UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

PARIS 1 PANTHÉON-SORBONNE
ECOLE D'ECONOMIE DE LA SORBONNE
MASTER 1: ECONOMÉTRIE STATISTIQUES

THESIS

ARMA modeling

YAPI Wilfried
ZARICINII Xenia
GAKOSSO Chaveline

Prefessor : M.DE PERETTI

Paris 2020

Abstract

It is common to use information criteria to select ARIMA models, such as AIC, AICc, AICu, BIC or BICc. If asymptotically, these criteria are all equivalent, on small samples, they will not give the same information, and sometimes differ frankly from the ACF and PACF.The purpose of our study is to assess the relevance of the different information criteria on simulated series by varying the sample size and the type of process.We will proceed by simulations of Monte Carlo to estimate the number of times the criteria refers information that corresponds to the data generating process.
Then, we are going to make a forecast for problematic ARMA processes where information criteria choose a different model in most of cases and examine which forecast is better. Lastly we will make a forecast for a real time series CAC 40 index.

Contents

Introduction

ARIMA is an ARMA model with order of integration d. ARIMA can be a non stationary model with ARMA as a stationary model. An ARMA is a model that means an autoregressive–moving-average model that is used In the statistical analysis of time series. This model provide a description of a (weakly) stationary stochastic process in terms of autoregression (AR(p)) model and moving average (MA(q)) model.

AR(p) model is a time series model that uses observations from previous time steps as input to a regression equation to forecast values on the next time steps. This model is trying to describe certain time-varying processes in economics, finance, nature etc.

MA(q) model is a time series model that uses past values of a stochastic term as input to the equation to forecast values on the next time steps.

ARMA model is a combination of both AR and MA models that is used to describe certain time-varying processes.

For the first time ARMA model was describes in Peter Whittle's thesis in 1951 "Hypothesis testing in time series analysis" and met it's popularity in the 1970 book by George E. P. Box and Gwilym Jenkins.

This model has acquired its significance for finance, but there was a problem how to determine order p and q. ACF and PACF couldn't give an exact response, so scientists developed information criteria such as AIC(Akaike criterion) or BIC(Schwarz information criterion) in order to be able to choose p and q correctly. Over the years, new criteria version such as AICc, AICu or BICc began to appear.

In our thesis we are going to verify which criteria are working better among them.

Chapter 1

Theory

1.1 Stationarity

1.1.1 Stationarity notion

To be able to receive correct results while working with time series, some basic assumptions have to be verified. One of common assumptions is that time series should be stationary.
The process can be weakly stationary or second-order stationary and strongly stationary.
The theory for time series is based on the assumption of second-order stationarity. It means that the process $\{Y_t\}_{t=1}^{T}$ is called weakly stationary if for all t and h integers:

Expected value(mean) $\mathbb{E}(Y_t) = \mu$
Covariance $\mathrm{Cov}(Y_t, Y_{t+h}) = \gamma(h)$

Weak Stationarity implies stationarity in mean and variance. It means $\mathrm{Var}(Y_t) = \gamma(0) = \sigma^2$

Process can be also strongly stationary. The process is strict stationary or strongly stationary if

$$(Y_{t_1}, \ldots, Y_{t_k}) \quad \text{and} \quad (Y_{t_1+h}, \ldots, Y_{t_k+h})$$

have the same distribution for all sets of time points $t_1, \ldots, t_k$ and all integers h.
A strict stationary process is automatically weakly stationary, but the converse of this is not generally true.

However, as we are studding ARIMA model and time series, we will only check if our process is weakly stationary. A common example of weak stationary process is a white noise, which is a series of centered variables not correlated with the constant mean.

$$m = 0 \text{ and } \forall h \neq 0, \gamma(h) = 0$$

Below we can observe the plot representing a common example of stationary time series, which is a White Noise. The data with mean zero where all variables have the same variance and each value has a zero correlation with all other values in the series.

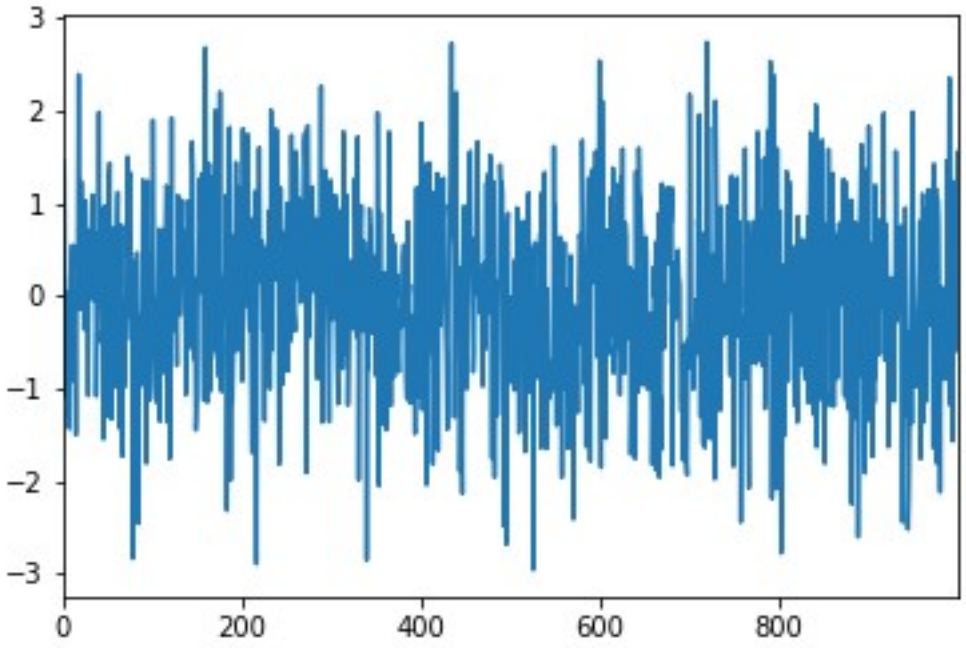


Figure 1.1: White noise plot

1.1.2 Order of integration d.

In this project we are studding an Autoregressive integrated moving average called shortly ARIMA (p,d,q) model where:

p the number of autoregressive terms

d the number of differences

q the number of moving averages

The estimation of ARIMA model suppose time series stationarity. However, real-life data are often not stationary. As an example, series with a seasonal effect or series can exhibit a linear trend over time. Some non-stationary series can be stationary in difference. For that, we subtract from the observation an observation at the previous step.

Y_t can be non-stationary, but $Y_t - Y_{t-1} = \Delta Y_t$ is a stationary process. If a time series needs to be differentiated to become stationary, it is considered as an integrated version of a stationary series.

In the case then we need to differentiate Yt one time, we can say that the process $\{Y_t\}_{t=1}^{T}$ is integrated of order 1 specified $I(1)$ and we can note it as ARIMA (p,1,q). After differentiation and reduction to a stationary form we can note the model as stationary model ARMA(p,q) (Autoregressive moving average).

Therefore, $Y_t \sim ARIMA(p, 0, q)$ and $Y_t \sim ARMA(p, q)$ are equivalents, as the process is integrated of order 0 and thus it is stationary.

For $Y_t \sim ARIMA(p, 1, q)$ a stationary equivalent is $\Delta Y_t \sim ARMA(p, q)$.

Thus, ARIMA model is a generalization of an autoregressive moving average ARMA model.

The process $\{Y_t\}_{t=1}^{T}$ is integrated order d, specified $I(d)$ if the differentiation d times return a stationary process, d being the number of unit roots.

1.1.3 Stationarity statistical tests

As stationarity is one the common assumption for our model ARIMA as well, the test for stationarity is obligatory if the real life data is chosen for the analysis. In the case of data simulation, it is enough to simulate well data. However, real life data or simulated one, the first thing to check and validate is stationarity. For this purpose, we define below some tests to check time series stationarity.

There is a few ways to check if the series is stationary or not. Often we can observe easily stationarity on the plots. Usually we observe on the plots if there is a trend or seasonal effect in the series.

The other way to find out if we have a stationary process or not is to use statistical tests such as Augmented Dichey-Fuller test (ADF), KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test or Phillips–Perron test.

Augmented Dichey-Fuller test (ADF)

Augmented Dichey-Fuller test is the most popular test used to check the presence of a unit root in time series sample. The presence of the unit root indicates that the statistical properties of a given series are not constant with time qnd series is not stationary in time. As the other statistic tests ADF test has a null hypothesis (H0) and alternative hypothesis (H1) for level of confidence 95 % :

H0: $\gamma = 0$ Null hypothesis suggests a non-stationary time series. In other words, the time series has a unit root and the series has some time dependent structure.

H1: $\gamma < 0$ Alternative hypothesis suggests stationarity of the time series.

Test statistic : $DF = \frac{\hat{\gamma}}{\hat{\sigma}(\hat{\gamma})}$

The decision rule: if test statistic $<$ critical value for the Dickey–Fuller test we reject H0

Otherwise we observe the p-value:

p-value > 0.05: We accept H0 and non-stationarity of the series.

p-value ≤ 0.05: We reject H0. The series doesn't have a unit root and it is stationary.

ADF statistic is a negative number and more negative it is more likely we reject the hypothesis H0 that there is a unit root.

Phillips–Perron test (PP test).
Phillips-Perron test is another unit root test used to check the process stationarity. As for the test ADF, its H0

Null hypothesis suggests the presence of the unit root in the univariate time series and the fact that time series is integrated of order 1. This test is built on the Dickey-Fuller test of the null hypothesis $\rho = 0$ for the model $\Delta Y_t = (\rho - 1)Y_{t-1} + u_t$ where Δ is the first difference operator. This test is robust with respect to unspecified autocorrelation and heteroscedasticity in the disturbance process of the test equation.

p-value > 0.05: We accept H0 and non-stationarity of the series.

p-value ≤ 0.05: We reject H0. The series doesn't have a unit root and it is stationary.

As this test is based on the asymptotic theory, it works well with large samples and less well with small samples.

Both ADF and PP tests have the same disadvantage. They are sensitive to structural breaks.

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test.

KPSS test is used to test the stationarity of the process. However, contrary to previous tests described before, the KPSS test Null hypothesis suggests stationarity of the series and alternative hypothesis suggests non-stationary process.

H0: Stationary process (trend stationary).

H1: Non-stationary process. Presence of unit root.

In this test the absence of a unit root is not a proof of stationarity but, a proof of trend-stationarity.

Received t statistic is compared to table of critic values for this test. If the t statistic > Critic value, the Null hypothesis is rejected and the data is not stationary.

This test has one major disadvantage. It has a high rate of Type I errors, which means test tends to reject H0 too often. The way to deal with this problem is to combine this test with ADF and compare results.

