Annexes Chapitre 1

La Désaisonnalisation des Séries Temporelles

Olivier DARNÉ

Olivier DARNÉ

La moyenne mobile centrée à 12 termes $M_{2\times 12}$ est définie avec $p = 6$

$$M_{2\times 12}(X_t) = \frac{1}{2p}\left(\frac{X_{t-p}}{2} + \sum_{i=-p+1}^{p-1}(X_{t+i}) + \frac{X_{t+p}}{2}\right)$$

$$\begin{aligned}
M_{2\times 12} &= \frac{1}{2p}\left(\frac{L^{-p}}{2} + \sum_{i=-p+1}^{p-1}(L^i) + \frac{L^p}{2}\right) = \frac{1}{12}\left(\frac{L^{-6}}{2} + \sum_{i=-6+1}^{6-1}(L^i) + \frac{L^6}{2}\right) \\
&= \frac{1}{12}\left(\frac{L^{-6}}{2} + L^{-5} + L^{-4} + \cdots + L^5 + \frac{L^6}{2}\right) \\
&= \frac{1}{24}\left(L^{-6} + 2L^{-5} + 2L^{-4} + \cdots + 2L^5 + L^6\right) \\
&= \frac{1}{24}L^{-6}\left(1 + 2L + 2L^2 + \cdots + 2L^{11} + L^{12}\right) \\
&= \frac{1}{24}L^{-6}(1+L)\left(1 + L + L^2 + \cdots + L^{11}\right)
\end{aligned}$$

Les moyennes mobiles de Henderson

Dans X-11 l'estimation de la tendance-cycle TC_t s'obtient par l'extraction de TC_t de la série désaisonnalisée $SA_t = TC_t + I_t$

- filtre conservant la TC_t
- filtre réduisant la composate irrégulière (bruit) I_t

Afin d'assurer d'obtenir une estimation lisse (*smooth*) de la tendance-cycle Henderson (1916) propose un programme de minimisation sur les coefficients θ_i des moyennes mobiles :

Minimise :

$$H_k = \sum_{i=-p}^{p} (\nabla^3 \theta_i)^2$$

sous les contraintes

$$\sum_{i=-p}^{p} \theta_i = 1 \qquad \text{et} \qquad \sum_{i=-p}^{p} i\theta_i = 0 \qquad \text{et} \qquad \sum_{i=-p}^{p} i^2\theta_i = 0$$

Les coefficients de ces moyennes mobiles peuvent aussi être calculés de manière explicite pour des moyennes mobiles d'ordre $k = 2p+1$, avec

$$\theta_i = \frac{315\left[(n-1)^2 - i^2\right]\left[n^2 - i^2\right]\left[(n+1)^2 - i^2\right]\left[3n^2 - 16 - 11i^2\right]}{8n(n-1)^2(4n^2-1)(4n^2-9)(4n^2-25)}$$

Dans X-11 l'estimation de la tendance-cycle TC_t s'obtient par l'application d'un des trois filtres linéaires de Henderson disponibles dans le logiciel : les filtres à 9, 13 et 23 termes, notés respectivement H9, H13 et H23.

Le tableau donne les poids θ_i des moyennes mobiles symétriques telles que $\theta_i = \theta_{-i}$

Table 1: Weight functions a_j for some common weighted moving averages.

Name	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
3 MA	.333	.333										
5 MA	.200	.200	.200									
2×12 MA	.083	.083	.083	.083	.083	.083	.042					
3×3 MA	.333	.222	.111									
3×5 MA	.200	.200	.133	.067								
S15 MA	.231	.209	.144	.066	.009	−.016	−.019	−.009				
S21 MA	.171	.163	.134	.037	.051	.017	−.006	−.014	−.014	−.009	−.003	
H5 MA	.558	.294	−.073									
H9 MA	.330	.267	.119	−.010	−.041							
H13 MA	.240	.214	.147	.066	.000	−.028	−.019					
H23 MA	.148	.138	.122	.097	.068	.039	.013	−.005	−.015	−.016	−.011	−.004

