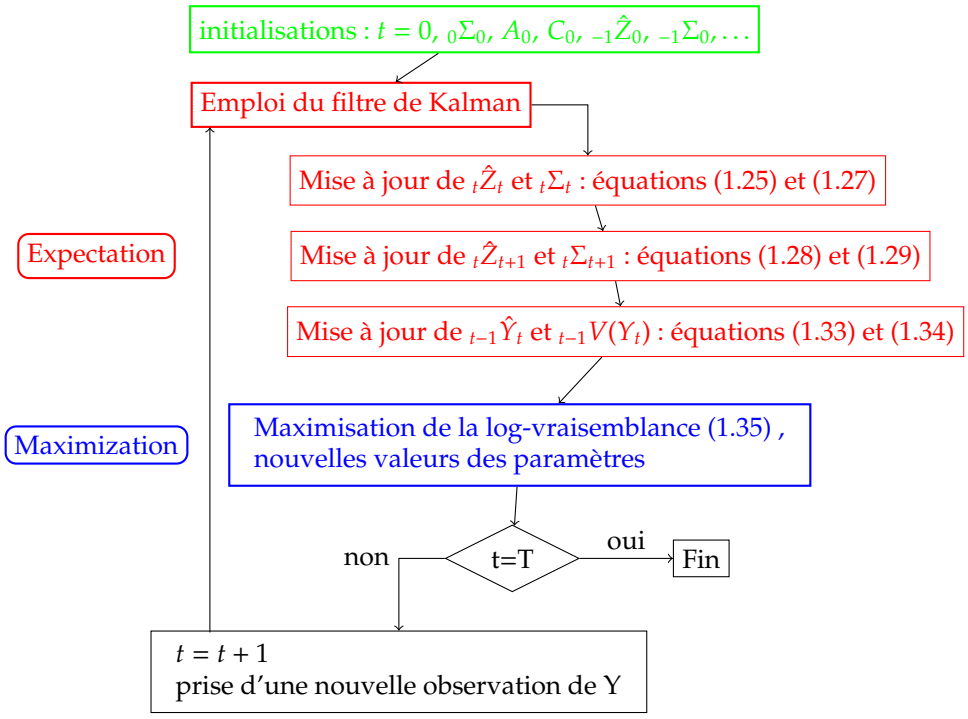

FIGURE 1.12 – Maximisation de la vraisemblance avec l'algorithme EM

- `LEAD`=n_2 : précise le nombre de périodes pour lequel on demande la construction de prévisions, ici n_2, cette fenêtre de prévisions débutant à l'observation précisée par `BACK`=n_1 ou SKIPLAST=n_1.

- `ALPHA`= α : ici α est le seuil de risque que l'on désire utiliser pour la construction des intervalles de confiance autour des prévisions qui seront à $100(1-\alpha)\%$. Par défaut $\alpha = 0.05$ et donc des IC à 95%.

- `PLOT`= : permet d'obtenir les graphes des prévisions avec `PLOT=FORECASTS`, et celles des estimations lissées du trend avec `PLOT=TREND`, des autres composantes avec `PLOT=DECOMP` et de leurs variances avec `PLOT=DECOMPVAR`. Les graphes des estimations filtrées des composantes et de leurs variances sont créés avec `PLOT=FDECOMP` et `PLOT=FDECOMPVAR`.

- `PRINT`= : gère l'impression des valeurs lissées ou filtrées estimées ainsi que les prévisions. Avec `PRINT=DECOMP` on demande l'impression des valeurs lissées du trend, du trend augmenté de la composante régression si elle est présente, de l'ensemble trend+régression+cycle et enfin de la somme de toutes les composantes à l'exception de la composante résiduelle. Les observations filtrées correspondantes sont imprimées avec `PRINT=FDECOMP` et les prévisions avec `PRINT=FORECASTS`. On peut supprimer toutes les impressions précédentes avec `PRINT=NONE`.

- `OUTFOR`= : permet de donner le nom d'une table qui sera créée et contiendra les erreurs de prévisions à l'horizon d'une période sur la fenêtre d'estimation, les prévisions et leur intervalle de confiance sur la fenêtre de prévision, les valeurs lissées de toutes les composantes et leur somme.