1.2 ARIMA

1.2.1 AR

An AR process, called autoregressive, is characterized by the fact its explanatory components are lag of the dependant variable.
An autoregressive process of order p, AR(p), is written like this:

$$\begin{array}{l} Y_t = \mu + \sum_{i=1}^{p} \theta_i Y_{t-i} + \varepsilon_t \\ \text{with } \{\varepsilon_t\}_{t\in\mathbb{Z}} \text{ Weak White Noise} \end{array} \tag{1.1}$$

We can rewrite the model like this:

$$Y_t - \sum_{i=1}^{p} \theta_i Y_{t-i} = \mu + \varepsilon_t \iff \underbrace{(1 - \sum_{i=1}^{p} \theta_i B^i)}_{\Theta(B)} Y_t = \mu + \varepsilon_t \tag{1.2}$$

B: Lag Operator
$\Theta(B)$: Characteristic Polynomial
B_i :Root of Characteristic Polynomial

$$\forall i,\ |B_i| > 1 \iff Y_t \text{ stationary, I(0)} \tag{1.3}$$

Let's take an example of AR(1) without constant:

$$\begin{array}{l} Y_t = \theta Y_{t-1} + \varepsilon_t \ \text{ where } |\theta| < 1 \\ (1-\theta B)Y_t = \varepsilon_t \iff Y_t = (1-\theta B)^{-1}\varepsilon_t \\ Y_t = \sum_{i=0}^{+\infty} \theta^i B^i \varepsilon_t = \sum_{i=0}^{+\infty} \theta^i \varepsilon_{t-i} \end{array} \tag{1.4}$$

Let's check the process is stationary:

$$\begin{array}{l} \mathbb{E}(Y_t) = \frac{\mathbb{E}(\varepsilon_{t-i})}{1-\theta} = 0 \ \ \forall t \in \mathbb{Z} \\ Cov(Y_t, Y_{t-h}) = \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} \theta^i\theta^j \underbrace{Cov(\varepsilon_{t-i}, \varepsilon_{t-h-j})}_{\begin{cases} 0 \text{ if } i \neq h+j \\ \sigma_\varepsilon^2 \text{ if } i = h+j \end{cases}} = \theta^h \frac{\sigma_\varepsilon^2}{1-\theta^2} = \Gamma(h) \quad \forall t \in \mathbb{Z}\ , \forall h \in \mathbb{N} \end{array} \tag{1.5}$$

In this case, white noise is the process of innovations which means that:

$$\begin{array}{l} \underline{Y_{t-1}} = \{Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, ...\} \\ \text{Cov}(\underline{Y_{t-1}}, \varepsilon_t) = 0 \end{array} \tag{1.6}$$

1.2.2 MA

An MA process, called "Moving Average", is characterized by the fact that its explanatory components are a linear combination of weak white noises.
A moving average process of order q, MA (q), is written in this way:

$$\begin{array}{l} Y_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j} \\ \text{avec } (\varepsilon_t)_t \text{ Weak White Noise} \end{array} \tag{1.7}$$

We can rewrite the model like this:

$$Y_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j} \Longleftrightarrow Y_t = \mu + \underbrace{(1 + \sum_{j=1}^{q} \varphi_j B^j)}_{\varphi(B)} \varepsilon_t \tag{1.8}$$

Unlike the AR model, there is no assumption on the coefficients associated with the regressors because the process is always stationary.
Take the example of an MA (1) without constant, $Y_t = \varepsilon_t + \varphi \varepsilon_{t-1}$
Let's check the process is stationary:

$$\begin{array}{l} \mathbb{E}(Y_t) = \mathbb{E}(\varepsilon_t) + \varphi \times \mathbb{E}(\varepsilon_{t-1}) = 0 \ \ \forall t \in \mathbb{Z} \\ \text{Cov}(Y_t, Y_{t+h}) = \sigma_\varepsilon^2 (1 + \varphi^2) \times \mathbb{1}_{h=0} + \sigma_\varepsilon^2 \times \varphi \times \mathbb{1}_{h=1} = \Gamma(h) \ , \forall h \in \mathbb{N} \ \forall t \in Z \end{array} \tag{1.9}$$

We can see that the expression of a stationary AR (1) can be rewritten in MA (∞) as demonstrated in the previous subsection.

The difference between an AR and an MA is the fact that white noise, at time t, has an indirect but infinite effect on the variable to be explained in the AR process, while in the MA process this effect is direct but only on q dates.
The inversion of the characteristic polynomial allows us to pass from an AR model to an MA model and vice versa. This method allows us to choose the model where the order is the least important and therefore to have fewer parameters to estimate.

1.2.3 ARMA

An ARMA process is a process that has an AR component and an MA component. An ARMA (p,q) is written:

$$\begin{array}{l} Y_t = \underbrace{\sum_{i=1}^{p} \theta_i Y_{t-i}}_{AR(p)} + \underbrace{\varepsilon_t + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j}}_{MA(q)} \\ \Longleftrightarrow Y_t - \sum_{i=1}^{p} \theta_i Y_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j} \\ \Longleftrightarrow \Theta(B) Y_t = \varphi(B) \varepsilon_t \end{array} \tag{1.10}$$

The ARMA model can allow us to make dynamic forecasts (like the AR process and the MA process) in the following way:

${}_t\hat{Y}_{t+h}$ is the forecast, at time t, of the value in t + h

$${}_t\hat{Y}_{t+h} = \sum_{i=1}^{p} \hat{\theta}_i \tilde{Y}_{t+h-i} + \sum_{j=0}^{q} \hat{\varphi}_j \tilde{\varepsilon}_{t+h-j} \tag{1.11}$$

where estimates are made by OLS With

$$\tilde{Y}_{t+h-i} = \begin{cases} Y_{t+h-i} \text{ if } t+h-i < t \\ \hat{Y}_{t+h-i} \text{ if } t+h-i \geq t \end{cases} \tag{1.12}$$

$$\tilde{\varepsilon}_{t+h-j} = \begin{cases} \hat{\varepsilon}_{t+h-j} \text{ if } t+h-j < t \\ 0 \text{ if } t+h-j \geq t \end{cases} \tag{1.13}$$

Here we consider that the process is stationary, which amounts to saying that the AR part has all its roots which are greater than 1 in absolute value.

This process $\{Y_t\}_{t=1}^{T}$ is defined like this:

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_T \end{bmatrix}}_{\substack{Y \\ (T,1)}} = \underbrace{\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ Y_1 & 0 & 0 & \dots & \varepsilon_1 & 0 & 0 & \dots \\ Y_2 & Y_1 & 0 & \dots & \varepsilon_2 & \varepsilon_1 & 0 & \dots \\ Y_3 & Y_2 & Y_1 & \dots & \varepsilon_3 & \varepsilon_2 & \varepsilon_1 & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots \\ Y_{T-1} & Y_{T-2} & Y_{T-3} & \dots & \varepsilon_{T-1} & \varepsilon_{T-2} & \varepsilon_{T-3} & \dots \end{pmatrix}}_{\substack{X \\ (T,p+q)}} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \\ \varphi_1 \\ \vdots \\ \varphi_q \end{bmatrix}}_{\substack{\beta \\ (p+q,1)}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_T \end{bmatrix}}_{\substack{\varepsilon \\ (T,1)}} \tag{1.14}$$

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \\ \hat{\varphi}_1 \\ \vdots \\ \hat{\varphi}_q \end{bmatrix} \tag{1.15}$$

where $\tilde{X} = \underset{(T-\max\{p,q\},p+q)}{X}$ and $\tilde{Y} = \underset{(T-\max\{p,q\},p+q)}{Y}$

1.2.4 ARIMA

An ARIMA (p, d, q) is an ARMA which is I (d), which therefore has a unit root of order d. So we will write the model this way by adding a term to the characteristic polynomial of the AR part:

$$(1 - \sum_{i=1}^{p} \Theta_i B^i) \underbrace{(1-B)^d Y_t}_{Y_{d,t}} = (1 + \sum_{j=1}^{q} \varphi_j B^j)\varepsilon_t \tag{1.16}$$

$Y_{d,t} = (1-B)^d Y_t = \Delta^d Y_t$ is stationary because it is the method of differentiation which transforms into I (0).

1.3 AutoCorrelation Function and Partial AutoCorrelation Function

1.3.1 AutoCorrelation Function

When we compare several series, it is more useful to consider the autocorrelation, which is the corrected covariance of the variance.
The autocorrelation function is the noted function k that measures the correlation of the series with itself lagged from k periods:

$$\rho_k = \frac{\mathrm{Cov}(Y_t, Y_{t-k})}{\sigma_{Y_t}\sigma_{Y_{t-k}}} = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y}_n)(Y_{t-k} - \overline{Y}_n)}{\sqrt{\sum_{t=k+1}^{n}(Y_t - \overline{Y}_n)^2}\sqrt{\sum_{t=k+1}^{n}(Y_t - \overline{Y}_n)^2}} = \frac{\gamma(k)}{\gamma(0)} \tag{1.17}$$

We can deduce from this that:
ρ_0=1 and $-1 \leq \rho_k \leq 1 \ \forall k$

In general, the autocorrelation is used to characterize linear dependencies in residual series (i.e. trend-adjusted time series and the season). Indeed, the trend and the season are deterministic components and it makes little sense to estimate statistical properties of deterministic quantities. Moreover, if the characteristics of the series studied change over the time, it can be difficult to estimate its statistical properties because there is usually a single completion of the process which is not sufficient to make the estimate. However, it is very useful to understand the pace of empirical autocorrelation of a raw series with a trend and/or season.

1.3.2 Partial AutoCorrelation Function

The partial autocorrelation measures the correlation between Y_t and Y_{t-k}, without the influence of other variables $(Y_{t-1}, Y_{t-2}, \dots, Y_{t-k-1})$.

$$PACF(k) = Corr(Y_t, Y_{t-k} | Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-k-1}) \tag{1.18}$$

1.3.3 ARMA Properties and Correlogram

Property 1: the autocorrelation coefficients (k) of a MA(q) are null for k>q.
Property 2: the partial autocorrelation coefficients PAC(h) of an AR(p) are null for k>p.

The graph of the autocorrelation function is called a correlogram. The partial correlogram represents the partial autocorrelation coefficient depending on the lag. To establish a rule as to the behaviour of the autocorrelation function , two cases are to be considered:
• If p > q, the autocorrelation function behaves like a mixture exponential/sinusoidal functions cushioned.
• If q ≥ p, the q−p+1 first values of the auto-correlation have their own behaviour, and for k ≥ q−p + 1, autocorreogram tends to 0. Symmetrical properties exist for partial autocorreogram.

It should be noted that for economic series, it is rare to have d≥ 3. Therefore, we can conclude that there are 5 scenarios of evolution of the ACF and the PACF: (1) A autoregressive parameter (p): ACF - exponential decomposition; PACF - a peak in period 1, no correlation for other periods.

(2) Two autoregressive parameters (p): ACF - a sinusoidal form component or a set of exponential decompositions; PACF - peaks at periods 1 and 2, no correlation for other periods.

(3) A moving average parameter (q): ACF - peak in period 1, no correlation for other periods; PACF - exponentially depreciated.

(4) Two moving average parameters (q): ACF - peaks at periods 1 and 2, no correlation for other periods; PACF - a sinusoidal shape component or a set of exponential decompositions.

(5) A autoregressive parameter (p) and a moving average (q): ACF - exponential decomposition beginning in period 1; PACF - exponential decomposition beginning in period 1.

1.3.4 Identifying AR and MA terms in an ARIMA process

In addition, the ACF and the PACF are easy-to-use tools to recognize autoregressive and moving average models.

1.3.4.1 Identifying AR terms in an ARIMA process

A sudden extinction of partial autocorrelation associated with a more gradual decline in autocorrelation is the sign of an autoregressive process. In particular, the partial autocorrelation of k lag is equal to the AR(k) coefficient estimated in a model containing k AR terms. Multiple regression AR coefficients could be determined by predicting $(y_t - y_{t-1})$ from k samples representing k lags. The lag at which the partial autocorrelation disappears indicates the number of autoregressive terms to be included.

Generally, this pattern is associated with a positive 1 lag autocorrelation, a sign that the series remains under-differentiated. A slight under-differentiation can therefore be compensated by the addition of a autoregressive term.