S = Spencer's weighted moving average
H = Henderson's weighted moving average

Les moyennes mobiles asymétriques de Musgrave

Lorsque l'on applique une moyenne mobile symétrique d'ordre $2p+1$ (par ex. moyenne mobile de Henderson) il n'est pas possible d'obtenir, par construction les p premières et dernières observations de la série *TC* lissée

Pour les observations en début et en fin de série, X-11 emploie des moyennes mobiles asymétriques dérivées de la méthode de Musgrave (1964) afin de prolonger la série

Cette méthode construit des moyennes mobiles asymtriques qui minimise les révisions des estimations. Il pose les hypothèses suivantes :

- la série à lisser peut se modéliser linéairement sous la forme $X_t = a + bt + \varepsilon_t$, où a et b sont des constantes, et les ε_t sont des variables aléatoires non corrélées, de moyenne nulle et de variance σ^2
- on dispose d'une série de poids $\{\omega_1, \ldots, \omega_N\}$ (par ex. une moyenne mobile centrée de Henderson) et on cherche une série de poids $\{\upsilon_1, \ldots, \upsilon_M\}$, avec $M < N$, $\sum_{i=1}^{N} \omega_i = 1$ et $\sum_{j=1}^{N} \upsilon_j = 1$
- cette nouvelle moyenne mobile doit minimiser les révisions des estimations en minimisant le critère :

$$E\left(\sum_{i=1}^{M} \upsilon_i X_i - \sum_{i=1}^{N} \omega_i X_i\right)^2$$

Sous ces hypothèses, les poids peuvent être calculés explicitement en fonction du rapport bruit/signal (I/C) $D = b^2/\sigma^2$

$$\upsilon_j = \omega_j + \left[\frac{1}{M}\sum_{i=M+1}^{N}\omega_i\right] + \left[\frac{\left(j - \frac{M+1}{2}\right)D}{1 + \frac{M(M-1)(M+1)}{12}D}\sum_{i=M+1}^{N}\left(i - \frac{M+1}{2}\right)\omega_i\right]$$

Elles sont obtenues pour des valeurs pré-définies du rapport bruit/signal (I/C) et déterminent ainsi la longueur du filtre de tendance-cycle de Henderson à appliquer.

Algorithme de la méthode X-11

Série brute mensuelle : $X_t = C_t + S_t + I_t$

1. **Estimation de la tendance-cycle par une moyenne mobile** 2×12 **:**
   $C_t^{(1)} = M_{2\times12}(X_t)$
2. **Estimation de la composante saisonnier-irrégulier :**
   $(S_t + I_t)^{(1)} = X_t - C_t^{(1)}$
3. **Estimation de la composante saisonnière par une moyenne** 3×3 **sur chaque mois :**
   $S_t^{(1)} = M_{3\times3}\left[(S_t + I_t)^{(1)}\right]$
   et normalisation
   $\tilde{S}_t^{(1)} = S_t^{(1)} - M_{2\times12}\left(S_t^{(1)}\right)$
4. **Estimation de la série corrigée des variations saisonnières :**
   $A_t^{(1)} = (C_t + I_t)^{(1)} = X_t - \tilde{S}_t^{(1)}$

5. **Estimation de la tendance-cycle par une moyenne mobile de Henderson sur 13 termes :**
   $C_t^{(2)} = H_{13}\left(A_t^{(1)}\right)$
6. **Estimation de la composante saisonnier-irrégulier :**
   $(S_t + I_t)^{(2)} = X_t - C_t^{(2)}$
7. **Estimation de la composante saisonnière par une moyenne mobile** 3×5 **sur chaque mois :**
   $S_t^{(2)} = M_{3\times5}\left[(S_t + I_t)^{(2)}\right]$
   et normalisation
   $\tilde{S}_t^{(2)} = S_t^{(2)} - M_{2\times12}\left(S_t^{(2)}\right)$
8. **Estimation de la série corrigée des variations saisonnières :**
   $A_t^{(2)} = (C_t + I_t)^{(2)} = X_t - \tilde{S}_t^{(2)}$

TAB. 2.2 – L'algorithme de base de X-11.