- `BOOTSTRAP(NREP=integer <SEED=integer>)`. Option encore expérimentale dans SAS 9.4 qui permet d'obtenir des intervalles de confiance autour des prévisions à partir de NREP échantillons bootstrappés et prise des fractiles sur les NREP séries de prévisions associées à ces échantillons. Sans cette option, les intervalles de confiance sont construits, comme dans la procédure ARIMA, en supposant que les coefficients estimés sont les vraies valeurs des paramètres. La technique du bootstrap introduit les variations des paramètres estimés dans la construction des IC et devrait donc donner à ceux-ci des étendues plus proches de leurs vraies valeurs [15]. L'entier correspondant à SEED, par défaut 123, permet d'initialiser le générateur de nombre au hasard et surtout de répéter les résultats obtenus.

On aura par exemple :

- Si on veut apprécier la capacité du modèle à prévoir en dynamique les 12 dernières observations, celles-ci n'étant pas utilisées pour l'estimation des paramètres [16], on aura :

15. Cette option devrait vous être compréhensible après le cours "Boostrap et simulations" du second semestre

16. Les fenêtres d'estimation et de prévision s'étendent donc respectivement sur les plages d'observations (1, T-12) et (T-11, T).

```
estimate back=12;
forecast lead=12;
```

- Sur données mensuelles, demande de création de prévisions pour les 2 années suivant la fin de la période d'observation avec :

  ```
  forecast lead=24;
  ```

- graphe des prévisions de la série étudiée et de ses composantes lissées :

  ```
  forecast lead=12 plot=(forecasts decomp);
  ```

Chapitre 2

Un exemple de construction d'un modèle UCM

2.1 Rappel de la liste des paramètres

La table 2.1 rappelle quelques-uns des paramètres afférents aux diverses composantes d'un modèle UCM. On se souvient qu'en forçant à zéro une variance, on contraint la composante concernée à être non aléatoire.

On rappelle qu'en plus de ceux indiqués dans cette table, on peut également trouver

- les coefficients des éventuelles variables explicatives listées dans la commande `Model`,
- ceux du processus autorégressif éventuel afférent à la variable expliquée qui peut être précisé dans la commande `Deplag`,
- ceux de l'éventuel processus SARMA, $\Phi(L^s)\phi(L)y_t = \Theta(L^s)\theta(L)\epsilon_t$ imposé sur le résidu du modèle via la commande `Irregular`.

2.2 La production industrielle de l'industrie chimique

Pour illustrer la construction d'un modèle UCM, on utilise l'indice brut de la production industrielle de l'industrie chimique observé mensuellement de janvier 2000 à juillet 2017. Cette série est présentée dans le graphe 2.1. Dans notre table, nous avons en plus de la variable "production" contenant ces observations une variable contenant leur date de réalisation.

composante	paramètre	signification	equation
Trend	σ^2_η	variance du trend	(1.3) ou (1.4)
Pente	σ^2_ξ	variance de la pente	(1.5)
Cycle	σ^2_v	variance	
Cycle		d'un cycle périodique	(1.17)
Cycle		ou d'une composante AR(1) inobservée	(1.18)
Cycle	ρ	coefficient autorégressif (*damping factor*)	
Cycle		d'un cycle périodique	(1.17)
Cycle		ou d'une composante AR(1) inobservée	(1.18)
Saisonnalité	σ^2_ω	variance des indicatrices	
Saisonnalité		de type dummy variables	(1.19)
Saisonnalité		ou tirées de fonction périodiques	(1.21)
Explicatives à coefficient variable	σ^2_η	variance des variations du coefficient	(1.22)

