L'estimation des processus ARMA & les tests de validation

Gilbert Colletaz

19 octobre 2020

Résumé

On présente ici les trois méthodes les plus couramment utilisées pour estimer les paramètres d'un ARMA : l'estimation des moindres carrés conditionnels, des moindres carrés non conditionnels et du maximum de vraisemblance. Une fois l'estimation effectuée on s'intéresse naturellement à la qualité de l'ajustement et donc à la mise en oeuvre de tests validant ou non le modèle que l'on vient d'estimer.

Table des matières

1 L'estimation : Moindres carrés conditionnels, non conditionnels et maximum de vraisemblance

On illustrera ces méthodes en considérant en détail le cas d'un processus AR(1), ce qui devrait permettre de comprendre chacune des trois méthodes sans se perdre dans des complexités inutiles. Seuls quelques exemples illustreront les cas des autres processus.

On supposera donc, sauf dans les exemples particuliers, avoir à estimer le vecteur $(\theta, \sigma_u^2) = (\phi, \mu_x, \sigma_u^2)$ dans un modèle d'écriture :

$$(x_t - \mu_x) = \phi(x_{t-1} - \mu_x) + u_t$$

(notez que le vecteur des paramètres à estimer est scindé en deux : d'un côté θ qui contient les coefficients de l'équation du modèle, *i.e.* ce que l'on appelle souvent constante et pentes, et de l'autre la variance résiduelle, σ_u^2).

1.1 moindres carrés conditionnels

Ici, les valeurs des paramètres estimés sont obtenues en satisfaisant le critère usuel des moindres carrés ordinaires à savoir la minimisation d'une somme des carrés des résidus. On les qualifie de moindres carrés conditionnels car, d'une part la somme des carrés considérés porte sur $T - p$ termes en présence d'une composante autorégressive d'ordre p, *i.e.* on n'intègre pas dans la somme les carrés des p premiers résidus et, d'autre part, les valeurs de la variable $x - \mu_x$ et des innovations qui précèdent la première date d'observation de l'échantillon sont égalées à leur espérance non conditionnelle, c'est-à-dire zéro. Avec ces deux règles, les résidus intégrés dans le calcul de la somme des carrés conditionnelle, RSS_c, ne

dépendent que de valeurs de la variable x ou de l'innovation u connues. Pour illustrer ces règles, on peut considérer les exemples suivants :

— AR(1) : $RSS_c = \sum_{t=2}^{T} u_t^2 = \sum_{t=2}^{T} \left[(x_t - \mu_x) - \phi_1(x_{t-1} - \mu_x)\right]^2$
— AR(2) : $RSS_c = \sum_{t=3}^{T} u_t^2 = \sum_{t=3}^{T} \left[(x_t - \mu_x) - \phi_1(x_{t-1} - \mu_x) - \phi_2(x_{t-2} - \mu_x)\right]^2$
— MA(1) : $RSS_c = \sum_{t=1}^{T} u_t^2$, avec $x_t - \mu_x = u_t - \theta_1 u_{t-1}$ et donc

$$\begin{aligned}
u_1 &= x_1 - \mu_x + \theta_1 u_0 = x_1 - \mu_x, \text{ car } u_0 = 0 \text{ par application de la règle,} \\
u_2 &= x_2 - \mu_x + \theta_1 u_1, \\
u_3 &= x_3 - \mu_x + \theta_1 u_2, \\
&\vdots \\
u_T &= x_T - \mu_x + \theta_1 u_{t-1},
\end{aligned}$$

— ARMA(1,1) : $RSS_c = \sum_{t=2}^{T} u_t^2$, avec $(x_t - \mu_x) = \phi_1(x_{t-1} - \mu_x) + u_t - \theta_1 u_{t-1}$, et :

$$\begin{aligned}
u_1 &= x_1 - \mu_x \text{ par application de la règle,} \\
u_2 &= x_2 - \mu_x - \phi_1(x_1 - \mu_x) + \theta_1 u_1 \\
u_3 &= x_3 - \mu_x - \phi_1(x_2 - \mu_x) + \theta_1 u_2 \\
&\vdots \\
u_T &= x_T - \mu_x - \phi_1(x_{T-1} - \mu_x) + \theta_1 u_{t-1}
\end{aligned}$$