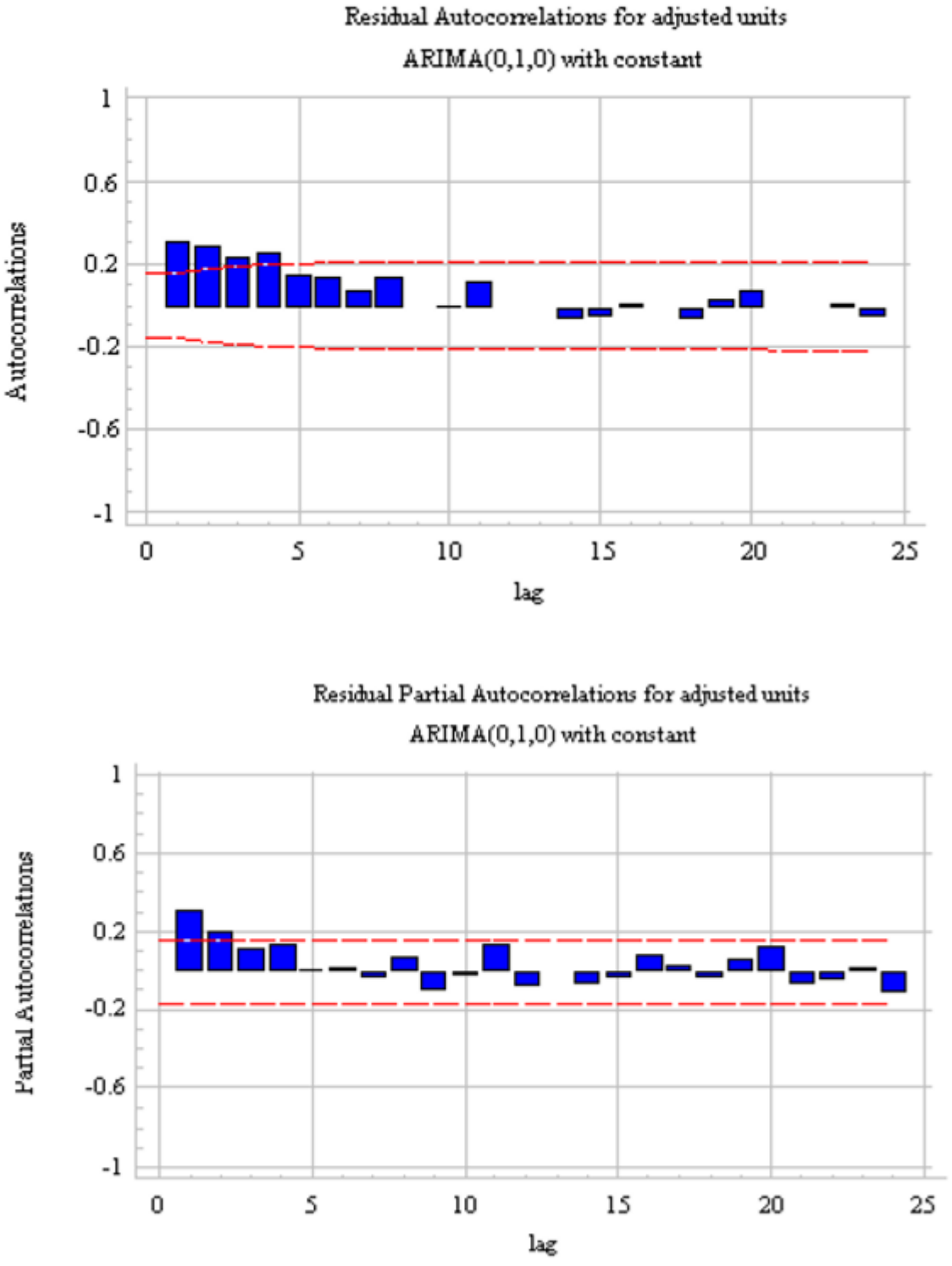


Figure 1.2: Identifying AR terms using ACF and PACF plots

We can note that the autocorrelation of lag 1 is significant and positive, and that the partial autocorrelation presents a more marked extinction than the autocorrelation. The partial autocorrelation actually has only two significant peaks, while the autocorrelation has four. The differentiated series therefore presents a sign of autoregressive processes of order 2.
The adjusting of the series with an ARIMA model (2,1,0) gives the following ACF and PACF functions:

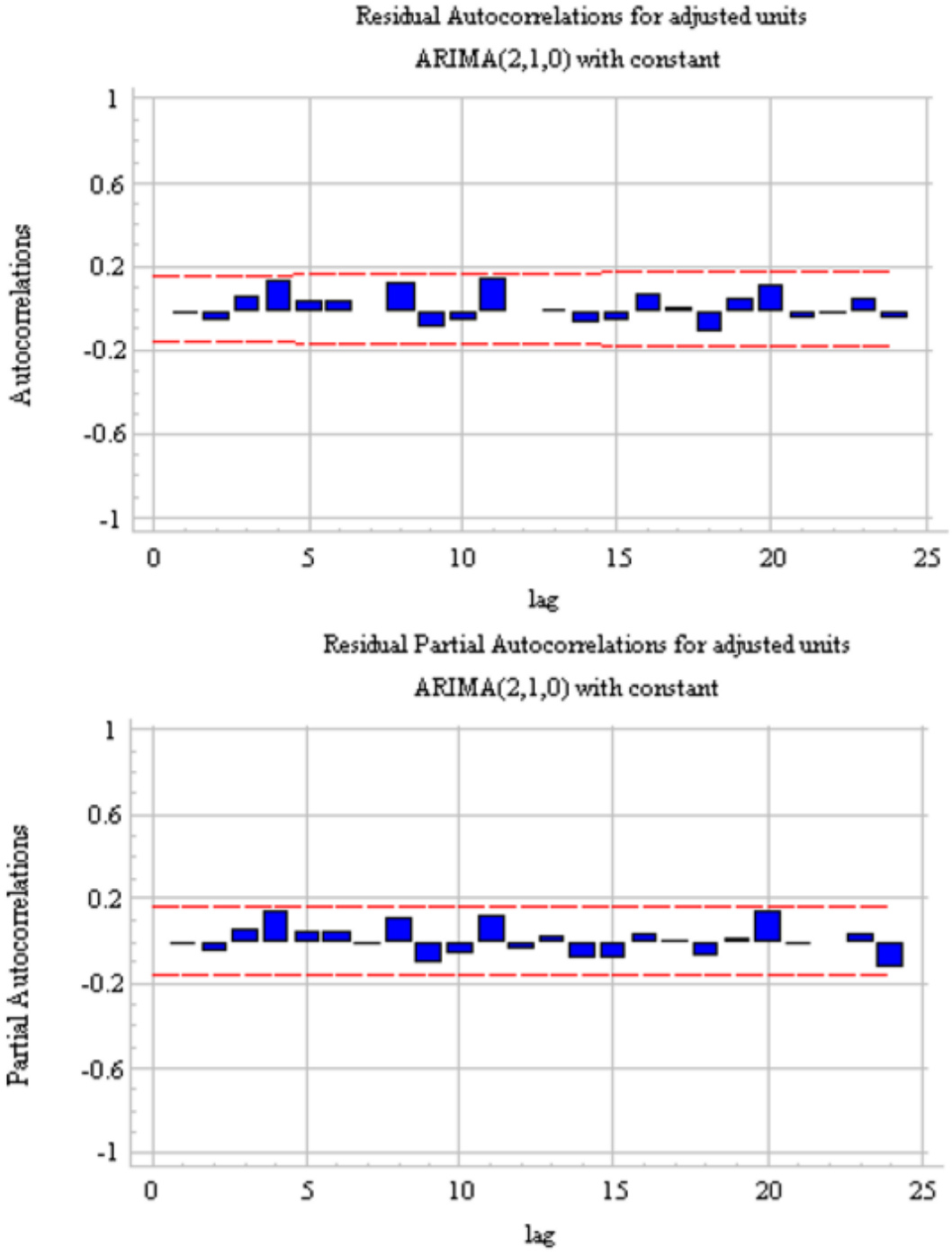


Figure: Adjusting of the series using AR terms

The autocorrelations to lags 1 and 2 have been removed, and no autocorrelation appears at the upper lags.

1.3.4.2 Identifying MA terms in an ARIMA process

If the autocorrelation is significant to the k lag but more to the k-1 lag, this indicates that k moving average terms need to be added to the model.

We can note that while AR coefficients can be estimated by a multiple regression analysis, such an approach is impossible for MA coefficients. On the one hand, because the prediction equation is non-linear, and on the other hand errors cannot be specified as independent variables. Errors should be calculated step by step based on current parameter estimates.

An MA signature is usually associated with a negative autocorrelation to lag 1, a sign that the series is over-differentiated. A slight over-differentiation can therefore be compensated by the addition of a moving average term.

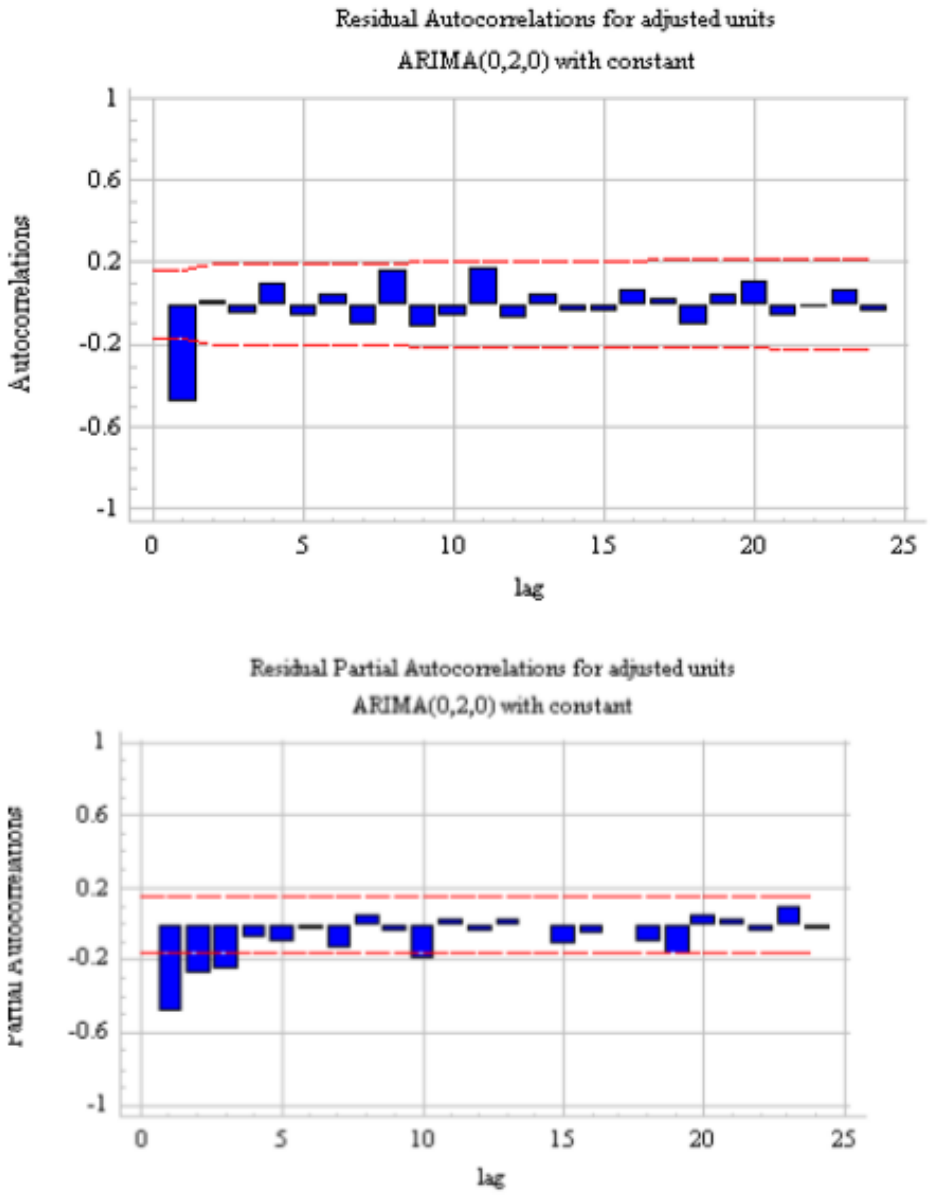


Figure: Adjusting of the series using MA terms

These functions are typical of the moving average processes. The single, negative peak of the autocorrelation indicates an MA process(1). The relevant model is therefore an ARIMA (0.2,1).

Knowing that the terms AR can compensate for a slight under differentiation, and the MA terms a slight over-differentiation, it is common that two alternative models are possible to adjust the starting series: one with 0 or 1 order of differentiation combined with AR terms, and another with the higher level of differentiation, combined with MA terms. The choice of the one or other model may be based on theoretical assumptions related to the observed phenomenon.

Thanks to ACF and PACF, it's possible to obtain the order of MA and AR. However, it could be difficult to distinguish them. In that case, it's possible to take the upper bounds for the orders and select a model minimizing a penalized criterion of type AIC or BIC.

1.4 Information criteria

An information criterion is a criterion based on the predictive power of the model that considers the number of parameters to be estimated. They are constructed as functions of the variance of the estimated residues of the model and the number of parameters to be estimated. The aim is to minimize these functions in relation to these two arguments (application of the principle of parsimony).

We will keep only two: the Akaike information criterion (1973) and the Schwarz information criterion (1978).