Schéma de la méthode X-11

Partie A :	**Ajustements préalables**
	- pour aléas connus et importants
	- pour jours ouvrables
Partie B :	**Première correction automatique de la série**
	- Estimation de la composante irrégulière
	- Détection et correction automatique des points atypiques
	- Correction des effets de jours ouvrables
Partie C :	**Seconde correction automatique de la série**
	- Estimation de la composante irrégulière
	- Détection et correction automatique des points atypiques
	- Correction des effets de jours ouvrables
Partie D :	**Désaisonnalisation**
	1 Calcul de la série désaisonnalisée provisoire (tableaux D1 à D6)
	2 Lissage de la série désaisonnalisée par une moyenne mobile de Henderson et nouvelle estimation des coefficients saisonniers (tableaux D7 to D10)
	3 Calcul de la série désaisonnalisée définitive (tableau D11), extraction de la composante tendance-cycle (tableau D12) et de la composante irrégulière (tableau D13)
Partie E :	**Composantes corrigées des valeurs très atypiques**
Partie F :	**Mesures de qualité de la désaisonnalisation**
Partie G :	**Graphiques**

TAB. 2.3 – Schéma simplifié du fonctionnement de X-11.

Le logiciel STAMP

Koopman et alii (1995, 2010) ont développé le logiciel STAMP (*Structural Time series Analyser, Modeller and Predictor*) à la London School of Economics and Political Science.

STAMP, dans son option par défaut, suggère d'utiliser le modèle structurel fondamental (BSM, *Basic Structural Model*) proposé par Harvey (1981), cad un modèle sans élément exogène ni cycle, pour modéliser les composantes inobservables.

La composante irrégulière est généralement supposée être un bruit blanc Gaussien de moyenne zéro et de variance σ_ε^2 :

$$I_t = \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

La tendance est modélisée comme une marche aléatoire avec dérive, où la dérive (la pente de la tendance) suit également une marche aléatoire.

$$\begin{aligned} T_t &= T_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \end{aligned}$$

- $\eta_t \sim N(0, \sigma_\eta^2)$ et $\zeta_t \sim N(0, \sigma_\zeta^2)$ sont deux processus de bruit blanc mutuellement non corrélés
- si $\sigma_\zeta^2 = 0 \Rightarrow$ processus $I(1)$: $T_t = T_{t-1} + \beta_1 + \eta_t$
- si $\sigma_\zeta^2 = 0$ et $\sigma_\eta^2 = 0 \Rightarrow$ tendance linéaire : $T_t = T_1 + \beta_1 t$

Dans le BSM, la composante saisonnière est un modèle trigonométrique stochastique (Harvey, 1989) :

$$S_t = \sum_{j=1}^{[s/2]} \gamma_{j,t}$$

où chaque $\gamma_{j,t}$ est générée par :

$$\begin{bmatrix} \gamma_{j,t} \\ \gamma^*_{j,t} \end{bmatrix} = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix} \begin{bmatrix} \gamma_{j,t-1} \\ \gamma^*_{j,t-1} \end{bmatrix} + \begin{bmatrix} \omega_{j,t} \\ \omega^*_{j,t} \end{bmatrix}$$

- $\lambda_j = 2\pi j/s$, $j = 1, \ldots, [s/2]$, s : le nombre d'observations par année, et $t = 1, \ldots, T$.
- les innovations saisonnières $\omega_{j,t}$ et $\omega^*_{j,t}$ sont mutuellement non corrélées, de moyenne zéro et de variance commune σ^2_ω