TABLE 2.1 – récapitulatif des paramètres d'un modèle UCM

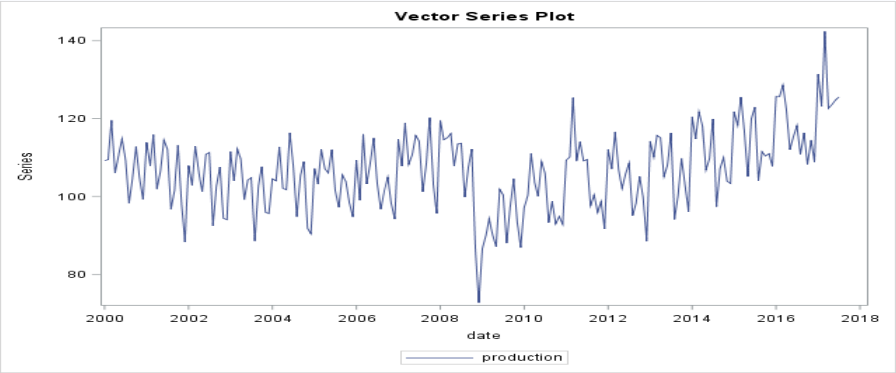

FIGURE 2.1 – production de l'industrie chimique, janvier 2000-juillet 2017

Outlier Summary							
Obs	date	Break Type	Estimate	Standard Error	Chi-Square	DF	Pr > ChiSq
107	NOV2008	Level	-19.73532	3.7712252	27.39	1	<.0001
108	DEC2008	Level	-17.13492	3.771095	20.65	1	<.0001
130	OCT2010	Additive Outlier	-10.12577	3.4938113	8.40	1	0.0038

TABLE 2.2 – Résultats de la recherche d'outliers, option `checkbreak`

2.3 Modèle de base et repérage des outliers

On conseille généralement de partir d'un modèle basique ayant le minimum de composantes et de l'enrichir par étapes. Ce modèle de base peut contenir par exemple un trend, une saisonnalité et une composante irrégulière. Sur données mensuelles, on peut ainsi partir avec :

```
proc ucm data=chimie;
  id date interval=month;
  irregular;
  level checkbreak;
  slope;
  season type=dummy lenght=12;
  model production;
  estimate;
run;
```

L'ajout de l'option `checkbreak` réclame la recherche d'observations atypiques, qui peuvent affecter la pente du trend ou simplement une observation particulière. Les résultats de cette recherche sont donnés dans la table 2.2.
Un changement de pente est donc repéré à la fin de l'année 2008, ce qui semble raisonnable à la vue du graphique 2.1. Pour cette raison, nous travaillerons par la suite sur la période janvier 2009-juillet 2017 avec les observations représentées dans le graphe 2.2.
Par ailleurs, l'observation d'octobre 2010 dévie significativement de sa valeur prévue. La correction usuelle est de créer une variable indicatrice via une étape data et de l'introduire comme explicative dans le modèle.
Cette première étape se termine donc avec :

```
data chimie;
  set chimie;
  where date>"31dec2008"d;
  ao_10_2010=(month(date)=10 and year(date)=2010);
  run;
proc ucm data=chimie;
  id date interval=month;
  irregular;
  level;
```

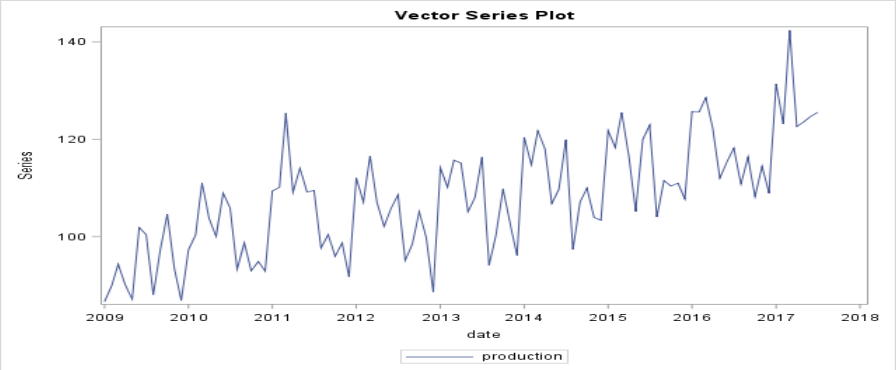

FIGURE 2.2 – production de l'industrie chimique, janvier 2009-juillet 2017

Final Estimates of the Free Parameters					
Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	0.00060825	0.0035060	0.17	0.8623
Level	Error Variance	4.18011	1.29424	3.23	0.0012
Slope	Error Variance	0.00000256	.	.	.
Season	Error Variance	4.58748	1.14236	4.02	<.0001
ao_10_2010	Coefficient	-8.91652	3.27067	-2.73	0.0064

TABLE 2.3 – Estimations des paramètres libres, janvier 2009 - juillet 2017
Modèle M0

```
  slope;
  season type=dummy lenght=12;
  model production = ao_10_2010;
  estimate;
run;
```

Les paramètres à estimer dans ce premier modèle, nommé M0, sont alors dans l'ordre d'apparition des commandes du programme précédent σ^2_ϵ, σ^2_η, σ^2_ξ, σ^2_ω et β le coefficient de l'indicatrice a0_10_2010. Après exécution, on récupère la table 2.3.