Dans tous ces exemples, vous devez remarquer qu'étant donné un ensemble d'observations, $\{x_1, x_2, \ldots, x_T\}$, si on attribue des valeurs aux paramètres inconnus du modèle $\hat{\theta}$ alors chacune des RSS_c est calculable. En conséquence cela a un sens de rechercher quelles valeurs de ces paramètres minimisent la somme des carrés des résidus conditionnelle. En cas de présence de termes MA dans le modèle à estimer, la difficulté est qu'il faut mettre en oeuvre des algorithmes de moindres carrés non linéaires. Pour vous en convaincre, il suffit par exemple dans le cas de l'ARMA(1,1) ci-dessus de remplacer u_2 par son équation de définition dans l'expression de u_3. On voit alors apparaître des termes de la forme θ_1^2 et $\theta_1\phi_1$: le modèle est donc évidemment non linéaire en ϕ_1, θ_1. En conséquence, l'estimation des ARMA(p,q) et MA(q) avec $q > 0$ peut rapidement poser des problèmes de convergence même pour des valeurs faibles de q et de fait, des processus pour lesquels $q > 3$ sont rarement rencontrés. En revanche, vous devez pouvoir vérifier en considérant les cas des processus AR(1) et AR(2), qui se caractérisent par l'absence de termes MA, que le problème ne se pose pas : il suffit d'appliquer un algorithme de moindres carrés linéaires pour minimiser la somme des carrés des résidus de ces deux modèles.

1.2 maximum de vraisemblance et moindres carrés non conditionnels

On considère encore ici notre modèle type : $X_t \sim$ AR(1) stationnaire. La fonction de vraisemblance qui est associée à une trajectoire observée de ce processus $X_1, X_2, \ldots, X_T$ a pour écriture : $L(X_T, X_{T-1}, \ldots, X_3, X_2, X_1; \beta, \sigma_u^2)$ où $\beta = (\mu_X, \phi_1)$. On peut simplifier l'écriture de cette vraisemblance en faisant notamment apparaître les vraisemblances conditionnelles des observations. En effet, avec Bayes, nous avons :

$$\begin{aligned}
f(X_T, X_{T-1}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) &= f(X_T | X_{T-1}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) f(X_{T-1}, X_{T-2}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2), \text{ et pour un AR(1),} \\
&= f(X_T | X_{T-1}; \theta, \sigma_u^2) f(X_{T-1}, X_{T-2}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) \\
&= f(X_T | X_{T-1}; \theta, \sigma_u^2) f(X_{T-1} | X_{T-2}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) f(X_{T-2}, X_{T-3}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) \\
&= f(X_T | X_{T-1}; \theta, \sigma_u^2) f(X_{T-1} | X_{T-2}; \theta, \sigma_u^2) f(X_{T-2}, X_{T-3}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) \\
&= f(X_T | X_{T-1}; \theta, \sigma_u^2) f(X_{T-1} | X_{T-2}; \theta, \sigma_u^2) \ldots f(X_3 | X_2; \theta, \sigma_u^2) f(X_2 | X_1; \theta, \sigma_u^2) f(X_1)
\end{aligned}$$

Pour aller plus loin et connaître les expressions des densités du membre de droite, il faut connaître la distribution de chacune des aléatoires. Comme souvent, nous ferons une hypothèse de normalité sur l'innovation : $u_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2)$. Nous sommes donc maintenant en présence d'un bruit blanc gaussien. Comme X possède une écriture de Wold sur u, cela

implique que X est lui-même gaussien. Avec $(X_t - \mu_x) = \phi_1(X_{t-1} - \mu_x) + u_t$, vous devez pouvoir montrer que :

