Réponses et commentaires aux TP 5 et 6

TP5

1. Estimation du saisonnier

   (a) La commande *monthplot* permet d'observer les valeurs du saisonnier, mais aussi l'évolution par an vis-à-vis de la moyenne qui estime le saisonnier. On remarque ainsi que tous les mois exhibent le même type évolution par rapport à cette moyenne. Ceci s'explique par la tendance qui est pour l'instant est présente dans ce signal (notez que l'évolution par an vis-à-vis de la moyenne qui estime le saisonnier est proche de la tendance estimée plus loin).

   (b) *s$remainder* représente la série corrigée des variations saisonnières.

   (c) Pour choisir le paramètre de lissage du saisonnier on peut se référer aux deux limites : s'il n'y a pas de lissage (voir plus haut), le saisonnier estimé est entaché de discontinuités qui peuvent être liées aux erreurs d'estimation (visibles notamment lorsque l'on estime sur une dizaine d'année); si on lisse trop, on peut revenir à un saisonnier totalement nul. En conséquence, on voudra conserver l'allure globale du saisonnier estimé en évitant les brutales discontinuités (cela dit, rien ne nous garantit que sur des données réelles, pour lesquelles l'hypothèse de stationnarité du saisonnier peut aussi être mise en doute, il n'y ait pas en réalité de brutales discontinuités; on préférera cependant toujours un peu lisser notamment lorsqu'il y a comme ici peu d'années...). En pratique, ici on choisira la valeur *s.window* $\simeq 6$. En conclusion, il appert qu'il y a plus e déménagement en été qu'en hiver.

2. Exercice

   (a) Pour choisir entre la séquence estimation de tendance puis du saisonnier, et estimation du saisonnier puis de la tendance, plusieurs facteurs interviennent. Si la saisonnalité se détache clairement du bruit (ce qui est le cas ici), on peut commencer par estimer le saisonnier. Sinon on préférera commencer par estimer la tendance. Deuxième facteur, s'il y a beaucoup de données, la tendance globale (de type linéaire) du signal pourra être estimer directement sur la série initiale, puis sur la série détendancialisée, l'estimation du saisonnier sera meilleure que si elle avait été faite initialement. Rien empêche alors d'estimer sur le bruit restant une tendance plus locale.

3. Introduction aux processus ARMA

   (a) On a $X^{(1)}$ qui vérifie $X_n^{(1)} = 0.6X_{n-1}^{(1)} + \varepsilon_n$. C'est un processus AR(1). On obtient facilement que $r(k) = 0.6r(k-1)$ pour $k \geq 1$ et $\mathbb{E}\{(X_n^{(1)})^2\} = r(0) = 0.36r(0) + 1$. En conséquence,

   $$r(n) = \frac{25}{16}(0.6)^n \quad \text{et} \quad \rho(n) = \frac{r(n)}{r(0)} = (0.6)^n.$$

   En comparant ces corrélations théoriques avec les résultats du corrélogramme, on trouve à peu près les mêmes valeurs pour $\rho(1)$ et $\rho(2)$.

   (b) On a $X^{(2)}$ qui vérifie $X_n^{(2)} = \varepsilon_n - 0.6\varepsilon_{n-1}$. C'est un processus MA(1). On obtient facilement que $r(0) = 1.36$, $r(1) = -0.6$ et $r(k) = 0$ pour $|k| \geq 2$. Ainsi,

   $$\rho(1) = -\frac{15}{34} \quad \text{et} \quad \rho(k) = 0 \quad \text{pour} \quad |k| \geq 2.$$

En comparant ces corrélations théoriques avec les résultats du corrélogramme, on trouve à peu près la même valeur pour $\rho(1)$, puis les $\rho(n)$ sont (en général) entre les deux barres permettant de les considérer comme nulles pour $n \geq 2$. Il peut arriver que certaines corrélations soient légèrement hors de ces barres, mais on pourra tout de même les considérer comme nulles. En général, sauf s'il y a vraiment beaucoup de données, il y a peu de sens à estimer des corrélations pour des écarts temporels supérieurs à 10.

(c) Enfin, on a $X^{(3)}$ qui vérifie $X_n^{(3)} = 0.1X_{n-1}^{(3)} + 0.2X_{n-2}^{(3)} + \varepsilon_n - \varepsilon_{n-1} - \varepsilon_{n-2}$. C'est un processus ARMA(2,2). On obtient facilement (équation de Yule-Walker) que $r(k) = 0.1r(k-1) + 0.2r(k)$ pour $k \geq 3$, soit $r(k) = a(0.5)^n + b(-0.4)^n$ avec $(a,b) \in \mathbb{R}^2$. Pour déterminer les valeurs de $r(0)$, $r(1)$ et $r(2)$, on utilise les corrélations entre $X_n^{(3)}$ et les ε_n. En effet, d'après l'équation initiale, on a :