1.4.1 Akaike Information Criterion

The Akaike information criterion (AIC) proposed by Hirotugu Akaike in 1973, is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data.Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

$$AIC = 2k - 2\ln(L) \tag{1.19}$$

When a statistical model is estimated, it is possible to increase the likelihood of the model by adding a parameter. In order to ensure that there is no overfitting of the model, a penalty term is added to make a trade-off between the number of parameters and the minimum variance.
Therefore, the Akaike information criterion tends to choose a larger number of parameters than the real model, which leads us to a smaller variance of estimated residues. The model is then chosen with the lowest Akaike information criterion.

1.4.2 AICc

When the sample size is small, there is a strong probability that AIC will select models that have too many parameters, i.e. that AIC will overfit. To address such potential overfitting, AICc was developed. Proposed by Hurvich and Tsai in 1989, AICc is AIC with a correction for small sample sizes.

The formula for AICc depends upon the statistical model. Assuming that the model is univariate, is linear in its parameters, and has normally-distributed residuals (conditional upon regressors), then the formula for AICc is as follows:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{1.20}$$

where n denotes the sample size and k denotes the number of parameters. Thus, AICc is AIC with an extra penalty term for the number of parameters.

However, if the assumption that the model is univariate and linear with normal residuals does not hold, then the formula for AICc will generally be different from the formula above. For some models, the formula can be difficult to determine. Indeed, For every model that has AICc available, though, the formula for AICc is given by AIC plus terms that includes both k and k^2. In comparison, the formula for AIC includes k but not k^2. In other words, AIC is a first-order estimate (of the information loss), whereas AICc is a second-order estimate.

Moreover, when the number of k parameters is the same for all models, then the AICc and the AIC will have identical relative values. In this situation the AIC can always be used. Note that when $n \to \infty$, the extra penalty term converges to 0, and thus AICc converges to AIC.

Burnham and Anderson (2002) strongly recommend the use of the AICc instead of the AIC if n is small and/or k large and show that a bootstrapping estimate allows the AICc to tend towards the AIC when n becomes large. Brockwell and Davis (1991) recommend using the AICc as the primary criterion for selecting ARMA models for time series. McQuarrie and Tsai (1998) confirmed the interest of the AICc with many simulations on regressions and time series.

1.4.3 Bayesian Information Criterion

There are many criteria for information based on the Akaike criterion. Of these, the Bayesian information criterion, also known as Schwarz information criterion (SBC) is one of the most popular. It is defined as:

$$BIC = -2\ln(L) + \ln(n)k \tag{1.21}$$

with n the number of observations in the sample studied, L the likelihood of the sample and k the number of parameters. This criterion has the advantage of penalizing models with excess parameters more heavily and selecting models of smaller size than the AIC criterion (when n$>$ 7).

1.4.4 Under ARIMA Model

To determine the order of a non-seasonal ARIMA model, a useful criterion is the **Akaike information criterion (AIC)** . It is written as :

$$AIC(p,q) = \ln\left(\hat{\sigma}_{\varepsilon}^2\right) + 2\frac{(p+q)}{T} \tag{1.22}$$

where L is the likelihood of the data, p is the order of the autoregressive part and q is the order of the moving average part. The k represents the intercept of the ARIMA model. For AIC, if k = 1 then there is an intercept in the ARIMA model (c $\neq$ 0) and if k = 0 then there is no intercept in the ARIMA model (c = 0).

The **corrected AIC (AICc)** for ARIMA models can be written as :

$$AICc(p,q) = AIC + 2\frac{(p+q+c)(p+q+c+1)}{T-p-q-c-1} \tag{1.23}$$

The **corrected AIC (AICu)** using an unbiased variance estimate :

$$AICu(p,q) = \ln\left(s_{p+q}^2\right) + \frac{T+p+q}{T-p-q-2} \tag{1.24}$$

where s_{p+q}^2 is the unbiased variance and $p+q$ the parameters

The **Bayesian Information Criterion (BIC)** can be written as :

$$BIC(p,q) = \ln\left(\hat{\sigma}_{\varepsilon}^2\right) + ln(T)\frac{(p+q)}{T} \tag{1.25}$$

The **Corrected BIC (BICc)**:

$$BICc(p,q) = \ln\left(\hat{\sigma}_{\varepsilon}^2\right) + (p+q)\frac{\ln(T)}{T-p-q-1} \tag{1.26}$$

Comparaison between AIC and BIC

The objective is to minimize the AIC, AICc or BIC values for a good model. The lower the value of one of these criteria for a range of models being studied, the better the model will suit the data. Generally the simulated models are very simple, therefore , the BIC criterion selects the real model and AIC the real model or a larger model, leading the authors to conclude that BIC is more efficient for choosing the real model. However, when the model is more complex, for example composed of a multitude of "small effects", BIC becomes less efficient than AIC because even for large sample sizes, BIC selects under-adjusted models.
The AIC and the BIC are used for two completely different purposes . While the AIC tries to approximate models towards the reality of the situation, the BIC attempts to find the perfect fit. The BIC approach is often criticized as there never is a perfect fit to real-life complex data ; however, it is still a useful method for selection because it penalizes models with a large number of parameters more heavily than the AIC would. AICc can only be used to compare ARIMA models with the same orders of differencing. For ARIMAs with different orders of differencing, RMSE can be used for model comparison.

However we can see all the indicators converge towards the same value and are thefore identical.

Chapter 2

Series Simulation

2.1 Simulation

We will explain what is the code and the logic of our simulation. First, we will simulate data with the Armasim function having a precise ARMA identification. In a second step, we will use 2 different methods to try to find the ARMA. And finally, we will compare these 2 methods by looking if there is a difference and, if so, which is the most effective method.

The first step is to create a matrix allowing us to obtain the lags.
We will initialize a matrix on which, through a loop, we will incorporate different orders of the Identity Matrix.

```
proc iml;
start ARMA(y,p,q);
p1=4;
Free X1;
do i=1 to p1;
S=j(nrow(Y),nrow(Y),0); /* Selection Matrix*/
id=I(nrow(Y)-i);    /* Identity Matrix*/
S[i+1:nrow(Y),1:nrow(Y)-i]=id;
X1=X1||S*Y;
end;
/* OLS */
/* beta = inv(X'X)X'Y */
beta=inv(t(X1[p1+1:nrow(X1),]) * X1[p1+1:nrow(X1),])*t(X1[p1+1:nrow(X1),]) *y[p1+1:nrow(X1),];
Residu=Y-X1*beta;
```

Let's take an example where we take three lags of the series.

$$S_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \boxed{1} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad S_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \boxed{1} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad S_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \boxed{1} & 0 & 0 & 0 \end{pmatrix} \tag{2.1}$$

By multiplying the simulated vector by S matrices, we obtain the lag vectors.

$$S_1 \times Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \quad S_2 \times Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ Y_1 \\ Y_2 \end{bmatrix} \quad S_3 \times Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ Y_1 \end{bmatrix}$$

We concatenate those vectors to obtain a matrice of lags.

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ Y_1 & 0 & 0 \\ Y_2 & Y_1 & 0 \\ Y_3 & Y_2 & Y_1 \end{pmatrix} \tag{2.2}$$

OLS formula allow us to obtain residual. Those residuals can replace efficiently the errors because of convergence in probability.

$$\begin{cases} \hat{\tilde{\beta}} = (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{Y} \\ \hat{\varepsilon} = Y - \tilde{X}_1\hat{\tilde{\beta}} \end{cases} \tag{2.3}$$

The second step consists in repeating this procedure on the simulated series but also on the estimated residuals, which allows us to have the Explanatory Matrix incorporating terms AR (p) and terms MA (q).

```
/*AR(p)*/
Free X2;
do i=1 to p;
S=j(nrow(Y),nrow(Y),0); /* Selection Matrix */
id=I(nrow(Y)-i);   /* Identity Matrix*/
S[i+1:nrow(Y),1:nrow(Y)-i]=id;
X2=X2||S*Y;
end;
/*MA(q)*/
Free X3;
do i=1 to q;
S=j(nrow(Y),nrow(Y),0); /* Selection Matrix */
id=I(nrow(Y)-i);   /* Identity Matrix*/
S[i+1:nrow(Y),1:nrow(Y)-i]=id;
X3=X3||S*residu;
end;
X=X2||X3;
beta=inv(t(X[a+1:nrow(X),]) * X[a+1:nrow(X),])*t(X4[a+1:nrow(X),]) *y[a+1:nrow(X),];
return(beta);
Finish Arma;
```

$$X = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ Y_1 & 0 & 0 & \dots & \hat{\varepsilon}_1 & 0 & 0 & \dots \\ Y_2 & Y_1 & 0 & \dots & \hat{\varepsilon}_2 & \hat{\varepsilon}_1 & 0 & \dots \\ Y_3 & Y_2 & Y_1 & \dots & \hat{\varepsilon}_3 & \hat{\varepsilon}_2 & \hat{\varepsilon}_1 & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots \\ Y_{T-1} & Y_{T-2} & Y_{T-3} & \dots & \hat{\varepsilon}_{T-1} & \hat{\varepsilon}_{T-2} & \hat{\varepsilon}_{T-3} & \dots \end{pmatrix} \tag{2.4}$$

Again, we use the OLS formula. It will allow us to have the estimation of the terms AR and of the terms MA.

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \\ \hat{\varphi}_1 \\ \vdots \\ \hat{\varphi}_q \end{bmatrix}$$

where $\tilde{X} = \underset{(T-\max\{p,q\},p+q)}{X}$ and $\tilde{Y} = \underset{(T-\max\{p,q\},p+q)}{Y}$

We have finished initializing the Armasim Function, we are going to incorporate it into a precise DGP by enumerating values for the beta vector as long as it allows us to have an ARMA process (ARIMA which is I (0)).
For example, we will assign the ARMASIM function to the following realization: ARMA (2,1) such as :
$Y_t = 0.5Y_{t-1} - 0.8Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1}$

```
/*ARMA(2,1) with T=500*/
phi={1 -0.5  0.8}; /*AR*/
theta={1 0.5}; /*MA*/
y21500=armasim(phi,theta, 0,1,500,0); /*  y=armasim(phi,theta,mean,standard error,T,seed)*/
y21500=y21500-y21500[:] ; /* Center the serie */
beta21500 = ARMA(y21500,2,1);
print beta21500;
```

Thereafter, we will try to find the DGP through different methods.

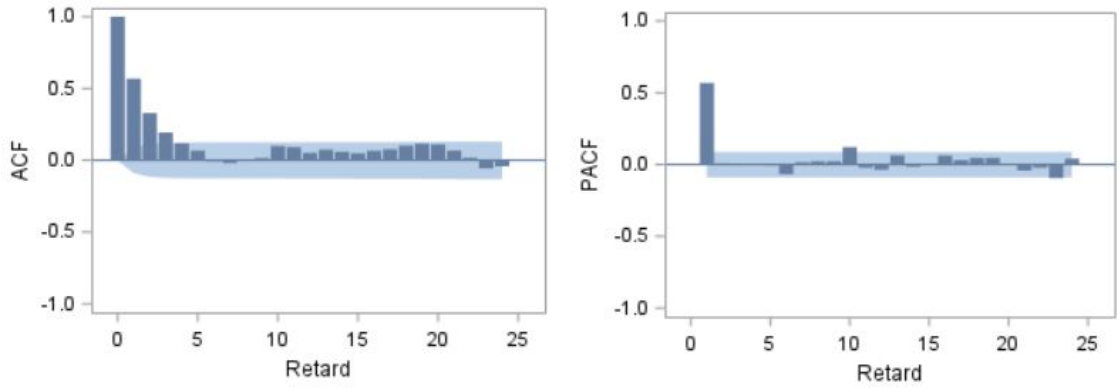
2.2 Analysis with ACF and PACF

	ACF	PACF
MA(q)	Break after q periods	No break but constant decreasing
AR(p)	No break but constant decreasing	Break after p periods

In this section, we will use the previous table to qualify the characterization of the different series when possible. We will use a graphical and non-analytical approach