$$
\begin{aligned}
X_1 \sim \mathcal{N}(\mu_x, \frac{\sigma_u^2}{1-\phi_1^2}) \Rightarrow f(X_1) &= \frac{1}{\frac{\sigma_u\sqrt{2\pi}}{\left(1-\phi_1^2\right)^{1/2}}} exp\left(-\frac{1}{2}\frac{(X_1-\mu_x)^2}{\frac{\sigma_u^2}{1-\phi_1^2}}\right) \\
X_{2|X_1} \sim \mathcal{N}(\mu_x + \phi_1(X_1-\mu_x), \sigma_u^2) \Rightarrow f(X_{2|X_1}) &= \frac{1}{\sigma_u\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{(X_2-\mu_x-\phi_1(X_1-\mu_x))^2}{\sigma_u^2}\right) \\
X_{3|X_2} \sim \mathcal{N}(\mu_x + \phi_1(X_2-\mu_x), \sigma_u^2) \Rightarrow f(X_{3|X_2}) &= \frac{1}{\sigma_u\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{(X_3-\mu_x-\phi_1(X_2-\mu_x))^2}{\sigma_u^2}\right) \\
&\vdots \\
X_{T|X_{T-1}} \sim \mathcal{N}(\mu_x + \phi_1(X_{T-1}-\mu_x), \sigma_u^2) \Rightarrow f(X_{T|X_{T-1}}) &= \frac{1}{\sigma_u\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{(X_T-\mu_x-\phi_1(X_{T-1}-\mu_x))^2}{\sigma_u^2}\right)
\end{aligned}
$$

Soit encore $f(X_{t|X_{t-1}}) = \frac{1}{\sigma_u\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{u_T^2}{\sigma_u^2}\right)$ pour $t = 2, 3, \ldots, T$. En conséquence, la vraisemblance s'écrit encore :

$$
L(X_T, X_{T-1}, \ldots, X_3, X_2, X_1; \theta, \sigma_u^2) = \frac{(1-\phi_1^2)^{1/2}}{(\sigma_u\sqrt{2\pi})^T} \prod_{t=2}^{T} exp\left(-\frac{u_t^2}{2\sigma_u^2}\right) exp\left(-\frac{(1-\phi_1^2)(X_1-\mu_x)^2}{2\sigma_u^2}\right)
$$

et donc une log-vraisemblance pour un AR(1) :

$$
\begin{aligned}
l(\theta, \sigma_u^2; X1, X_2, \ldots, X_{t-1}, X_T) &= \frac{1}{2} log\left(\frac{1-\phi_1^2}{(\sigma_u\sqrt{2\pi})^T}\right) - \frac{1}{2\sigma_u^2}\left((1-\phi_1^2)(X_1-\mu_x)^2 + \sum_{t=2}^{T} u_t^2\right) \\
&= \frac{1}{2} log\left(\frac{(1-\phi_1^2)}{(\sigma_u\sqrt{2\pi})^T}\right) - \frac{1}{2\sigma_u^2}\left((1-\phi_1^2)(X_1-\mu_x)^2 + RSS_c\right)
\end{aligned}
$$

où RSS_c est, ainsi que nous l'avons vu dans la section précédente, la somme des carrés des résidus que minimise les estimateurs des moindres carrés conditionnels lorsqu'on ajuste un AR(1), somme qui ne considère pas la première observation mais seulement celles de rang 2 à T. Pour cette raison, le terme $(1-\phi_1^2)(X_1-\mu_x)^2 + RSS_c$ qui lui ajoute une quantité fondée sur cette première observation est qualifiée de somme des carrés non conditionnelle, et les estimateurs qui minimisent cette somme $RSS_u = (1-\phi_1^2)(X_1-\mu_x)^2 + RSS_c$ sont naturellement nommés estimateurs des moindres carrés non conditionnels. Pour résumer, l'estimation d'un AR(1) peut se faire au moyen :

- des estimateurs des moindres carrés conditionnels qui minimisent $RSS_c = \sum_{t=2}^{T}(X_t - \mu_x - \phi_1(X_{t-1}-\mu_x))^2$,
- des estimateurs des moindres carrés non conditionnels qui minimisent $RSS_u = (1-\phi_1^2)(X_1-\mu_x)^2 + RSS_c$,
- des estimateurs du maximum de vraisemblance qui maximisent $\frac{1}{2} log\left(\frac{(1-\phi_1^2)}{(\sigma_u\sqrt{2\pi})^T}\right) - \frac{1}{2\sigma_u^2}(RSS_u)$.

Les deux dernières techniques nécessitent l'emploi d'algorithmes itératifs de maximisation de fonctions non linéaires et les logiciels utilisent souvent les solutions des moindres carrés conditionnels comme valeurs initiales pour ces itérations. Les trois types d'estimateurs peuvent être obtenus avec la proc ARIMA de SAS via l'option METHOD à faire apparaître dans la commande ESTIMATE en choisissant un des trois mots clefs selon : METHOD=CLS|ULS|ML. Par défaut, METHOD=CLS.
Il est généralement admis que les estimateurs du maximum de vraisemblance devraient être préférés aux deux autres.