$$\begin{aligned} X_n^{(3)} &= 0.1X_{n-1}^{(3)} + 0.2X_{n-2}^{(3)} + \varepsilon_n - \varepsilon_{n-1} - \varepsilon_{n-2} \\ &= 0.21X_{n-2}^{(3)} + 0.02X_{n-3}^{(3)} + \varepsilon_n - 0.9\varepsilon_{n-1} - 1.1\varepsilon_{n-2} - 0.1\varepsilon_{n-3} \\ &= 0.041X_{n-3}^{(3)} + 0.042X_{n-4}^{(3)} + \varepsilon_n - 0.9\varepsilon_{n-1} - 0.89\varepsilon_{n-2} - 0.31\varepsilon_{n-3} - 0.21\varepsilon_{n-4}, \end{aligned}$$

pour $n \in \mathbb{Z}$. Par suite,

$$\begin{aligned} \mathbb{E}\{X_n^{(3)}\varepsilon_n\} &= 0+0+1+0+0 = 1 \\ \mathbb{E}\{X_n^{(3)}\varepsilon_{n-1}\} &= 0+0+0-0.9+0+0 = -0.9 \\ \mathbb{E}\{X_n^{(3)}\varepsilon_{n-2}\} &= 0+0+0+0-0.89+0+0 = -0.89, \end{aligned}$$

en utilisant le fait que les ε sont indépendants et que X_n ne dépend que de tous les ε_k avec $k \leq n$. Alors,

$$\begin{aligned} r(0) = \mathbb{E}\{X_n^{(3)}\{X_n^{(3)}\} &= \mathbb{E}\{X_n^{(3)}(0.1X_{n-1}^{(3)} + 0.2X_{n-2}^{(3)} + \varepsilon_n - \varepsilon_{n-1} - \varepsilon_{n-2})\} \\ &= 0.1r(1) + 0.2r(2) + 1 + 0.9 + 0.89 = 0.1r(1) + 0.2r(2) + 2.79, \\ r(1) = \mathbb{E}\{X_n^{(3)}\{X_{n-1}^{(3)}\} &= \mathbb{E}\{X_{n-1}^{(3)}(0.1X_{n-1}^{(3)} + 0.2X_{n-2}^{(3)} + \varepsilon_n - \varepsilon_{n-1} - \varepsilon_{n-2})\} \\ &= 0.1r(0) + 0.2r(1) + 0 - 1 + 0.9 = 0.1r(0) + 0.2r(1) - 0.1 \\ r(2) = \mathbb{E}\{X_n^{(3)}\{X_{n-2}^{(3)}\} &= \mathbb{E}\{X_{n-2}^{(3)}(0.1X_{n-1}^{(3)} + 0.2X_{n-2}^{(3)} + \varepsilon_n - \varepsilon_{n-1} - \varepsilon_{n-2})\} \\ &= 0.1r(1) + 0.2r(0) + 0 + 0 - 1 = 0.1r(1) + 0.2r(0) - 1 \end{aligned}$$

A partir des trois équations précédentes, on trouve : $r(0) \simeq 2.72$, $r(1) \simeq 0.22$ et $r(2) \simeq -0.43$. En comparant ces corrélations théoriques avec les résultats du corrélogramme, on trouve à peu près les mêmes valeurs pour $\rho(1)$ et $\rho(2)$.

(d) Le modèle simulé est un ARMA(1,1) d'équation : $X_n = 0.3X_{n-1} + \varepsilon_n - 0.9\varepsilon_{n-1}$. Pour simuler un ARMA(2,2), il faut dans $list$ mettre $ar = c(0.3, 0.7)$ et $ma = c(0.9, -04)$ (par exemple). Enfin, pour simuler avec un bruit uniforme, il faut créer auparavant une fonction telle que $rand2\_function(n)$ $runif(100, min = -1, max = 1)$ et rajouter ce bruit dans $arima \sim (..., rand.gen = rand2)$.

(e) La procédure d'estimation par minimum de contraste est convergente en général lorsque l'ordre n'est pas trop grand ou trop éloigné du vrai ordre. Plus on augmente la taille, plus cela se vérifie et à la limite il n'y aura plus convergence que lorsque le bon ordre a été choisi.

(f) La matrice de covariance permet de connaître l'écart-type de l'estimation, et décroît lorsque la taille de la série augmente (on sait même que la décroissance est en $\sqrt{n}$).

TP6

1. Corrélogrammes

   (a) Le processus généré est un AR(2), d'équation $X_n = 0.9X_{n-1} - 0.2X_{n-2} + \varepsilon_n$.

   (b) Pour simuler, on peut simuler à partir d'un bruit gaussien centré réduit, et multiplier par $\sqrt{2}$ les X.