Le modèle d'espace d'état

Pour estimer les paramètres (σ^2_η, σ^2_ζ, σ^2_ω, σ^2_ε), la tendance et la saisonnalité, le modèle structurel est mis sous forme d'espace d'état

Les modèles espace-état intègrent la distinction entre les variables observées (le signal) et les variables inobservables. Ils sont constitués :

- une équation de mesure ou équation d'observation (*measurement equation*), décrivant la manière dont les variables observées sont générées par les variables inobservables et les résidus

$$Y_t = C_t Z_t + \eta_t$$

- une équation d'état (*transition equation*), décrivant la manière dont les variables inobservables sont générées à partir de leur retard et d'innovations

$$Z_t = A_{t-1} Z_{t-1} + \varepsilon_t$$

Le système d'espace d'état est définit par ces deux équations :

$$Y_t = C_t Z_t + \eta_t \quad (1)$$
$$Z_t = A_{t-1} Z_{t-1} + \varepsilon_t \quad (2)$$

- Y_t : vecteur de mesure
- Z_t : vecteur d'état
- C_t : matrice de mesure
- A_t : matrice de transition
- η_t : vecteur des innovations
- ε_t : vecteur des erreurs de mesures
- $C_t Z_t$: signal

Le système d'espace d'état est dit sous forme canonique ssi

$$E(\varepsilon_t \eta_s) = E(\varepsilon_t Z_0) = E(\eta_t Z_0) = 0 \qquad \forall t, s = 1, \ldots, T$$

Estimation des variables d'état par le filtre de Kalman

Il s'agit d'estimer à chaque instant t les variables inobservables (vecteur d'état) conditionnellement aux variables observées jusqu'à la date t (vecteur de mesure)

Pour calculer des estimations filtrées du vecteur d'état, l'algorithme optimal, appelé filtre de Kalman, est utilisé.

L'algorithme est structuré en deux étapes reprises d'itération en itération.

- les 2 premières équations (1 et 2) sont des équations de « mises à jour des mesures » (actualisation)
- les 2 équations suivantes (3 et 4) sont des équations de « mise à jour du temps » (prévision)

- la dernière équation (5) actualise la matrice gain K_t (gain de précision) qui intervient dans les équations précédentes
- les équations (2) et (4) sur les matrices de covariance sont appelées « équations de Riccati »

Chaque itération se résume par les 5 équations suivantes :

$$
\begin{aligned}
(1) \quad &: \quad Z^*_{t,t} = Z^*_{t-1,t} + K_t(Y_t - C_t Z^*_{t-1,t}) \\
(2) \quad &: \quad \Sigma_{t,t} = (I - K_t C_t)\Sigma_{t-1,t} \\
(3) \quad &: \quad Z^*_{t,t+1} = A_t Z^*_{t,t} \\
(4) \quad &: \quad \Sigma_{t,t+1} = A_t \Sigma_{t,t} A'_t + Q_t \\
(5) \quad &: \quad K_t = \Sigma_{t-1,t} C_t (C_t \Sigma_{t-1,t} C'_t + R_t)^{-1}
\end{aligned}
$$

- $Z^*_{t,t}$: estimation courante du vecteur d'état
- $\Sigma_{t,t} = V(Z_{t,t} - Z^*_{t,t})$: erreur quadratique moyenne sur Z_t
- $Z^*_{t-1,t}$: prévision du vecteur d'état faite à la date $t-1$
- $\Sigma_{t-1,t} = V(Z_{t-1,t} - Z^*_{t-1,t})$: erreur quadratique moyenne de prévision correspondante
- K_t : matrice de gain de Kalman