2 Les tests de validation

2.1 les tests d'orthogonalité des résidus

On sait que tout processus stationnaire possède une écriture de Wold unique sur un processus d'innovation qui est un bruit blanc. On sait également que si un processus MA inversible, AR stationnaire ou ARMA stationnaire et inversible gouverne une variable, alors cette modélisation est unique et est équivalente à l'écriture de Wold de cette variable. Dans ces conditions, si on a sélectionné le bon processus, son résidu doit se confondre avec le bruit blanc de cette écriture de Wold et doit donc être un bruit blanc. En d'autres termes, si le résidu d'un filtre n'est pas un bruit blanc, alors ce filtre

n'est pas le processus MA, AR ou ARMA adapté à la variable.
En conséquence à l'issue de l'étape d'estimation, la première question que l'on doit se poser est de savoir si on peut ou non rejeter l'hypothèse d'un bruit blanc sur la série résiduelle u. En cas de rejet, il est nécessaire de repartir à l'étape d'identification afin de proposer un autre processus.
On sait qu'un bruit blanc doit vérifier une hypothèse d'homoscédasticité et une hypothèse d'orthogonalité des résidus entre eux. La littérature admet que l'hypothèse la plus importante pour la qualité de l'ajustement est celle d'orthogonalité. Pour cette raison à l'issue de l'estimation, les logiciels d'économétrie réalisent sans intervention de l'utilisateur un certain nombre de tests et proposent des aides graphiques. Ces graphes et tests ne sont évidemment pas réalisés au moyen des résidus théoriques u_t mais sur les résidus empiriques, $\hat{u}_t$ calculés via $\hat{\phi}(L)(x_t - \hat{\mu}_x) = \hat{\theta}(L)\hat{u}_t$. Bien évidemment, rien ne vous interdit de réaliser également des tests d'homoscédasticité, par exemple un test de Chow.

2.1.1 les corrélogrammes

On retrouve en premier lieu les corrélogrammes représentant les fonctions ACF, PACF et IACF déjà vus lors de l'étape d'identification, mais ils portaient alors sur la série de travail x et l'objectif était de repérer et d'interpréter des évolutions caractéristiques sur des corrélations non nulles. Ils vont maintenant traiter de la série résiduelle obtenue après estimation d'un processus et il s'agit de vérifier que toutes ces corrélations sont nulles pour justifier le choix du modèle ajusté. Ainsi, pour chaque $k = 1, 2, 3, \ldots$ il s'agit de tester sur ses résidus empiriques :

— la nullité des autocorrélations,
— la nullité des corrélations partielles,
— la nullité des corrélations inverses.

Un modèle empirique sera naturellement considéré comme satisfaisant si toutes les corrélations estimées sont à l'intérieur de l'intervalle de confiance construit autour de zéro. Attention, l'inverse n'est pas forcément vrai : compte-tenu du seuil de risque pris et du nombre de corrélations calculées on peut s'attendre à ce que quelques-unes sortent de l'intervalle en question même si l'hypothèse nulle est vraie. La question est alors de savoir si l'observation d'une ou deux corrélations significatives justifie la remise en cause du modèle estimé. En pratique la décision à prendre dépend de l'ordre des corrélations. Par exemple s'il s'agit des corrélations de rang 1 et/ou 2, on considère souvent que le modèle doit être corrigé ; s'il s'agit des corrélations aux rangs 9 et 15, on considère généralement qu'elles ne justifient pas un changement de filtre car cela pourrait faire courir un risque de surapprentissage (overfitting). Naturellement s'il s'agit de corrélations aux ordres 12 et 24 sur données mensuelles, ou 7 sur données hebdomadaires, ou 4 sur données trimestrielles la correction doit être envisagée car elles pointent alors toutes sur l'existence d'une saisonnalité non prise en compte par le modèle initial.