1st simulation: ARMA(1,0) with 500 observations such as

$$Y_t = 0.5Y_{t-1} + \varepsilon_t \tag{2.5}$$

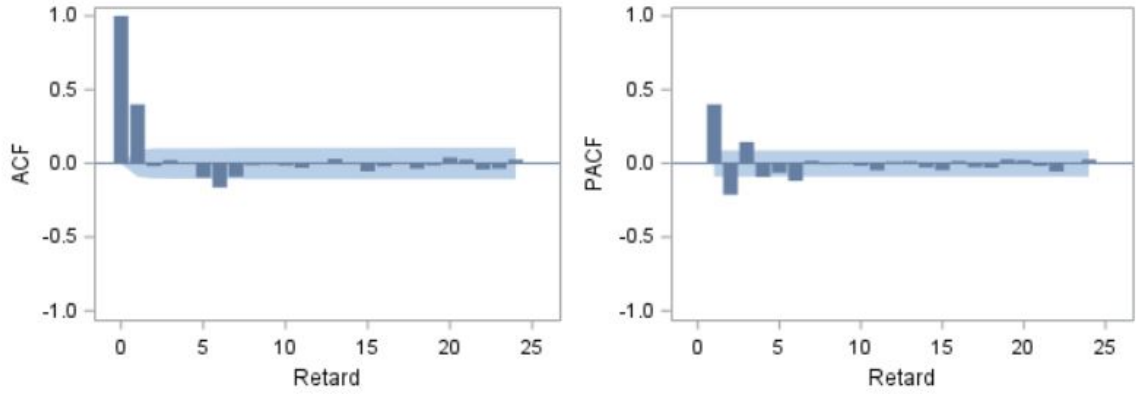


The graph tells us that the DGP could be an AR (1) defined like this:

$$\hat{Y}_t = 0.568\hat{Y}_{t-1}$$

The graphics do indeed determine the DGP. 2nd simulation: ARMA (0.1) with 500 observations such as

$$Y_t = \varepsilon_t + 0.5\varepsilon_{t-1} \tag{2.6}$$



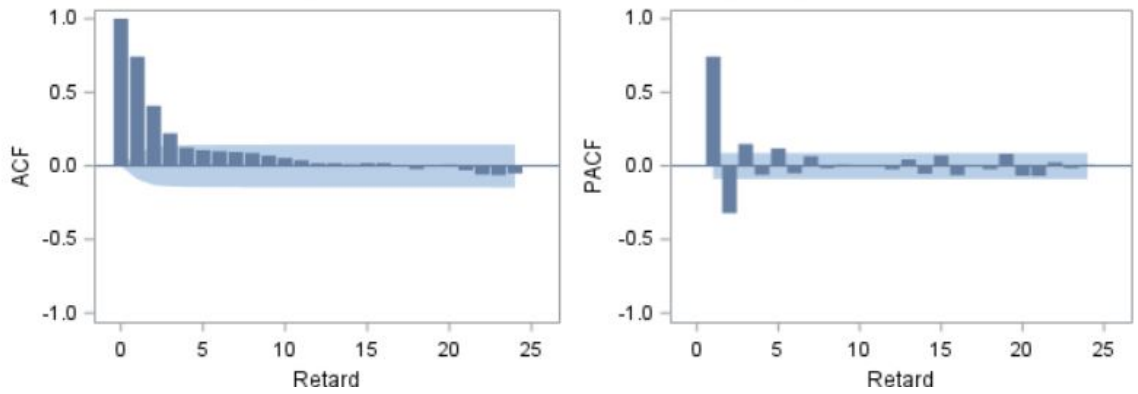
The graph tells us that the DGP could be an MA (1) defined like this:

$$\hat{Y}_t = \hat{\varepsilon}_t + 0.399\hat{\varepsilon}_{t-1}$$

The graphics do indeed determine the DGP.

3rd simulation: ARMA(1,1) with 500 observations such as

$$Y_t = 0.5Y_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1} \tag{2.7}$$



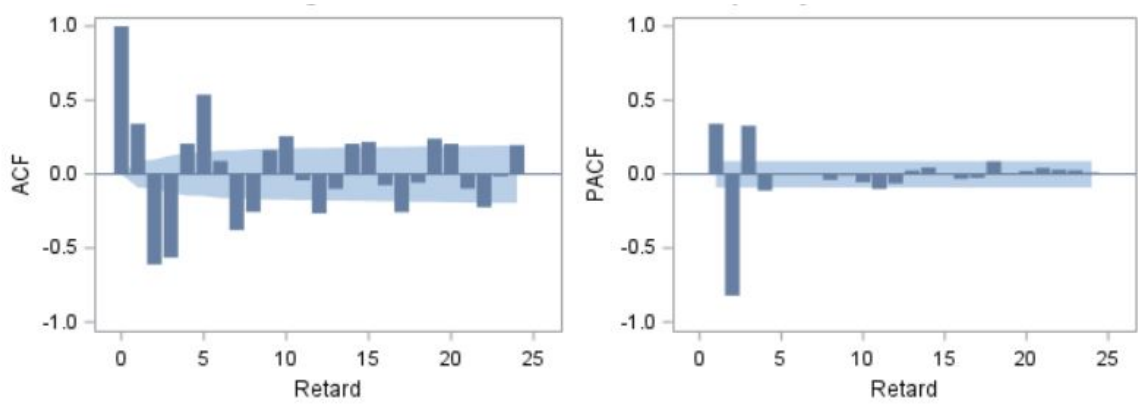
The graph tells us that the DGP could be an AR (2) defined like this:

$$\hat{Y}_t = 0.743\hat{Y}_{t-1} - 0.321\hat{Y}_{t-1}$$

The graph does not allow us to find the initial DGP

4th simulation: ARMA(2,1) with 500 observations such as

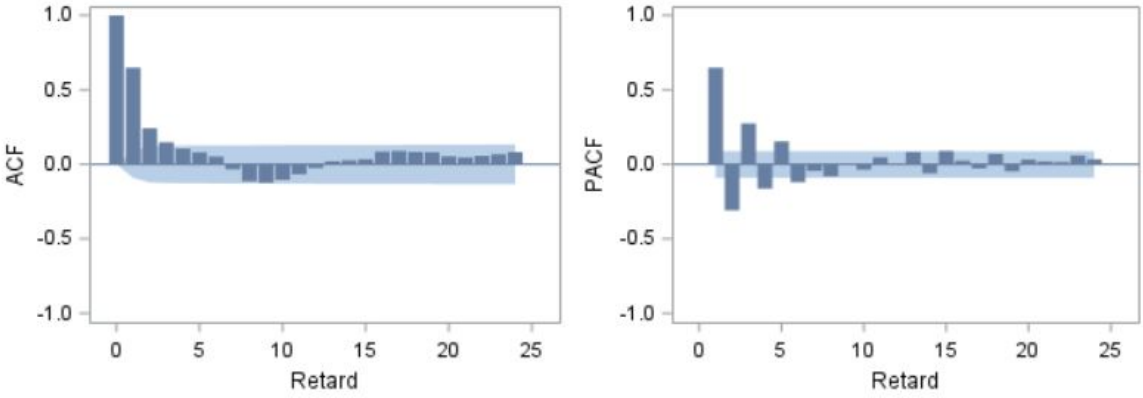
$$Y_t = 0.5Y_{t-1} - 0.8Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1} \tag{2.8}$$



The graphics do not give a clear indication so we cannot conclude.

5th simulation: ARMA (1,2) with 500 observations such as

$$Y_t = 0.5Y_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1} - 0.3\varepsilon_{t-2} \tag{2.9}$$



The graphics do not give a clear indication so we cannot conclude.
Finally, we can rewrite the table at the beginning in this form:

	ACF	PACF
MA(q)	Break after q periods	No break but constant decreasing
AR(p)	No break but constant decreasing	Break after p periods
ARMA(p,q)	Can't use, Need Information Criterion	Can't use, Need Information Criterion

2.3 Analysis with Information criterion

2.3.1 Maximum Likelihood estimation

In this part we have estimated a few ARMA models with Maximum likelihood estimation method in order to analyse information criterion. The ARMA models estimated are following:

- AR(1): $Y_t = 0.8Y_{t-1} + \varepsilon_t$
- MA(1): $Y_t = \varepsilon_t + 0.75\varepsilon_{t-1}$
- ARMA(1,1): $Y_t = 0.8Y_{t-1} + \varepsilon_t + 0.7\varepsilon_{t-1}$
- ARMA(1,2): $Y_t = 0.5Y_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1} - 0.25\varepsilon_{t-2}$
- ARMA(2,1): $Y_t = 0.5Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1}$

On the tables below we can observe AIC and BIC information criteria for all ARMA(p,q) processes simulated and estimated with Maximum likelihood method using PROC ARIMA. AR(p) in the tables represent the order p of estimated process and MA(q) represent respectively the order q of ARMA models.

AIC, T=50

	MA(0)	MA(1)	MA(2)
AR(0)		140.9976	
AR(1)	148.1873	160.0553	148.5981
AR(2)		150.2665	

BIC, T=50

	MA(0)	MA(1)	MA(2)
AR(0)		144.8216	
AR(1)	152.0114	165.7914	156.2462
AR(2)		157.9146	

AIC, T=100

	MA(0)	MA(1)	MA(2)
AR(0)		281.4846	
AR(1)	305.6951	279.4987	279.743
AR(2)		277.834	

BIC, T=100

	MA(0)	MA(1)	MA(2)
AR(0)		286.695	
AR(1)	310.9055	287.3143	290.1637
AR(2)		288.2547	

AIC, T=150

	MA(0)	MA(1)	MA(2)
AR(0)		423.0435	
AR(1)	427.3142	423.9051	401.9739
AR(2)		439.7415	

BIC, T=150

	MA(0)	MA(1)	MA(2)
AR(0)		429.0648	
AR(1)	433.3355	432.937	414.0164
AR(2)		451.784	

AIC, T=200

	MA(0)	MA(1)	MA(2)
AR(0)		563.9262	
AR(1)	592.1547	566.9041	566.0188
AR(2)		565.7725	

BIC, T=200

	MA(0)	MA(1)	MA(2)
AR(0)		570.5228	
AR(1)	598.7513	576.799	579.2121
AR(2)		578.9658	

AIC, T=500

	MA(0)	MA(1)	MA(2)
AR(0)		1397.169	
AR(1)	1375.187	1422.659	1413.908
AR(2)		1362.068	

BIC, T=500

	MA(0)	MA(1)	MA(2)
AR(0)		1405.598	
AR(1)	1383.616	1435.303	1430.766
AR(2)		1378.927	

Firstly, looking the AIC and BIC tables we can observe that AIC information criterion is always less than BIC criterion which is logically correct as in BIC formula we multiply our penalty by lnT (logarithm of the sample size) instead of 2 and LnT is always less than 2 in our cases.

$$AIC = \frac{-2l}{T} + \frac{2(P+q)}{T}$$
$$BIC = \frac{-2l}{T} + \frac{(P+q)\log(T)}{T}$$

Where l is a log likelihood function.
Then, we have estimated the same processes changing each time order p and q to check if the information criteria determine correctly the process. In the vast majority of cases both AIC and BIC determine correctly the ARMA estimated and have minimul values across all samples used. However, sometimes they define a different model than the model used. In particularly, they choose an incorrect process on the small samples. For instance, while testing T=50 sample for ARMA(1,2) both AIC(147.7602) and BIC(153.4963) consider it as ARMA(2,0). For ARMA(2,1) both AIC (148.0179) and BIC (153.754) consider it as ARMA(0,2).

Starting from T=200 sample AIC criterion shows a better result. It made only one mistake. For ARMA(1,2) AIC(563.9237) define this process as ARMA(2,0) as well as BIC(573.8186). However, BIC shows a worse result, because it made also a mistake for ARMA(2,1) and the BIC (576.6355) consider this process as ARMA(0,2).

On the large samples(T=500) AIC works better than BIC. Only BIC(1428.018) chose ARMA(2,0) instead of ARMA(1,2), while AIC chose all processes correctly and had a minimum value for the same order p,q that was simulated.

To conclude on the Maximum likelihood estimation, both AIC and BIC have the same mistakes and show the same result on small samples. When a large sample is used for the analysis, AIC criterion shows better results than BIC.

2.3.2 OLS estimation

Using the IML language, we will explain from the example of an ARMA (2.1) with 50 observations, what is the logic of our simulations.