Pour estimer les paramètres (σ^2_η, σ^2_ζ, σ^2_ω, σ^2_ε), la tendance et la saisonnalité, le modèle structurel est mis sous forme d'espace d'état, avec l'équation d'observation :

$$y_t = (1\ 0\ 1\ 0)\,\alpha_t + (\sigma_\varepsilon\ 0\ 0\ 0)\,u_t$$

où α_t, le vecteur d'état, dont les éléments sont inobservables, suit l'équation d'état suivante :

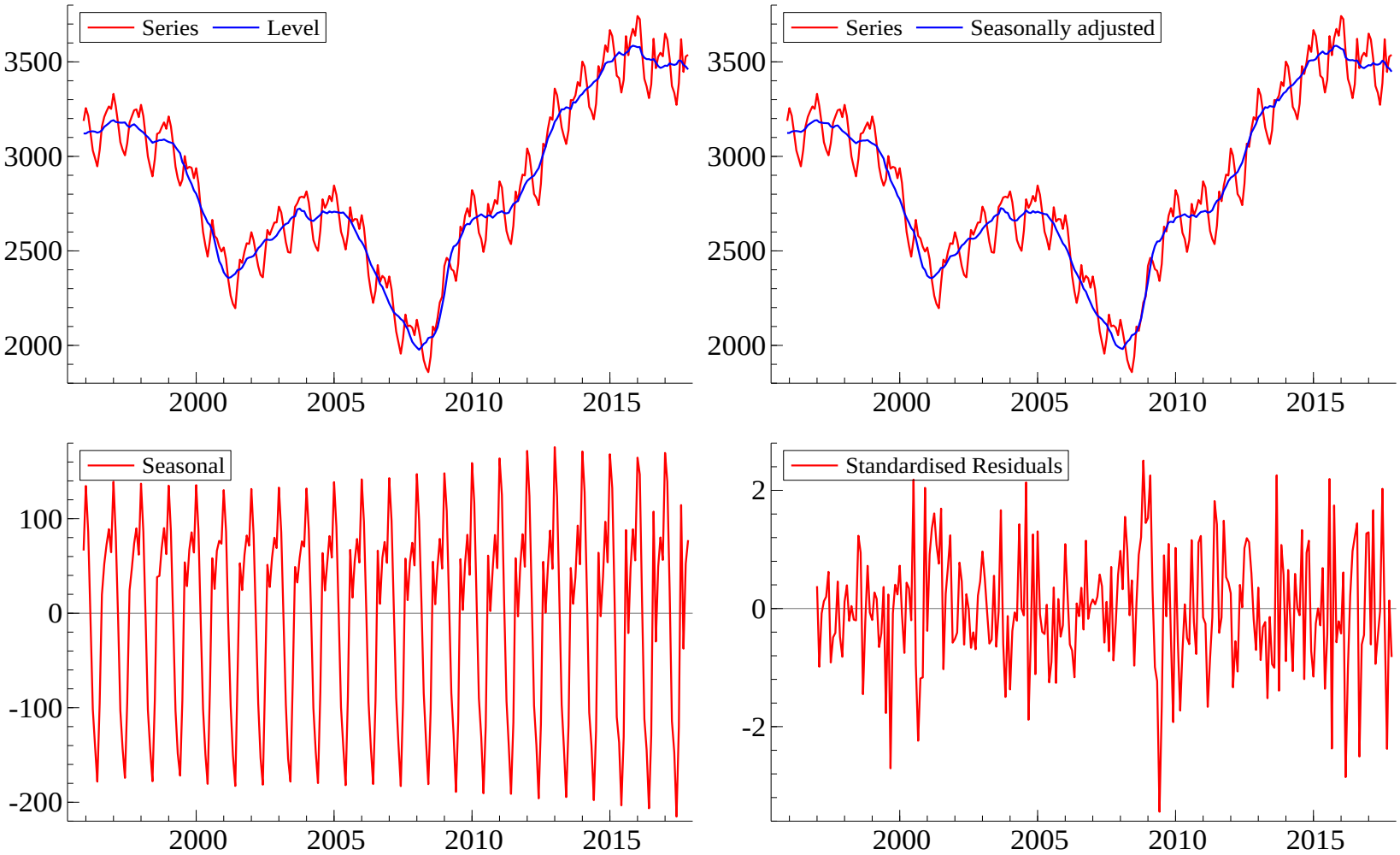
$$\alpha_t = \begin{pmatrix} T_t \\ \beta_t \\ \gamma_{j,t} \\ \gamma^*_{j,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\lambda_j & \sin\lambda_j \\ 0 & 0 & -\sin\lambda_j & \cos\lambda_j \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 0 & \sigma_\eta & 0 & 0 \\ 0 & 0 & \sigma_\zeta & 0 \\ 0 & 0 & 0 & \sigma_\omega \\ 0 & 0 & 0 & 0 \end{pmatrix} u_{t-1}$$