2.1.2 le test portmanteau de Ljung-Box

Les corrélogrammes précédents permettent de réaliser des tests ponctuels mais l'hypothèse de bruit blanc implique la nullité de la totalité des corrélations entre u_t et u_{t-k} pour tout $k \neq 0$. C'est donc plutôt un test joint de la forme

$$\rho_{u,k} = 0 \text{ pour } k = 1, 2, \ldots K$$

qui devrait être employé, *i.e.* un test de nullité simultané des K premiers coefficients d'autocorrélation, où K est un entier positif choisi par l'utilisateur. Un premier test est proposé par Box et Pierce. Il est fondé sur le résultat asymptotique selon lequel, sous l'hypothèse nulle, les estimateurs du maximum de vraisemblance $r_{\hat{u},k}$ sont des gaussiennes indépendantes, centrées de variance $1/T$. Comme sous H0, $\sqrt{T}\, r_{\hat{u},k} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, il est alors aisé de construire une statistique de Chi2 :

$$Q(K) = T \sum_{k=1}^{K} r_{\hat{u},k}^2 \tag{1}$$

Il reste à trouver son nombre de degrés de liberté. Pour cela, on retrouve un raisonnement que vous avez déjà rencontré : $Q(K)$ est construit comme une somme de K carrés de gaussiennes centrées réduites indépendantes. Elle devrait donc suivre un $\chi^2(K)$, cependant les corrélations sont estimées sur une série $\hat{u}$ dont la connaissance a nécessité l'estimation d'un ARMA(p,q) à p+q+1 paramètres ce qui, dans le problème de minimisation de la somme des carrés des résidus pour les méthodes CLS et ULS, ou de maximisation de la vraisemblance, a imposé p+q+1 contraintes, réduisant d'autant le nombre de degrés de liberté. Enfin, on sait que la présence d'un terme constant dans un modèle ARMA revient à une opération de centrage sur la variable traitée, et que la soustraction d'une constante à une série d'observations ne change rien à sa structure de corrélation. En conséquence, la présence d'une constante n'entraîne aucune perte de degrés de liberté.

Au final après l'estimation d'un ARMA(p,q), qu'il y ait ou non un terme constant, le nombre de degrés de liberté de la statistique de Box-Pierce $Q(K)$ est de $K-p-q$:

$$Q(K) = T\sum_{k=1}^{K} r^2_{\hat{u},k} \underset{\text{sous H0}}{\overset{d}{\rightarrow}} \chi^2_{(K-p-q)} \tag{2}$$

Peu de temps après la publication de cette statistique, Ljung et Box ont proposé une modification de l'estimation de la variance asymptotique des corrélations empiriques, modification censée améliorer l'adéquation de la distribution de Chi2 à la statistique de test sur petits échantillons. C'est cette statistique de Ljung-Box qui est aujourd'hui implémentée dans les logiciels d'économétrie. Son expression est :

$$Q(K) = T(T+2)\sum_{k=1}^{K} \frac{r^2_{\hat{u},k}}{T-k} \underset{\text{sous H0}}{\overset{d}{\rightarrow}} \chi^2_{(K-p-q)} \tag{3}$$

Dans la proc `ARIMA`, la valeur du nombre maximal de corrélations à considérer dans cette statistique, K, est précisé dans l'option `nlag=` qui doit être mise au niveau de la commande `identify`. Par exemple pour tester la nullité des 36 premiers coefficients de corrélation de la série des résidus on fera :

```
identify var=X nlag=36;
```

Par défaut sa valeur est égale à $Min(24, T/24)$. Notez aussi que la valeur de K étant indiquée au niveau de `identify`, ce nombre va s'imposer aussi bien pour les corrélogrammes de l'étape d'identification qui portent sur la série de travail x elle-même, que sur les corrélogrammes afférents à la série résiduelle $\hat{u}_t$.

Par ailleurs la statistique de Ljung-Box va être construite de façon à tester la nullité des K premiers coefficients corrélations en itérant par pas de 6 . Ainsi, en imposant $K = 26$, vous récupérerez en sortie une table qui testera la nullité des 6 premiers, des 12 premiers, des 18 premiers et pour finir des 24 premiers (dernier multiple de 6 inférieur à 26) coefficients de corrélations. Enfin, aux trois corrélogrammes usuels est ajouté un graphique indiquant les seuils de significativité marginaux de statistiques de Ljung-Box construites avec des nombres de corrélations allant de $p+q+1$ à K. Dans ce graphique des horizontales sont tracées aux niveaux des seuils de risque de 1% et 5%, permettant visuellement de statuer sur l'hypothèse nulle quel que soit le nombre de corrélations retenues et de repérer les corrélations responsables d'un éventuel rejet. Attention, dans ce graphique l'axe des ordonnées est inversé : plus une valeur est petite et plus elle est située en haut de cet axe.

