Réponses et commentaires aux TP 3 et 4

TP3

1. Prédiction par moindres carrés

   (a) La série y admet pour tendance : $a(x) = x - 0.1x^2$, pour saisonnalité : $s(x) = (-1)^x$ (de période 2) et pour bruit un bruit blanc ε_x qui suit une loi normale centrée de variance 25.

   (b) On compare les prédictions pour les temps $x = 25,\ 30$ et 50 avec les valeurs de $a(x)+s(x)$ car $\mathbb{E}(y_x|(y_1,..,y_{20}) = \mathbb{E}(a(x)+s(x)+\varepsilon_x|(\varepsilon_1,..,\varepsilon_{20})) = a(x)+s(x)$ du fait que le bruit est un bruit blanc. Donc $a(x)+s(x)$ est la “meilleure” prédiction possible au sens des moindres carrés. On trouve ainsi pour $x = 25,\ 30$ et 50 : -38.5, -59 et -199.

   (c) Théoriquement, comme la tendance est un polynôme de degré 2, le meilleur choix sur le degré du polynôme de régression est 2. Empiriquement, en général, on constate la même chose surtout dès que l'on fait des prédictions à long-terme ($x = 50$). Pour les prédictions, il vaut mieux choisir le degré du polynôme de régression le plus faible possible pour éviter les “explosions” à long terme.

   (d) La régression pondérée lorsque tous les poids sont uniformes donne le même résultat que précédemment. Si on ne met un poids que sur la première donnée, seule cette donnée est prise en compte dans la régression (on ne peut alors faire qu'une régression de degré 0); la même chose si le poids est uniquement sur la dernière. La prédiction est alors la donnée elle-même. Si maintenant le poids est exponentiel, on prend plus ou moins en compte les premières données. En général, lorsque l'on ne sait rien sur le modèle, on choisit un poids de l'ordre de 0.7 (en tout cas, dans $[0.5, 1]$ car sinon on ne prend vraiment en compte que les toutes dernières données).

2. Premier exercice

   (a) Pour la série X_1, théoriquement, on sait que le meilleur choix sur le degré du polynôme de régression est 1. On peut alors comparer la valeur obtenue avec cette régression et la valeur que l'on peut calculer (car le modèle est déterministe) en janvier 2000.

   (b) En mettant un poids, on peut améliorer un peu la prédiction car cela permet d'oublier le lointain passé où la partie fraction de X_1 intervenait et d'être ainsi dans la zone où c'est essentiellement la partie polynômiale qui domine dans la tendance.

3. Utilisation de filtres

   (a) X est une série dont la tendance est $1 + 0.01k$, la saisonnalité $2\sin(2\pi k/5)$ de période 5, et de bruit blanc gaussien centré de variance 4. Empiriquement, la période (et la saisonnalité) est difficilement identifiable du fait que le bruit est de même écart-type que l'amplitude de la saisonnalité.

   (b) La série filtrée se définit par $Y_n = X_{n+3} - X_{n-2}$ pour $n \in \{3, \cdots, 47\}$. En traçant la série filtrée on a réduit un peu l'irrégularité de X. Théoriquement, $\mathbb{E}Y_n = (1+0.01(n+3)) - (1+0.01(n-2)) + 0 = 0.05$, dont on a supprimé la saisonnalité et baissé d'un degré le polynôme de la tendance. Le bruit est donc $E'_n = E_{n-2} - E_{n+3}$, qui est gaussien centré de variance 8, mais qui n'est plus un bruit blanc, puisque $\mathbb{E}(E'_n E'_{n+5}) = -E(E^2_{n+3}) \neq 0$.

   (c) Pour reconstruire X à partir de (Y_n) et de $(X_{1,c}\,dots, X_5)$, on utilise $X_6 = Y_3 + X_1$, $X_7 = Y_4 + X_2$, puis le reste par itération.