On remarque des variances significatives pour les composantes *level* et *saisonnalité* signalant que ces composantes sont plutôt de type aléatoire [1]. Le coefficient de l'indicatrice est par ailleurs significativement différent de zéro. Une difficulté apparaît toutefois avec la composante slope du trend : l'algorithme n'a pas été en mesure de calculer l'écart-type de sa variance estimée. Cette dernière étant proche de zéro, nous décidons de traiter la pente comme étant une constante, i.e. d'imposer un trend déterministe ce qui ne semble pas déraison-

1. On se souvient que tout modèle doit a priori comporter une composante résiduelle aléatoire : la question de la non significativité ou de la significativité de la variance résiduelle ne se pose même pas, et cette composante sera présente même si sa variance est non significative.

Final Estimates of the Free Parameters					
Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	0.00313	0.15566	0.02	0.9840
Level	Error Variance	4.17924	1.29532	3.23	0.0013
Season	Error Variance	4.58630	1.14407	4.01	<.0001
ao_10_2010	Coefficient	-8.91708	3.27091	-2.73	0.0064

TABLE 2.4 – Estimations des paramètres libres, janvier 2009 - juillet 2017
Modèle M1

nable à la vue du graphe 2.2.
Les lignes de commandes correspondant à ce nouveau modèle, M1, sont alors :

```
proc ucm data=chimie;
  id date interval=month;
  irregular;
  level plot=smooth;
  slope var=0 noest plot=smooth;
  season type=dummy lenght=12 plot=smooth;
  model production = ao_10_2010;
  estimate plot=panel;
run;
```

Les estimations obtenues sont présentées dans la table 2.4. D'après celles-ci, les paramètres σ^2_η et σ^2_ω sont significatifs, ce qui plaide en faveur du caractère aléatoire et non pas déterministe des composantes de trend et de saisonnalité. Par ailleurs, le coefficient de l'indicatrice traitant le cas de l'observation atypique est également significatif.
Dans la table 2.8 récapitulative des modèles ajustés on peut lire les valeurs des critères de sélection d'Akaike, AIC, et de Schwarz, BIC. Les deux donnent la préférence au modèle M1 lorsqu'on le compare à M0. On retiendra donc un trend déterministe dans les modélisations ultérieures.

Les graphes 2.4, 2.5 et 2.6 permettent de visualiser les composantes *Level*, *Slope* et *Season* de ce modèle M1 estimées sur la totalité des observations. La pente étant contrainte à être constante, le second graphique n'a évidemment qu'un intérêt fort limité. La valeur estimée de cette pente apparaît dans la table 2.5 qui fournit les valeurs des composantes Level et Slope à la fin de la période d'estimation, respectivement 123.5 et 0.384. On notera que le caractère aléatoire des variables saisonnières autorise des variations dans leurs estimations au cours du temps : leurs profils annuels ne se répètent pas à l'identique. Dans ces sorties on trouve également la table 2.6 qui délivre des informations sur la significativité des différentes composantes du modèle à la fin de la fenêtre d'estimation. Enfin, l'option `plot=panel` de la commande `estimate` permet de récupérer en sortie dans le graphe 2.3 les indications habituelles sur les résidus du modèle : histogramme, q-q plot, auto-corrélations et autocorrélations par-