Firstly, we initialize the Armasim function as seen above, adding the formulas of the information criteria as well as the vector of the information criteria.

```
proc iml;
/*MATRIX DEFINITIONS*/
GresuAIC=j(4,4,0);
GresuAICc=j(4,4,0);
GresuAICu=j(4,4,0);
GresuBIC=j(4,4,0);
GresuBICc=j(4,4,0);

/*Module ARMA */
start ARMA(y,p,q);

if (p+q)=0 then do;
sygma2=(y-y[:])`*(y-y[:])/nrow(Y);
AIC=log(sygma2);
BIC=log(sygma2);
BICc=BIC;
AICc=AIC;
AICu=log(sygma2);/*ON PREND TOUJOURS LA VARIANCE DES RESIDUS*/
resu=AIC//AICc//AICu//BIC//BICc;
end;
else do;
/*Information criteria*/
AIC=log(sygma2)+2*(p+q)/nrow(resid);
BIC=log(sygma2)+2*(p+q)*log(nrow(resid))/nrow(resid);
BICc=BIC+(2*(p+q)*log(nrow(resid)))/(nrow(resid)-p-q-1);
AICc=AIC+(2*(p+q)*(p+q+1))/(nrow(resid)-p-q-1);
AICu=log(sygma2)+(2*(p+q+1))/(nrow(resid)-p-q-2);/*We will take the variance of the residus */
resu=AIC//AICc//AICu//BIC//BICc;
return(resu) ;
Finish ARMA;
```

$$resu = \begin{bmatrix} AIC \\ AICc \\ AICu \\ BIC \\ BICc \end{bmatrix}$$

After initializing the ARMASIM function,by taking 1000 iterations for the Monte Carlo simulation, we assign the ARMASIM function to the following realization: ARMA (2,1) such as :
$Y_t = 0.5Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1}$

```
do rep=1 to 1000;
/*ROUTINE*/
/*ARMA (2,1) with T=50*/

phi={1 -0.5 0.25 }; /*AR*/
theta={1 0.5 }; /*MA*/
y=armasim(phi,theta, 0,1,50,0); /*  y=armasim(phi,theta,mean,standard error,T,seed)*/
y=y-y[:] ;/* mean adjusted*/
```

Then, we store the different AIC/BIC in p row and q columns For example, the matrix in [2,1] gives the AIC for an ARMA(2,1). And we export our results to temporary tables.
After that,we highlight through contour plots, the density of each of the information criteria, and therefore the most p and q represented orders.
Finally,when the criteria do not give the same information we use the information criteria vector in order to choose the p and q orders of the information criterion with the lowest value.

```
/*ARMA (2,1) with T=50*/
phi={1 -0.5 0.25 }; /*AR*/
theta={1 0.5 }; /*MA*/
y=armasim(phi,theta, 0,1,50,0);/*   y=armasim(phi,theta,mean,standard error,T,seed)*/
```

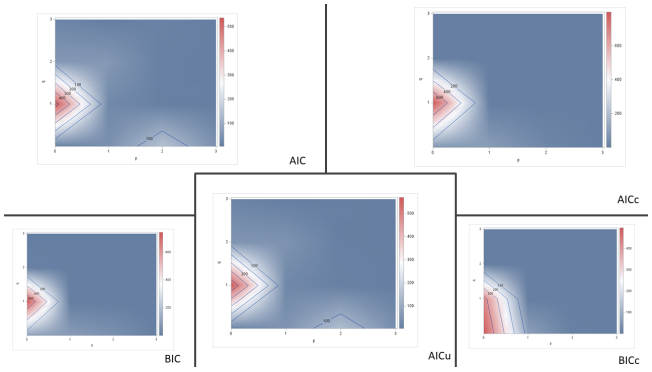
```
y=y-y[:] ;
Criteria21_50= ARMA(y,2,1) ; print Criteria21_50;
```

ARMA (0,1)

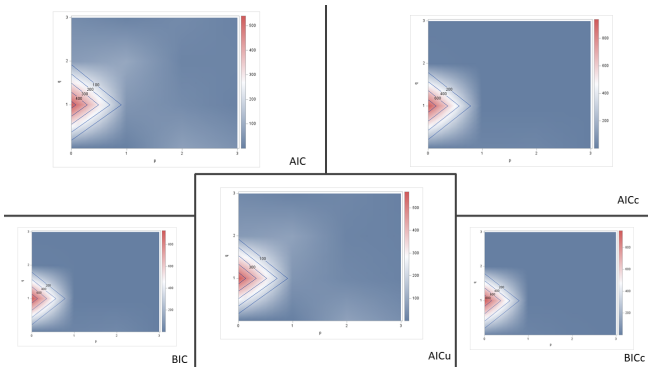
$$Y_t = \varepsilon_t + 0.75\varepsilon_{t-1}$$

According to the contour plots, all the information criteria choose p=0 and q=1 orders, and this, regardless of sample size.

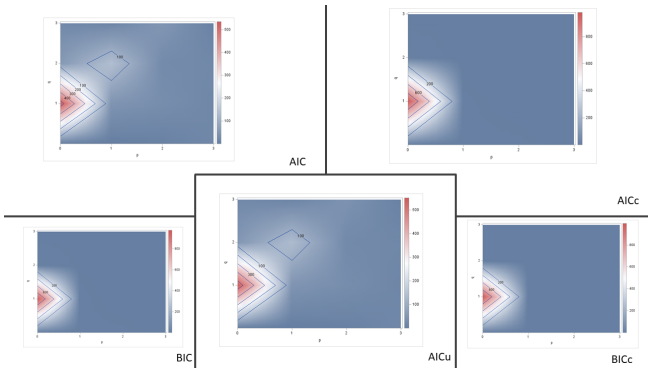
T=50



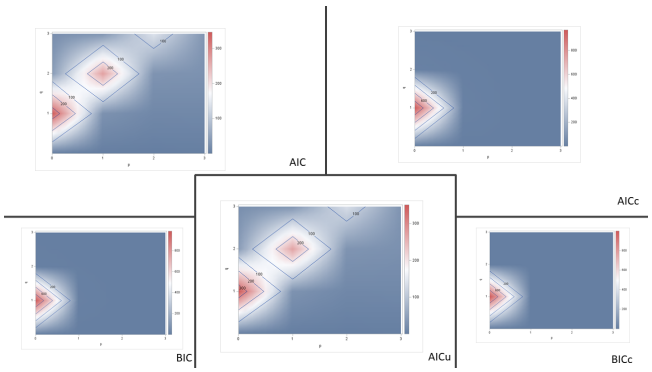
T=100



T=200



T=500



By minimizing the information criteria we will choose the p and q orders of AIC criterion and this, regardless of sample size.

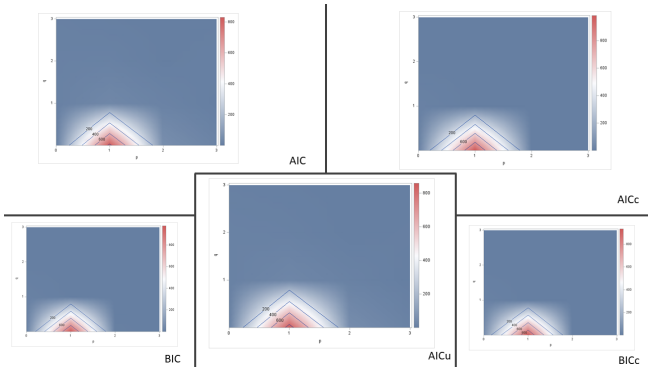
	Criteria01_50	Criteria01_100	Criteria01_150	Criteria01_200	Criteria01_500
AIC	0.1201027	-0.05526	0.2125132	0.0285269	-0.011931
AICc	0.2110118	-0.012707	0.240291	0.0491455	-0.003834
AICu	0.1696477	-0.033083	0.2267866	0.0390482	-0.00785
BIC	0.2430871	0.0189973	0.2670831	0.0721812	0.0090633
BICc	0.4171163	0.1161111	0.3362999	0.1265947	0.0341912

ARMA (1,0)

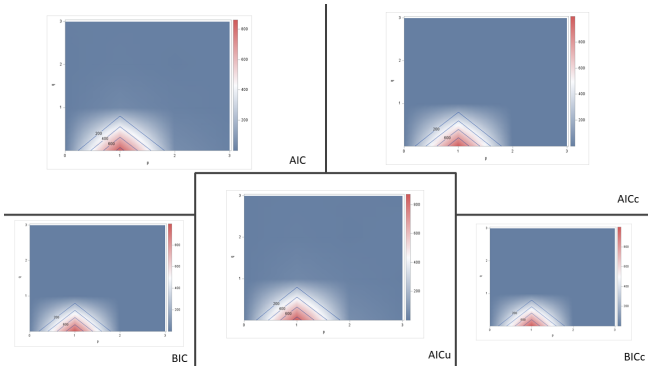
$$Y_t = 0.8Y_{t-1} + \varepsilon_t$$

According to the contour plots, all information criteria choose p=1 and q=0 orders, and this, regardless of sample size.

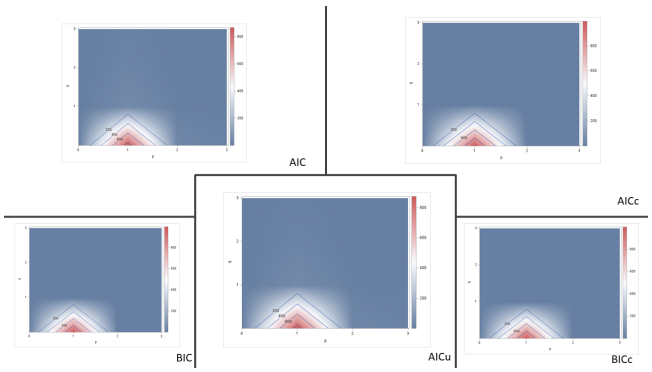
T=50



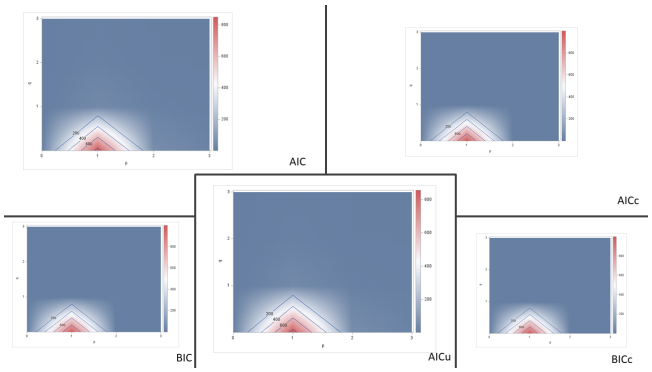
T=100



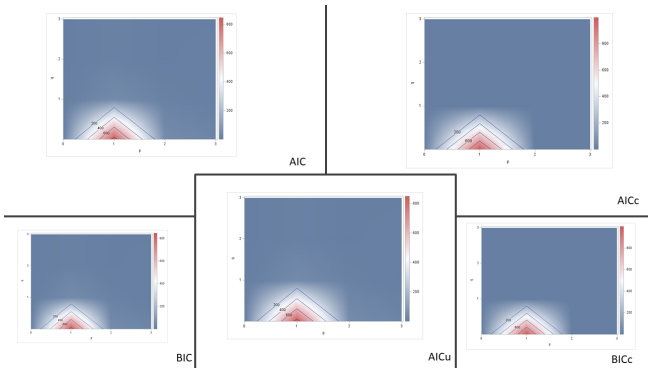
T=150



T=200



T=500



By minimizing the information criteria we will choose the p and q orders of AIC criterion and this, regardless of sample size.