   (c) Pour le corrélogramme théorique, les équations de Yule-Walker donnent :

   $$r(n+2) = 0.9r(n+1) - 0.2r(n) \quad \text{pour } n \geq 0 \quad \Longrightarrow \quad r(n) = a(0.4)^n + b(0.5)^n,$$

   avec $(a, b) \in \mathbb{R}^2$. Ensuite, on utilise les équations :

   $$\mathbb{E}(X_n^2) = r(0) = \mathbb{E}\{(0.9X_{n-1} - 0.2X_{n-2} + \varepsilon_n)^2\} = 0.85r(0) - 0.36r(1) + 2 \qquad (1)$$
   $$\mathbb{E}(X_n X_{n-1}) = r(1) = \mathbb{E}\{X_{n-1}(0.9X_{n-1} - 0.2X_{n-2} + \varepsilon_n)\} = -0.9r(0) - 0.2r(1). \qquad (2)$$

   De ces 2 équations, on en déduit que $r(0) = 100/21$, $r(1) = 75/21$, puis

   $$r(n) = \frac{50}{21}\left(-5(0.4)^n + 7(0.5)^n\right), \quad \text{pour } n \geq 0 \quad \Longrightarrow \rho(n) = \frac{r(n)}{r(0)} = \frac{1}{2}\left(-5(0.4)^n + 7(0.5)^n\right).$$

   On trouve donc $\rho(1) = 0.9$ et $\rho(2) = 0.475$, ce qui se retrouve avec le corrélogramme empirique.

   (d) Dans le cadre d'un modèle AR(p), les coefficients d'autocorrélation partielle sont les différents coefficients du modèle. C'est le cas ici, et on trouve des coefficients d'autocorrélation partielle empirique proches de ces coefficient (0.9 et -0.2).

   (e) Il vaut mieux utiliser pour un modèle AR les coefficients d'autocorrélation partielle qui donne directement les coefficients.

   (f) En augmentant la taille, les corrélations estimées sont plus proches des valeurs théoriques. Comme d'habitude, plus on a de points, mieux c'est.

2. Estimation des paramètres et de l'ordre par minimum de contraste.

   (a) On a convergence de la procédure et une estimation des coefficients assez convenable (les écarts-types de ces coefficients sont de l'ordre de 0.1).

   (b) Pour le "bon" ordre choisi (ordre (2,0)), on a normalement un bruit reconstitué qui est blanc.

   (c) En général, en changeant l'ordre, la procédure converge, sauf si on choisit un ordre trop grand. En revanche, le bruit n'est pas blanc lorsque l'on choisit de modéliser par un MA. Les ordres qui conviennent en général en taille 100 sont $(2,0)$, $(2,1)$ et $(1,0)$ (parfois $(0,2)$). En taille 1000, il ne reste souvent plus que $(2,0)$ et $(2,1)$. Par principe de parcimonie, on choisira toujours l'ordre validé le plus petit.

3. Estimation des paramètres et de l'ordre à partir des corrélations

   (a) On a convergence de la procédure et une estimation des coefficients assez convenable (les écarts-types de ces coefficients sont de l'ordre de 0.1).

   (b) Pour le "bon" ordre choisi (ordre (2,0)), on a normalement un bruit reconstitué qui est blanc.

   (c) En général, en changeant l'ordre, la procédure converge, sauf si on choisit un ordre trop grand. En revanche, le bruit n'est pas blanc lorsque l'on choisit de modéliser par un MA. Les ordres qui conviennent en général en taille 100 sont $(2,0)$, $(2,1)$ et $(1,0)$ (parfois $(0,2)$). En taille 1000, il ne reste souvent plus que $(2,0)$ et $(2,1)$. Par principe de parcimonie, on choisira toujours l'ordre validé le plus petit.

4. Exercice

   (a) Après simulation du processus qui est un ARMA(1,1), on peut estimer l'ordre. La méthode à partir des corrélogrammes donne de très mauvais résultats par rapport à l'ordre théorique (1,1), ce qui s'explique par le fait que les coefficients de la partie MA et de la partie AR de Y sont très proche de 1, donc aucune de ces 2 parties n'est négligeable.

   (b) En revanche, la méthode par minimum de contraste fournit encore des résultats attendus. Elle est plus robuste que la précédente.

5. Prédictions pour un processus ARMA

   (a) La variable $ParaZ$ est composé de l'estimation des coefficients de la partie MA et de la partie AR pour l'ordre $(1,1)$, par la méthode de minimum de contraste.

   (b) Les valeurs prévues sont à peu près correctes dans le très court terme (Y_{491} et parfois Y_{492}) mais tendent très vite vers O (qui est la "meilleure" prédiction possible pour le long-terme). La loi de l'estimateur est donnée par une loi gaussienne (le modèle ARMA est gaussien) de moyenne $pZ\$mean$ et de variance $pZ\$std$.

   (c) Enfin, on modéliser les coefficients à partir de leurs estimations par Yule-Walker. On obtient alors des prédictions assez proches des précédentes.