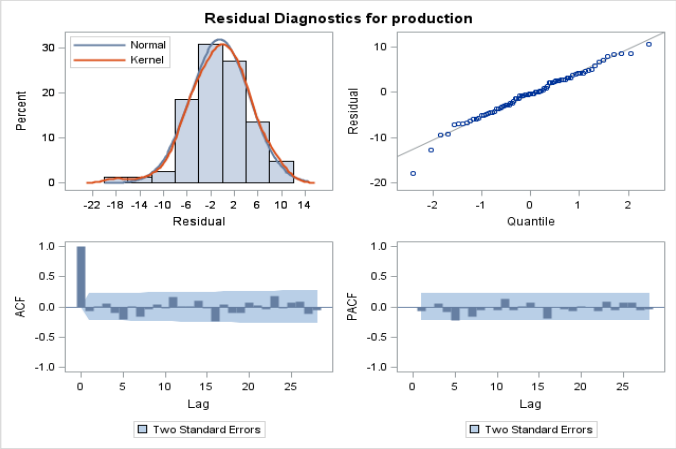

FIGURE 2.3 – Caractéristiques des résidus
Modèle M1

Trend Information (Based on the Final State)		
Name	Estimate	Standard Error
Level	123.5174277	1.8632835
Slope	0.383661292	0.2041109

TABLE 2.5 – Estimations du niveau et de la pente en juillet 2017 - Modèle M1

tielles. On voit qu'aucune anomalie majeure n'est repérée : le modèle M1 peut donc servir de base pour l'étude de variantes qui pourraient éventuellement l'améliorer.

2.4 Étude de variations du modèle initial

Disposant d'un modèle de base, nous pouvons chercher à l'améliorer en envisageant quelques variantes. Dans ce qui suit on indique la ligne de commande qui se substitue ou s'ajoute à celles du modèle M1 pour créer ces variations.

Significance Analysis of Components (Based on the Final State)			
Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	0.00	0.9937
Level	1	4394.39	<.0001
Slope	1	3.53	0.0602
Season	11	411.44	<.0001

TABLE 2.6 – Significativité des paramètres du modèle sur la fin de la période d'observation- Modèle M1

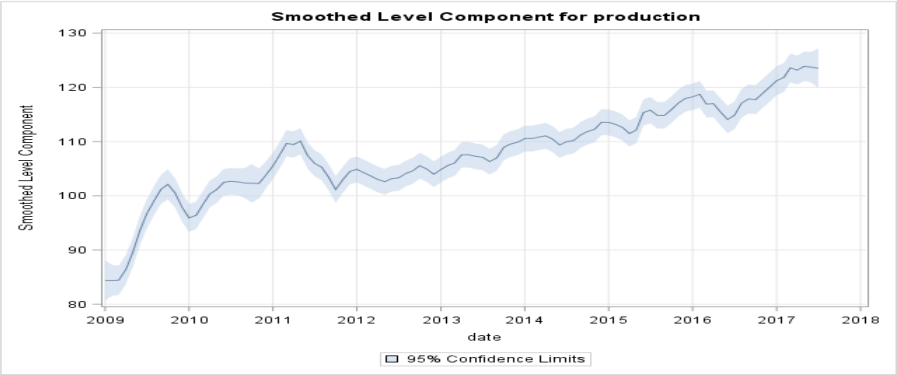

FIGURE 2.4 – Modèle M1 - Trend lissé estimé

FIGURE 2.5 – Modèle M1 - Estimation de la pente lissée du trend

Pour une variante donnée, l'absence de commentaires signifie que le modèle résultant est dominé au regard des deux critères par M1. La table 2.8 récapitule les diverses modélisations envisagées ainsi que les valeurs prises par les deux critères AIC et BIC.