2.1.3 exemples

Les figures 1 et 2 illustrent les développements qui précèdent. Le premier modèle estimé semble satisfaisant au regard des trois corrélogrammes. Seule la statistique de Ljung-Box signale une difficulté au seuil de 5% de risque lorsque l'on teste la nullité des 4 et des 8 premiers coefficients de corrélation. Apparemment ces rejets sont le fait des corrélations de rang 4 et 8. Sur données trimestrielles on peut être tenté d'introduire une composante saisonnière, même si la taille de ces corrélations montre que son pouvoir explicatif serait assez faible.

Les conclusions sont évidemment très différentes avec la figure 2 : le deuxième modèle ajusté est bien sûr inadéquat.

Notez aussi que ces graphes peuvent donner des indications sur les corrections à apporter lorsqu'un processus y apparaît comme insatisfaisant. Supposons par exemple que l'on vienne d'estimer un ARMA(p,q) et que corrélogrammes et test de Ljung-Box signalent la nécessité d'une correction mais :

- que les corrélogrammes indiquent que les résidus semblent gouvernés par un AR(1). On a alors d'une part l'équation initiale inadéquate, $\phi(L)x_t = \theta(L)u_t$, et d'autre part $(1-aL)u_t = \epsilon_t$ où a priori ϵ_t est un bruit blanc. En combinant les deux il vient :$(1-aL)\phi(L)x_t = \theta(L)(1-aL)u_t$, soit $(1-aL)\phi(L)x_t = \theta(L)\epsilon_t$: on peut tenter l'estimation d'un ARMA(p+1,q).
- que les corrélogrammes indiquent que les résidus semblent gouvernés par un MA(1). On a alors d'une part l'équation initiale inadéquate, $\phi(L)x_t = \theta(L)u_t$, et d'autre part $u_t = (1-aL)\epsilon_t$ où a priori ϵ_t est un bruit blanc. En combinant les deux il vient :$\phi(L)x_t = \theta(L)(1-aL)\epsilon_t$, et on peut tenter l'estimation d'un ARMA(p,q+1).

Enfin, lorsque les corrélogrammes ne donnent pas d'indication sur les corrections à apporter à un modèle ayant des résidus autocorrélés, la démarche usuelle est d'augmenter le nombre de retards pris en compte. L'intuition sous-jacente

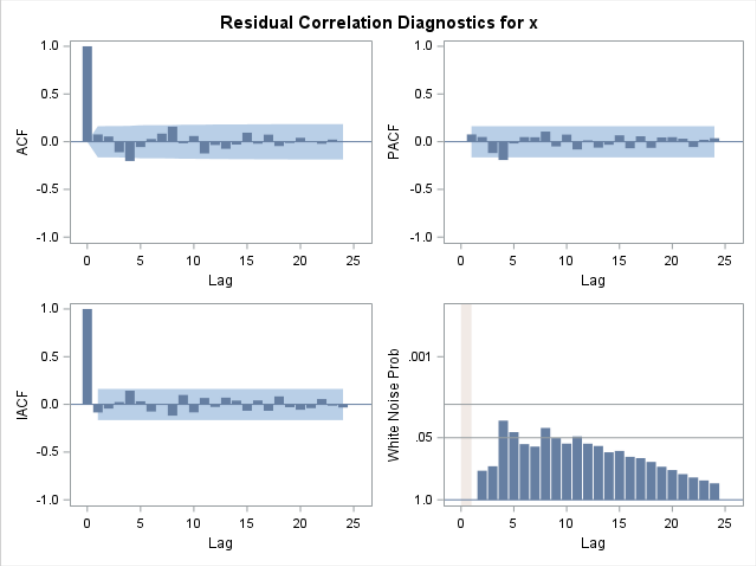

FIGURE 1 – Corrélogrammes et statistique de Ljung-Box n'invalidant pas obligatoirement le processus estimé

est que si les ordres retenus sont inférieurs aux ordres vrais, alors une partie de la structure de dépendance du présent aux observations passées n'est pas bien prise en compte par la partie expliquée et se retrouve donc présente dans la partie résiduelle. Sachant que l'augmentation de la dimension MA peut entraîner des difficultés de convergence des estimateurs, on préfère souvent augmenter la dimension AR. Par exemple, après un ARMA(1,1) repéré comme inadéquat pour cause de corrélation résiduelle, on pourra tenter d'ajuster un ARMA(2,1). Pour une raison explicitée au point suivant, notez cependant que l'on ne doit pas augmenter simultanément la dimension des deux composantes.