avec $u_t = (\varepsilon_t\ \eta_t\ \zeta_t\ \omega_t)'$.

Les paramè tres sont estimés par la méthode du maximum de vraisemblance.

L'estimation des composantes (optimales car ce sont des estimateurs MMSE, *Minimum Mean Squared Error*) est alors obtenue en utilisant le filtre et le lisseur de Kalman

Figure: Décomposition de la série par STAMP



Il existe principalement 4 types de points atypiques, définis de la manière suivante :

- *Additive Outliers* (AO) : affectent une seule observation à un moment du temps dans la série temporelle

$$f(t)_{AO} = \omega_{AO} I_t(\tau)$$

- *Level Shifts* (LS) : effet permanent sur le niveau de la série

$$f(t)_{LS} = [1/(1-L)]\,\omega_{LS} I_t(\tau)$$

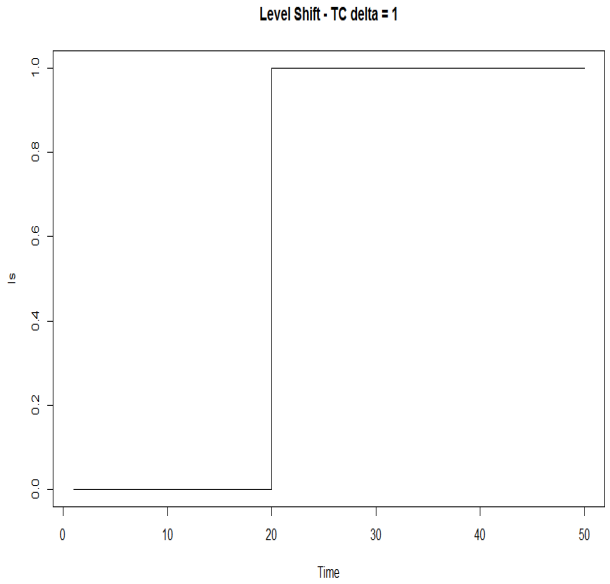
- *Temporary Changes* (TC) : effet temporaire la série qui retourne à son niveau précédent de manière exponentielle. Leur vitesse de retour dépend du paramètre δ dans le polynÃ´me

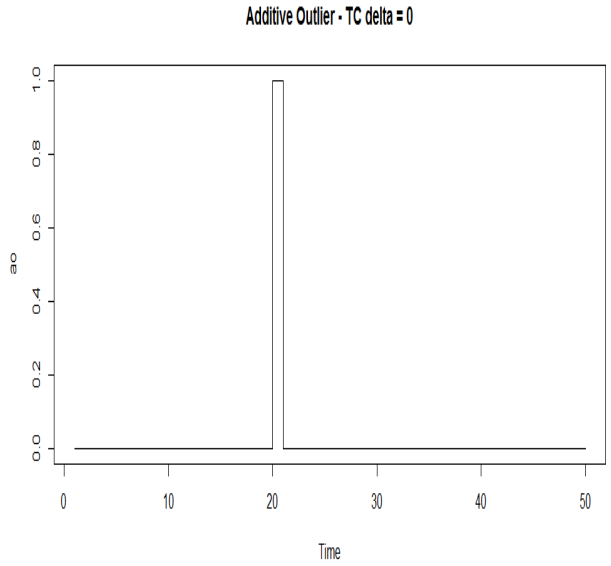
$$f(t)_{TC} = [1/(1-\delta L)]\,\omega_{TC} I_t(\tau) \qquad 0 < \delta < 1$$