1. M2 : Variables de saisonnalité de type trigonométriques :

   ```
   season length=12 type=trig;
   ```

2. M3 : Variables de saisonnalité déterministes :

   ```
   season length=12 type=dummy variance=0 noest;
   ```

3. M4 : Introduction d'un cycle stochastique :

   ```
   cycle;
   ```

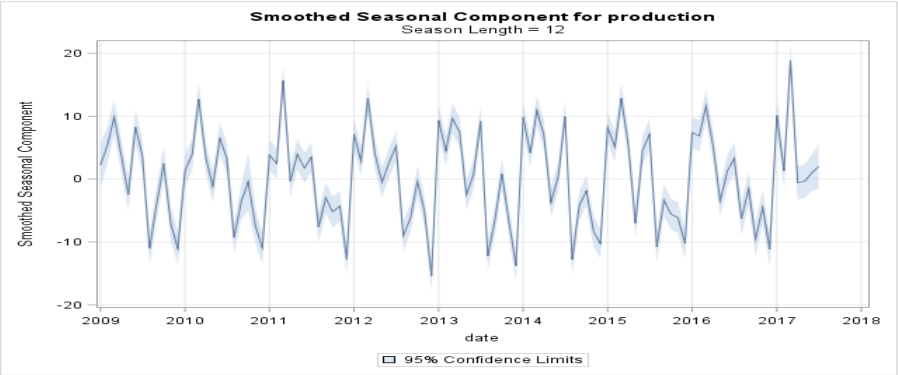

FIGURE 2.6 – Modèle M1 - Estimation de la saisonnalité lissée

Une contradiction apparaît : AIC privilégie M4 alors que BIC ferait retenir M1. Pour les départager, on peut imaginer de regarder lequel produit les meilleures prévisions sur les 12 derniers mois observés (août 2016-juillet 2017), ces douze derniers mois étant exclus de la période d'estimation. Pour ce faire, il suffit d'intégrer les commandes `estimate` et `forecast` suivantes :

```
estimate back=12;
forecast back=12 lead=12;
```

La première retire les 12 derniers mois de l'échantillon de la fenêtre d'estimation, la seconde demande la construction de 12 prévisions dynamiques commençant 12 mois avant la fin de l'échantillon, *i.e.* elle définit la fenêtre de prévision août 2016-juillet 2017. Sur cette sous-période, la RMSE et la MAE des erreurs de prévisions sont égales respectivement à 8.91 et 6.38 pour M1 et à 6.38 et 5.25 pour M4. Compte-tenu de ces chiffres, nous allons privilégier M4.
On peut toutefois remarquer qu'avec ce modèle on obtient aussi la table 2.7 renseignant sur la significativité des diverses composantes à la fin de la période d'estimation et faisant ressortir la non significativité de la composante *Cycle*. Il n'est donc pas assuré que les prévisions réalisées au moyen de M4 à partir de juillet 2017 soient très différentes de celles que nous obtiendrions à partir de M1.

4. M5 : Introduction d'un cycle déterministe :

   ```
   cycle variance=0 noest=variance;
   ```

5. M6 : Introduction d'un deuxième cycle stochastique :

Significance Analysis of Components (Based on the Final State)			
Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	0.00	0.9670
Level	1	5078.00	<.0001
Slope	1	8.66	0.0033
Cycle	2	1.07	0.5854
Season	11	461.16	<.0001

TABLE 2.7 – Modèle M4 - Significativité des composantes

Modèle	Paramètres	Variante	AIC	BIC
M0	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\xi, \sigma^2_\omega, \beta$	pente aléatoire	550.3	560.25
M1	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta$	pente constante	548.3	555.77
	variantes de M1			
M2	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta$	(season) type=trig	562.9	570.36
M3	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta$	(season) type=dummy variance=0 noest	553.8	558.78
M4	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta, \sigma^2_v, \rho$	cycle	544.48	559.41
M5	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta, \rho$	(cycle) variance=0 noest=variance ;	552.3	564.74
M6	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta, \sigma^2_{v_1}, \rho_1, \sigma^2_{v_2}, \rho_2$	cycle cycle	550.05	572.45
M7	$\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\omega, \beta, \sigma^2_v, \rho$	autoreg	550.73	563.18

TABLE 2.8 – Récapitulatif des variantes de modélisation

```
cycle;
cycle;
```

6. M7 : Introduction d'un cycle de type AR(1) :

```
autoreg;
```

2.5 La construction des prévisions

Avant de passer à l'étape de calcul des prévisions, nous allons effectuer une dernière correction sur le modèle utilisé. Celle-ci trouve son origine dans l'examen du graphe 2.7. Il n'est pas déraisonnable de penser que la valeur observée en mars 2017, supérieure à 140, est un outlier. En conséquence, sachant que dans M4 la composante saisonnière est aléatoire, la variable afférente au mois de mars va prendre une valeur anormalement élevée pour mars 2017, ce que confirme le graphe 2.8.