2.2 significativité des coefficients et test d'overfitting

2.2.1 construction de tests de student

Vous savez que la variance des estimateurs du maximum de vraisemblance est donnée par l'inverse de la matrice d'information de Fisher. Si on note β le vecteur des paramètres à estimer et $l()$ la log-vraisemblance alors :

$$V(\beta) = I(\beta)^{-1}$$
$$I(\beta) = -E\left[\frac{\delta^2}{\delta\beta\delta\beta^\top} l(X_1, X_2, \ldots, X_T; \beta)\right],$$

Le Hessien étant estimé par : $\widehat{\frac{\delta^2 l()}{\delta\beta\delta\beta^\top}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\delta^2 l(X_t,\beta)}{\delta\beta\delta\beta^\top}$, on trouve donc sur la diagonale de l'opposé de l'inverse de cette matrice les estimateurs,s_i^2, des variances des estimateurs des paramètres. On peut dès lors construire un test de nullité du i^{ème} paramètre au moyen de la statistique $t_i = \hat{\beta}_i / s_i$. L'aide de SAS invite toutefois à une interprétation prudente des résultats de ce test et rappelle que si la taille de la série est faible, que si le nombre de paramètres à estimer est élevé au regard de cette taille et que si le modèle dont le coefficient est testé n'est pas le bon, alors la distribution de la statistique t_i peut décaler sensiblement de la loi de student.
Malgré ces réserves, on préfère généralement retenir un modèle ayant tous ces coefficients significativement non nuls, ce qui suppose donc, à l'issue d'une étape d'estimation, de retirer des coefficients AR et/ou MA non significatifs pour ajuster un modèle plus parcimonieux. Naturellement ces retraits ne seront validés que s'ils n'induisent pas l'apparition d'autocorrélations résiduelles significatives.

2.2.2 overfitting

Cette procédure a été proposée par Box et Jenkins. Il s'agit de conforter le choix d'un processus estimé qui semble satisfaisant, *i.e.* qui aurait par exemple passé les tests précédents. L'overfitting consiste simplement à augmenter d'une

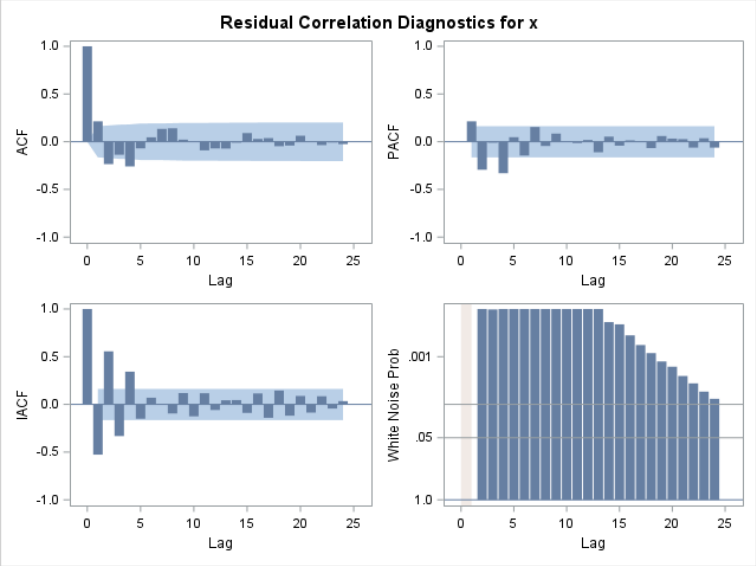

FIGURE 2 – Corrélogrammes et statistique de Ljung-Box invalidant le modèle estimé

unité et successivement les dimensions des polynômes AR et MA et à tester la nullité des coefficients ϕ et θ qui ont été ajoutés. Ainsi pour un ARMA(p,q) on procède en deux étapes :

1. estimation d'un ARMA(p+1,q) et test de H0 :$\phi_{p+1} = 0$.
2. estimation d'un ARMA(p,q+1) et test de H0 :$\theta_{q+1} = 0$.