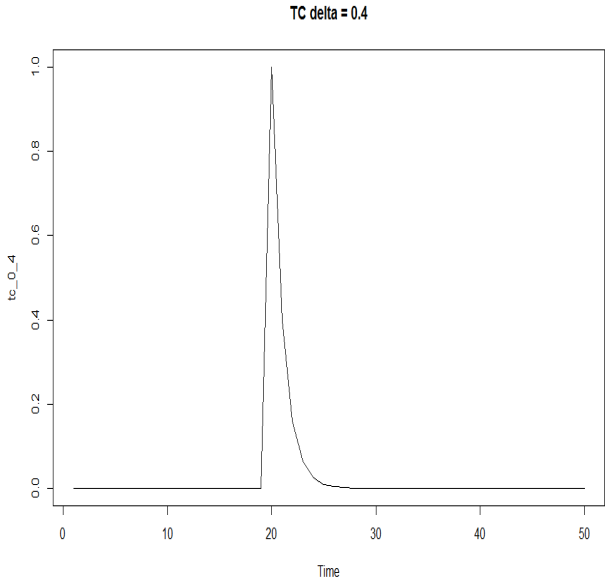
- *Innovative Outliers* (IO) : impact sur les innovations du modèle

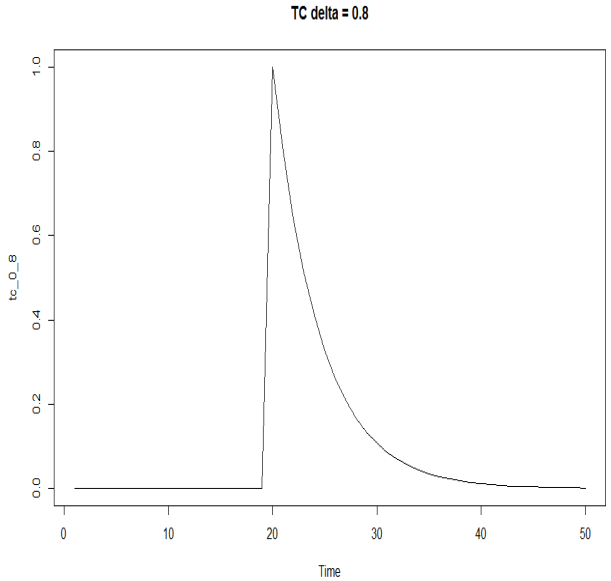
$$f(t)_{IO} = [\theta(L)/\alpha(L)\phi(L)]\,\omega_{IO} I_t(\tau)$$

Les points atypiques









Un modèle ARIMA est ajusté à X_t^* dans l'équation (**??**), et les résidus obtenus sont définis par :

$$\hat{a}_t = \pi(L)X_t^* \tag{3}$$

où $\pi(L) = \alpha(L)\phi(L)/\theta(L) = 1 - \pi_1 L - \pi_2 L^2 - \ldots.$

Pour les quatre types de points atypiques, l'équation (3) devient

$$\begin{aligned}
\text{AO:} \quad & \hat{a}_t = a_t + \omega_{AO} I_t(\tau) \\
\text{IO:} \quad & \hat{a}_t = a_t + \omega_{IO}\pi(L) I_t(\tau) \\
\text{LS:} \quad & \hat{a}_t = a_t + \omega_{LS}[\pi(L)/(1-L)] I_t(\tau) \\
\text{TC:} \quad & \hat{a}_t = a_t + \omega_{TC}[\pi(L)/(1-\delta L)] I_t(\tau)
\end{aligned}$$