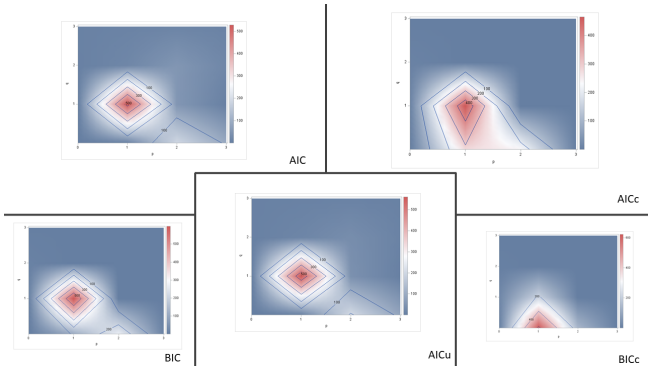
	Criteria10_50	Criteria10_100	Criteria10_150	Criteria10_200	Criteria10_500
AIC	0.1295297	-0.142171	0.0658412	-0.040259	0.3485036
AICc	0.2204388	-0.099618	0.0936189	-0.019641	0.3566008
AICu	0.1790747	-0.119994	0.0801146	-0.029738	0.3525849
BIC	0.2525142	-0.067914	0.1204111	0.0033948	0.3694979
BICc	0.4265433	0.0291996	0.1896279	0.0578083	0.3946257

ARMA (1,1)

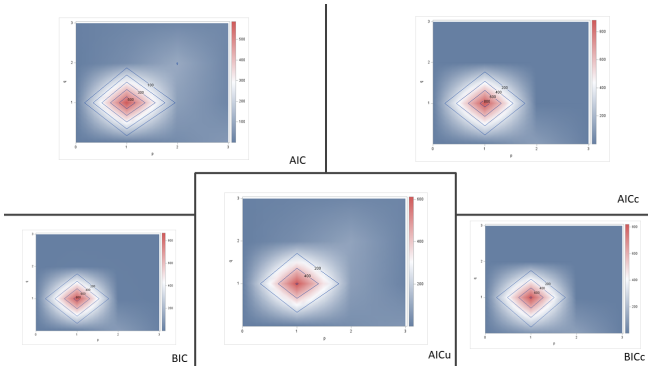
$$Y_t = 0.8Y_{t-1} + \varepsilon_t + 0.7\varepsilon_{t-1}$$

According to the contour plots, all the information criteria choose p=1 and q=1 orders, and this,for T=100, T=150, T=200 and T=500. Nevertheless, we observe that for T=50, BICc chooses p=1 and q=0 orders.

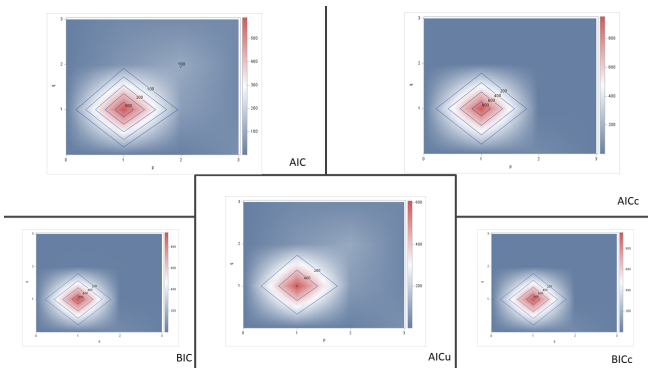
T=50



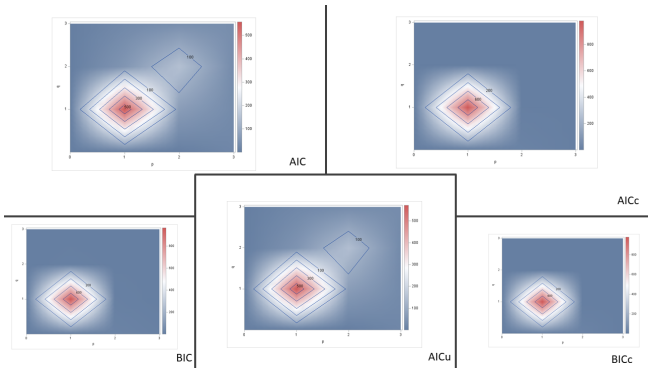
T=100



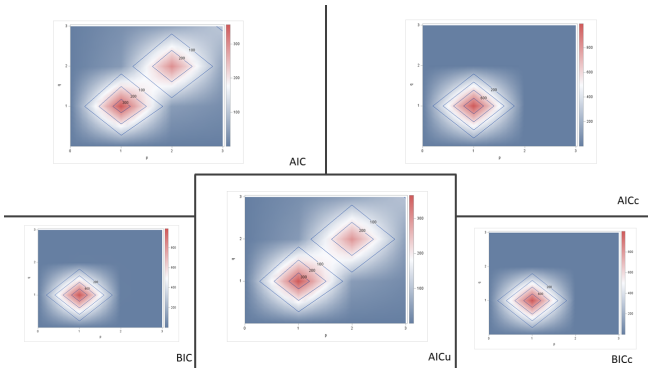
T=150



T=200



T=500



By minimizing the information criteria we will choose the p and q orders of AIC criterion and this, regardless of sample size.

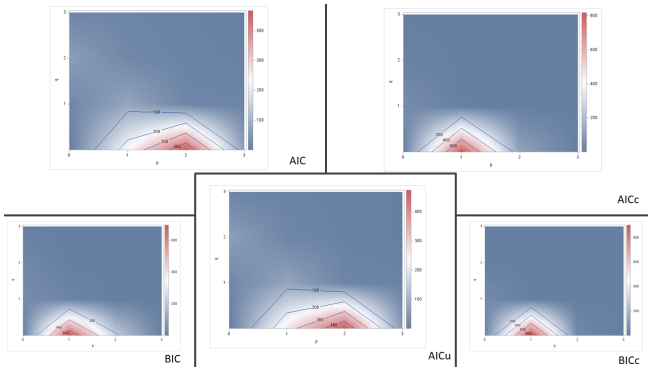
	Criteria11_50	Criteria11_100	Criteria11_150	Criteria11_200	Criteria11_500
AIC	0.1552407	0.0906593	-0.126383	0.1105592	0.0644431
AICc	0.4343105	0.2196915	-0.042467	0.2395915	0.0887839
AICu	0.2111414	0.11421	-0.111527	0.13411	0.0685737
BIC	0.4012096	0.2391738	-0.017243	0.2590737	0.1064316
BICc	0.7573623	0.4354898	0.1221582	0.4553898	0.1567892

ARMA (1,2)

$$Y_t = 0.5Y_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1} - 0.25\varepsilon_{t-2}$$

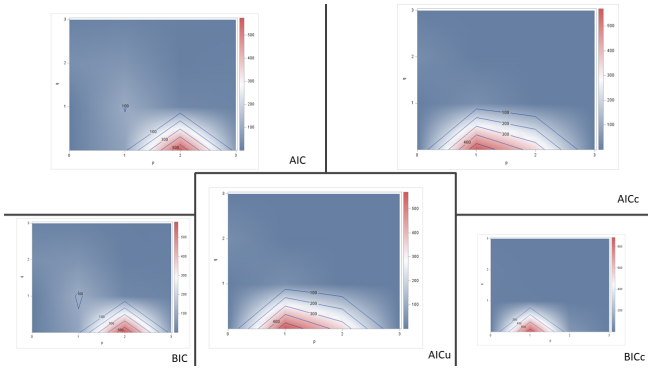
According to the contour plots,the choice of p and q orders differs for all the criteria.Indeed, for T=50, the AIC and AICu criteria choose p=2 and q = 0 orders while AICc, BIC and BICc criteria choose p=1 and q=0 orders.

T=50



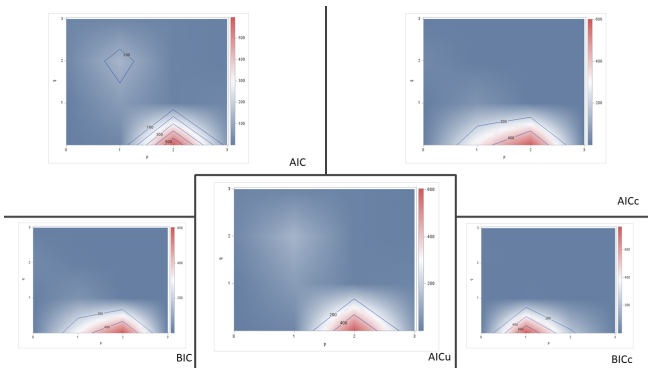
Then, for T=100, AICc, AICu and BICc criteria choose p=1 and q = 0 orders while AIC and BIC criteria choose p=2 and q = 0 orders.

T=100

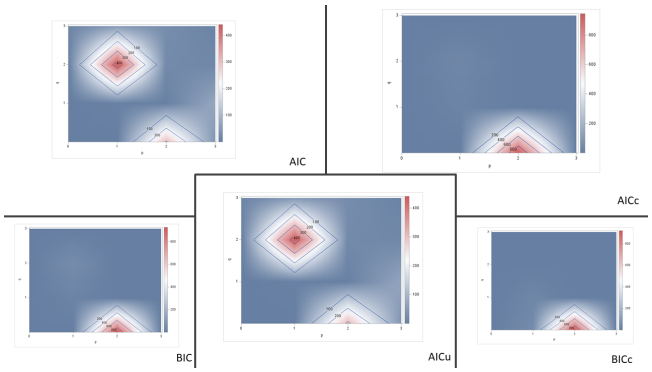


Moreover, for T=150 and T=200, all of them choose p=2 and q = 0 orders except to BICc criterion that choose p=1 and q=0 orders.

T=150

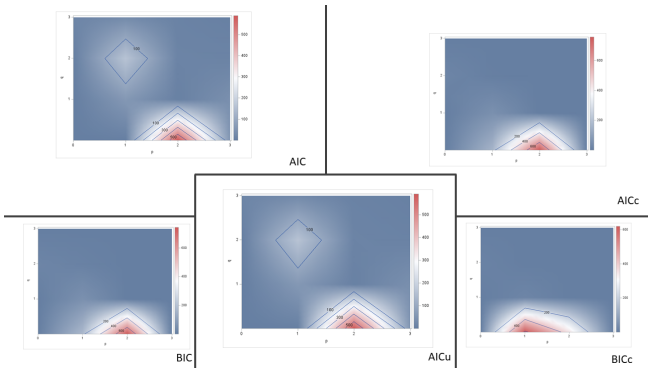


T=200



Finally, for T= 500, AIC and AICu criteria choose p=1 and q = 2 orders while AICc, BIC and BICc criteria choose p=2 and q=0 orders.

T=500



By minimizing the information criteria we will choose the p and q orders of AIC criterion and this, regardless of sample size.

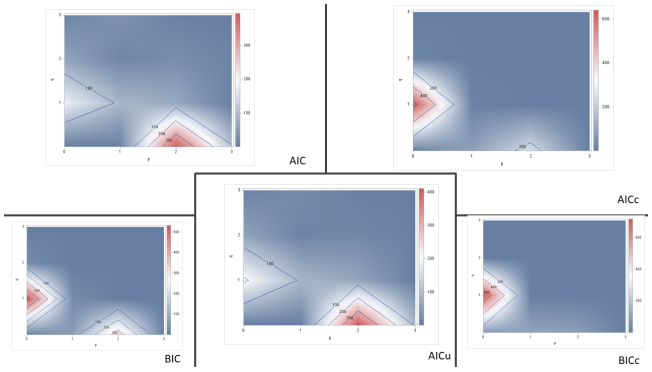
	Criteria12_50	Criteria12_100	Criteria12_150	Criteria12_200	Criteria12_500
AIC	0.0371364	-0.065296	-0.095265	0.158527	-0.014639
AICc	0.608565	0.1955731	0.0737494	0.283527	0.0341413
AICu	0.1018236	-0.039884	-0.079623	0.1697996	-0.010443
BIC	0.4060896	0.1574753	0.0684452	0.2894897	0.0483436
BICc	0.9530384	0.4551502	0.2790201	0.4544308	0.1240335

ARMA (2,1)

$$Y_t = 0.5Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1}$$

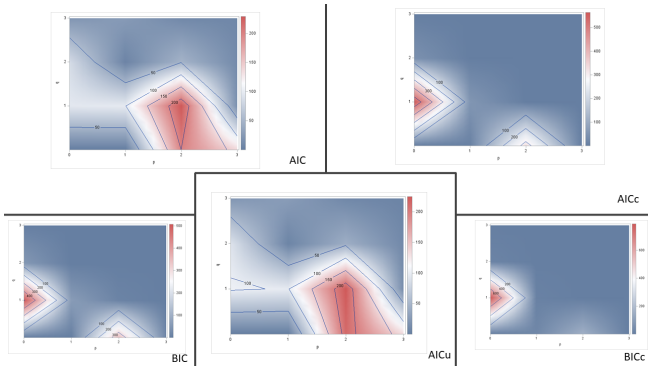
According to the contour plots,the choice of p and q orders differs for all the criteria.Indeed, for T=50, AIC and AICu criteria choose p=2 and q = 0 orders while AICc, BIC and BICc criteria choose p=1 and q=0 orders.