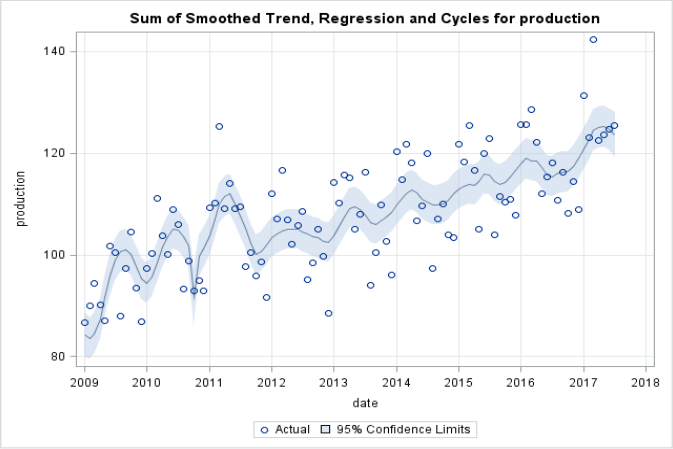

FIGURE 2.7 – Modèle M4 - Somme des estimations lissées du Trend, du Cycle et de la composante Régression

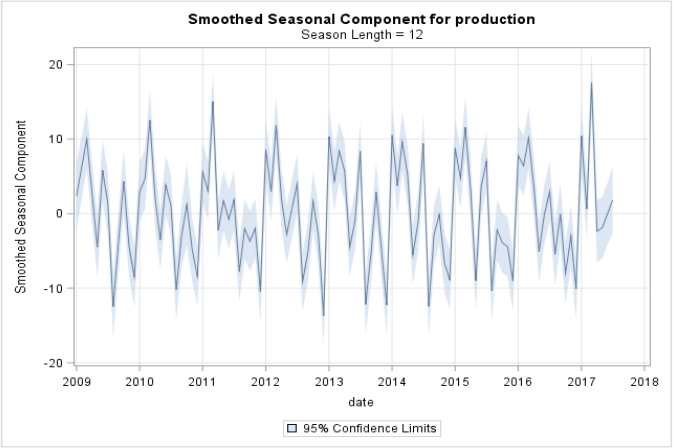

FIGURE 2.8 – Modèle M4 - Estimation lissée de la composante saisonnière

2.6 Ajustement de la composante saisonnière du modèle sélectionné

En l'absence de correction, comme les prévisions de cette composante saisonnière postérieures à la date de fin des observations se contentent de reproduire les douze dernières valeurs estimées, la valeur anormale serait reproduite en mars 2018, mars 2019,...Bien évidemment avec des prévisions dynamiques c'est la totalité des prévisions qui est alors affectée.
Pour éviter cela, on procède à une correction usuelle : création d'une variable, a0_03_2017 égale à l'unité en mars 2017 et à zéro pour toute les autres dates, variable utilisée comme explicative dans la composante de régression. Nous avons donc une étape `data` avec une commande de la forme :

```
ao_3_2017=(month(date)=3 and year(date)=2017);
```

suivie de

```
proc ucm data=chimie;
  id date interval=month;
  irregular;
  level;
  slope var=0 noest;
  season length=12 type=dummy plot=smooth;
  cycle;
  model production = ao_10_2010 ao_3_2017;
  estimate;
run;
```

Le coefficient de cette nouvelle variable estimé à 9.12552 avec un écart-type de 3.93425 est significatif au seuil de 5%. Surtout, comme le montre le graphique 2.9, elle permet de redonner à l'indicatrice du mois de mars 2017 une valeur non atypique.