Naturellement, la validation du modèle initial, ici l'ARMA(p,q) suppose qu'on ne rejette aucune des hypothèses H0. Notez bien qu'on ne doit pas augmenter les deux ordres p et q simultanément. On sait en effet que si x_t a une représentation ARMA(p,q) minimale, il est également gouverné par un ARMA(p+1,q+1), ARMA(p+2,q+2),.... Il n'y a donc rien à tester avec une augmentation simultanée.

3 La normalité des résidus

On pourrait naturellement effectuer un test standard tel le test de Jarque-Bera sur la série des résidus empiriques mais la proc ARIMA de SAS ne délivre automatiquement que deux informations graphiques en ce qui concerne leur normalité.

- un premier graphique superpose un histogramme construit au moyen des valeurs des résidus empiriques à la fonction de densité d'une gaussienne centrée ayant la même variance que celle estimée sur $\hat{u}_t$. Visuellement l'histogramme doit épouser la forme de la courbe en cloche de la densité. Des écarts importants peuvent signaler l'inadéquation de la distribution gaussienne. En plus de la densité précédente, ce même graphique reproduit un estimateur à noyau de la densité des résidus empiriques. L'interprétation est la même : plus l'estimateur non paramétrique est proche de la densité théorique de la gaussienne et moins on doutera de la normalité.
- un Q-Q plot : graphique sur lequel on positionne en abscisse les fractiles théoriques d'une gausienne $\mathcal{N}(0,1)$ et en ordonnées les fractiles empiriques observés dans cette série de résidus. Dans ce graphique, si l'hypothèse de normalité est parfaitement vérifiée, alors les fractiles vont se correspondre exactement et les points ayant en abscisse les fractiles théoriques et en ordonnée les fractiles empiriques devraient être alignés. Plus le nuage de ces points décale d'une droite et plus l'inadéquation de l'hypothèse de normalité est mise en évidence. Un avantage des Q-Q plot est qu'il peut mettre en évidence assez facilement les régions où les écarts à l'hypothèse théorique sont importants et celles où elles ne le sont pas.

La figure 3 reproduit une de ces sorties graphiques. Elle montre clairement que les résidus empiriques ne s'accordent pas ici avec l'hypothèse de normalité : dans le graphe de gauche l'histogramme ne se cale pas sur la densité théorique, notamment dans les queues de la distribution et par ailleurs l'estimation kernel de leur densité dévie de la densité théorique en particulier au voisinage de leur centre. Dans le Q-Q plot de droite, on peut voir que pour les valeurs négatives de

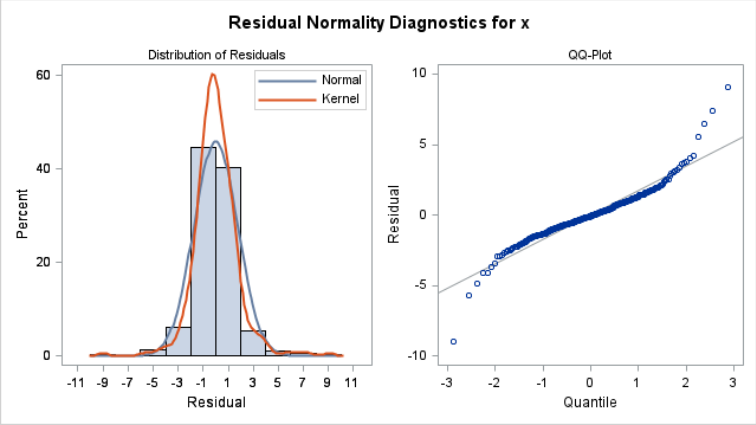

FIGURE 3 – Graphiques relatifs à la normalité des résidus

l'aléatoire, les fractiles empiriques sont atteints beaucoup trop tôt : les premiers points du nuage sont positionnés bien en-dessous de la droite d'alignement : il y a donc trop de grandes valeurs négatives dans l'échantillon des résidus relativement à l'effectif que ces grandes valeurs devraient atteindre en cas de normalité parfaite. Le graphique ne signale ensuite pas de déséquilibre flagrant sur toute la zone centrale. Une distorsion apparaît à nouveau sur les fortes valeurs positives des résidus empiriques : les fractiles empiriques sont au-dessus des valeurs qu'ils devraient avoir : il y a un effectif en excès de grandes valeurs dans cet échantillon. Ainsi, selon le Q-Q plot, l'hypothèse de normalité paraît inadéquat en raison d'un excès de kurtosis dans la série des $\hat{u}_t$.