Ces expressions peuvent Ã[a]tre vues comme un modèle de régression pour les résidus $\hat{a}_t$, cad

$$\hat{a}_t = \omega_i x_{i,t} + a_t \qquad i = \text{AO,IO, LS, TC,}$$

- $x_{i,t} = 0 \ \forall i$ et $t < \tau$
- $x_{i,t} = 1 \ \forall i$ et $t = \tau$,
- pour $t > \tau$ et $k \geq 1$,

$$\begin{aligned}
\text{AO:} \qquad & x_{AO,t+k} = 0 \\
\text{IO:} \qquad & x_{IO,t+k} = -\pi_k \\
\text{LS:} \qquad & x_{LS,t+k} = 1 - \sum_{j=1}^{k} \pi_j \\
\text{TC:} \qquad & x_{TC,t+k} = \delta^k - \sum_{j=1}^{k-1} \delta^{k-j} \pi_j - \pi_k
\end{aligned}$$

Les statistiques de test pour les 4 types de points atypiques sont basées sur des statistiques du ratio de vraisemblance (LR, *likelihood ratio*)

$$\text{AO:} \qquad \hat{\tau}_{AO}(\tau) = [\hat{\omega}_{AO}(\tau)/\hat{\sigma}_a]/\Big(\sum_{t=\tau}^{n} x_{AO,t}^2\Big)^{1/2}$$

$$\text{IO:} \qquad \hat{\tau}_{IO}(\tau) = \hat{\omega}_{IO}(\tau)/\hat{\sigma}_a$$

$$\text{LS:} \qquad \hat{\tau}_{LS}(\tau) = [\hat{\omega}_{LS}(\tau)/\hat{\sigma}_a]/\Big(\sum_{t=\tau}^{n} x_{LS,t}^2\Big)^{1/2}$$

$$\text{TC:} \qquad \hat{\tau}_{TC}(\tau) = [\hat{\omega}_{TC}(\tau)/\hat{\sigma}_a]/\Big(\sum_{t=\tau}^{n} x_{TC,t}^2\Big)^{1/2}$$

$$\text{avec} \qquad \hat{\omega}_i(\tau) = \sum_{t=\tau}^{n} \hat{a}_t x_{i,t} / \sum_{t=\tau}^{n} x_{i,t}^2 \qquad \text{for } i = \text{AO, LS, TC,}$$

$$\text{et} \qquad \hat{\omega}_{IO}(\tau) = \hat{a}_\tau$$

- $\hat{\omega}_i(\tau)$ avec $i =$ AO, IO, LS, TC : estimation de l'impact du point atypique au temps $t = \tau$
- $\hat{\sigma}_a$: une estimation de la variance des résidus

Les points atypiques

Les points atypiques sont identifiés lors d'une procédure de détection séquentielle, comprenant une itération interne et une autre externe

- dans l'itération externe, en supposant qu'il n'y a pas des points atypiques, un modèle ARIMA (p, d, q) est estimé, donnant ainsi les résidus
- les résultats de l'itération externe sont alors utilisés dans l'itération interne pour identifier les points atypiques
- les statistiques de test pour les 4 types de points atypiques sont calculées pour chaque observation
- si $\hat{\tau}_{max} = max|\hat{\tau}_i(\tau)| > VC \Rightarrow$ un point atypique est identifié au temps $t = \tau$

Quand un outlier est détecté au temps $t = \tau_1$ alors la série est corrigée de la manière suivante :

$$X_t^* = X_t - f(t)_{i^*}$$

La procédure est répétée jusqu'à plus aucun point atypique ne soit détecté

Ces détections et corrections de points atypiques sont mises en œuvre dans les logiciels TRAMO et RegARIMA (pré-ajustement de la série à modéliser)