2.7 Traitement pour présence d'explicatives exogènes dans la composante de régression

Dans un modèle UCM, la prévision de la variable d'intérêt est obtenue en calculant des prévisions sur chacune de ces composantes au moyen des équations qui les gouvernent, selon une mécanique semblable à celle employée sur les modèles ARIMA, puis à les ajouter conformément à l'équation de base des modèles UCM. Une fois les paramètres de ces équations estimés, les calculs ne posent pas de difficultés majeures. En revanche si une composante de régression est présente dans le modèle avec des explicatives supposées exogènes, cette mécanique s'enraye : les valeurs futures de l'expliquée dépendent alors de valeurs futures des explicatives qui ne sont pas modélisées par l'UCM. En d'autres termes, l'utilisateur doit préciser les trajectoires des exogènes sur la

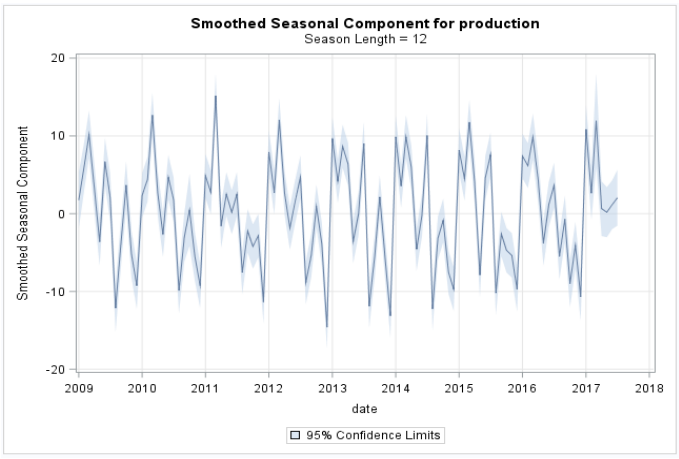

FIGURE 2.9 – Modèle M4 - Composante saisonnière avec correctif mars 2017

fenêtre de prévision.
Dans notre exemple les choses sont très simples : les deux explicatives, ao_10_2010 et ao_3_2017, sont des indicatrices égales à l'unité au mois d'octobre 2010 pour la première et en mars 2017 pour la seconde. Elles doivent être nulles pour toute autre date.

Ainsi, supposons que l'on veuille produire deux années de prévisions à partir d'août 2017, il conviendra de transmettre à la proc `UCM` une table semblable à celle employée pour les estimations du modèle sur la fenêtre janvier 2009-juillet 2017 mais prolongée de 24 observations égales à zéro pour les deux indicatrices. On aura par exemple, si *chimie* est le nom de la table initiale :

```
data ajout(drop=i);
  do i=1 to 24;
    date=intnx('month','01jul2017'd,i);
    ao_10_2010=0;
    ao_3_2017=0;
    output;
    end;
run;
data chimie;
  set chimie ajout;
run;
```

2.8 Construction des prévisions

Avec notre exemple, le programme précédent sera simplement poursuivi par :

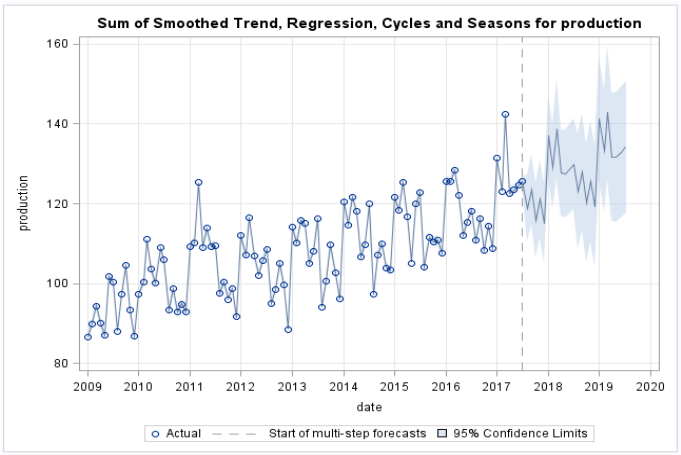

FIGURE 2.10 – production de l'industrie chimique

```
proc ucm data=chimie;
id date interval=month;
irregular;
level plot=smooth;
slope var=0 noest;
season length=12 type=dummy plot=smooth;
cycle plot=smooth;
model production = ao_10_2010 ao_3_2017;
estimate plot=panel;
forecast lead=24 plot=forecasts;
run;
```

On récupèrera alors notamment le graphe 2.10, présentant la production de l'industrie chimique observée jusqu'en juillet 2017 et prévue jusqu'en juillet 2019 avec son intervalle de confiance par défaut à 95%. Naturellement ces valeurs peuvent être imprimées ou/et sauvegardées.