(d) En utilisant l'autre filtre, on a la série $Z_n = (X_{n+2} + .. + X_{n-2})/5$. Théoriquement, Z_n a pour tendance $\mathbb{E}Z_n = 1 + 0.01n + 0$, pour saisonnier 0 et pour bruit $E''_n = -(E_{n+2} + .. + E_{n-2)/5}$, qui n'est pas un bruit blanc car $\mathbb{E}(E'_n E'_{n+1}) \neq 0$.

TP4

1. Normalement, on doit avoir les "bâtons" du corrélogrammes à l'intérieur des 2 bandes en dehors du premier coefficient (théoriquement le bruit x est un bruit blanc gaussien). Cependant, on peut avoir quelques "bâtons" très proches, voire légèrement hors ces bandes. On pourra alors encore conclure que le bruit est blanc. Il faut être manifestement hors ces bandes pour conclure à la non-nullité du coefficient, et donc au fait que le bruit n'est pas blanc.

2. On a $y_n = x_n + x_{n+1}$, pour $n = 1, .., 49$. Ainsi, y_n suit une loi normale centrée de variance 2. Quand on observe le corrélogramme, on a le second "bâton" qui est clairement hors des 2 bandes : le bruit ne semble pas blanc. Cela se confirme lorsque l'on calcule théoriquement la valeur du coefficient de corrélation linéaire pour l'écart 1 : $\rho(1) = \mathbb{E}(Y_0 Y_1)/2 = 1/2$. En revanche, les autres coefficients sont théoriquement nuls.

3. Le périodogramme est un estimateur non convergent mais sans biais de la densité spectrale. Cela se remarque normalement par le fait que le graphe présente de fortes oscillations autour de 0 (qui est la valeur du logarithme de la densité spectrale $*2\pi$ pour le bruit x considéré) pour les différentes fréquences calculées. Si on multiplie x par une constante, alors on décale la valeur moyenne de ces oscillations. Lorsqu'on lisse le périodogramme (en le convoluant par une fenêtre glissante par exemple) on améliore la régularité et on ainsi une meilleure estimation de la densité spectrale. Cependant, le résultat n'est pas très probant pour montrer que que le bruit est blanc.

4. En revanche, on observe clairement une différence en traçant le périodogramme de y. On observe un signal plus régulier, qui décroît régulièrement. Théoriquement, on obtient pour densité spectrale
$$f_y(\lambda) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} r_y(k) e^{-ik\lambda} = \frac{1}{\pi}(1 + \cos\lambda).$$
On peut ainsi tracer cette fonction (ou plutôt le logarithme...) et on observe le même type de comportement avec un maximum en 0 et une décroissance par la suite. Le signal est ainsi basse-fréquence, car l'énergie (obtenue par le module au carré de la densité spectrale) est surtout concentrée pour les basses fréquences, c'est-à-dire les fréquences proches de 0.

5. x est un bruit blanc gaussien de moyenne 2 et de variance 25. La droite de Henry de y est à peu près une droite. y est un bruit blanc uniforme sur $[1, 4]$. Cette fois-ci la droite de Henry exhibe un peu de non-linéarité aux bords. Cependant, la méthode est une méthode d'adéquation à une loi normale uniquement qualitative et donc un peu limitée.

6. Le test de Kolmogorov-Smirnov de S-plus marche "moyennement". Regarder en priorité la valeur de la p-value qui doit être supérieure à 0.05 pour que le test soit accepté. Attention au paramétrage automatique dans le cas gaussien : il est plus fiable de rentrer soi-même les paramètres de la gaussienne qui sont ses moyenne et variance empiriques. Il est à noter aussi que le test a une puissance qui augmente avec le nombre de données. Donc, il n'est pas surprenant d'accepter que y suive une loi gaussienne, car y n'est constitué que de 50 données. Si on prend maintenant, 1000 données, on a une p-value très très proche de 0.

7. Pour tester que y suit bien une loi uniforme, il ne faut pas publier de rentrer les paramètres de la loi uniforme qui sont min et max. Le teste d'adéquation des distributions de x et de y doit normalement être rejeté.