T=50



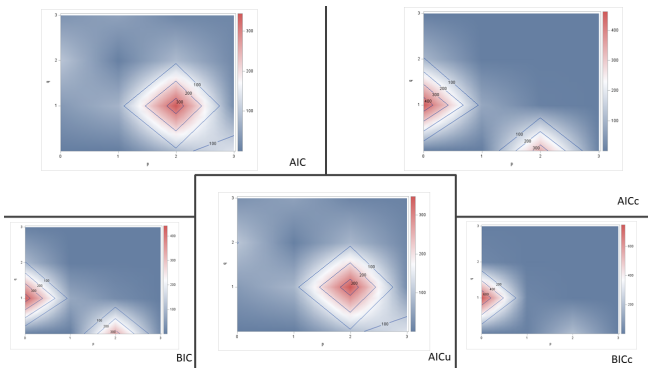
Then, for T=100, AICc, BIC and BICc criteria choose p=0 and q = 1 orders while AIC and AICu criteria choose p=2 and q = 1 orders.

T=100



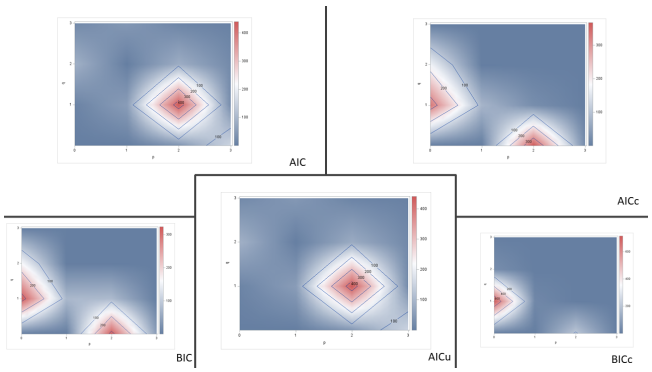
Moreover, for T=150 AIC and AICu criteria choose p=2 and q =1 orders while AICc, BIC and BICc criteria that choose p=0 and q=1 orders .

T=150



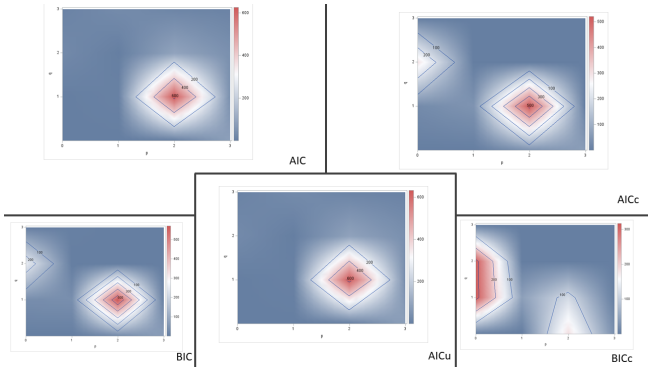
For T=200 AIC and AICu criteria choose p=2 and q =1 orders while the AICc chooses p=2 and q=0 orders and that BIC and BICc criteria choose p=0 and q=1 orders .

T=200



Finally, for T= 500, all of them choose p=2 and q = 1 orders except to BICc criterion chooses p=0 and q=2 orders .

T=500



By minimizing the information criteria we will choose the p and q orders of AIC criterion and this, regardless of sample size.

	Criteria21_50	Criteria21_100	Criteria21_150	Criteria21_200	Criteria21_500
AIC	0.1776189	-0.017958	0.153102	-0.028203	-0.005073
AICc	0.7490475	0.2429114	0.3221161	0.0967973	0.0437078
AICu	0.2423061	0.0074539	0.1687437	-0.01693	-0.000876
BIC	0.5465721	0.2048136	0.3168118	0.10276	0.0579101
BICc	1.0935209	0.5024884	0.5273868	0.2677011	0.1336001

Comparison between these methods

Firstly,the Maximum Likelihood estimation as we did by ARIMA procedure does not allow us to calculate some information criteria as AICc, AICu and BICc while this is possible with SAS/IML.
Moreover,the ACF and PACF are better at finding Pure AR or Pure MA models but as far as ARMA processes are concerned it is better to use the information criteria.
Finally, the best information criterion seems to be the Akaike's information criterion.

Chapter 3

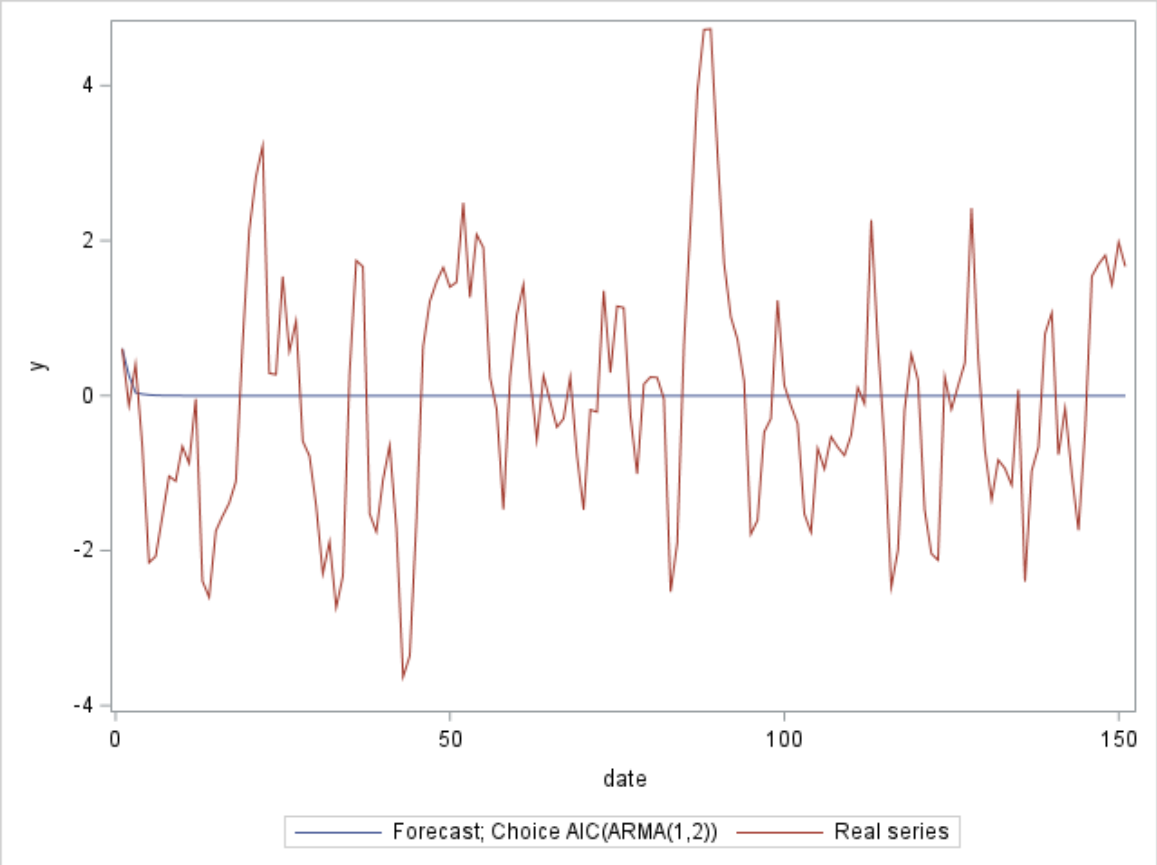
Series Forecast

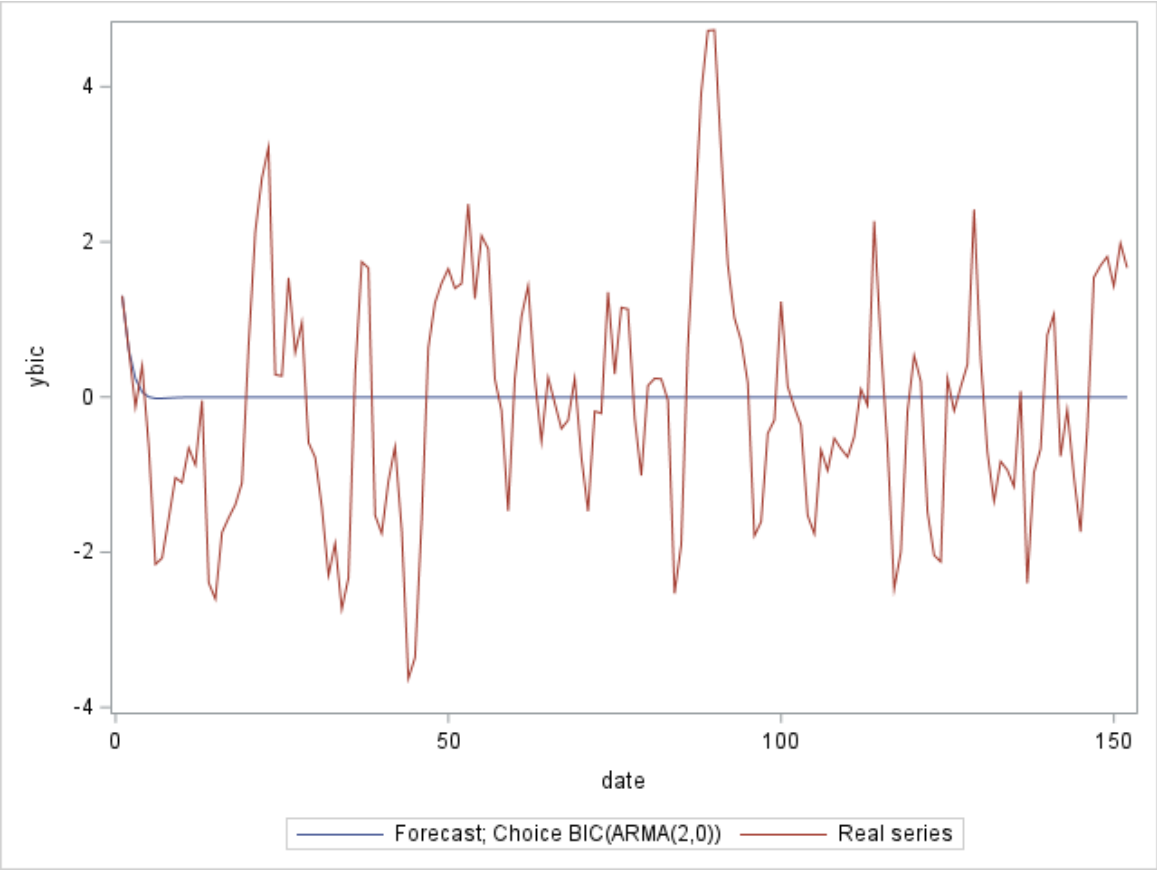
3.1 Forecasting simulated series

3.1.1 Forecast ARMA(1,2)

Forecast for ARMA(1,2) simulated process as ARMA(1,2) chosen by AIC criterion and ARMA(2,0) chosen by BIC criterion on the sample with 500 observations.

For this process we have made a forecast of 150 observations(300-450).





On the graphics we can observe that both forecasts are too similar and it is difficult to determine which forecast is better and which criterion determines better. We also observe that the forecast line converge to 0 fast as the process is stationary.Thus, there is an interest to forecast only a few periods.

For better comparison we have calculated a RMSE . The results for ARMA(1,2) for RMSE are following:

	RMSE
ARMA(1,2)	1,5203089
ARMA(2,0)	1,5145965

Observing the table, we can result that RMSE is less for ARMA(2,0) than for ARMA(1,2). Even though they are almost equal, the better forecast is the forecast for ARMA(2,0) though we have simulated ARMA(1,2) originally.

We can conclude that AIC choose correctly the series, but BIC is better for the forecast.

This result proves that AIC performes better than BIC.

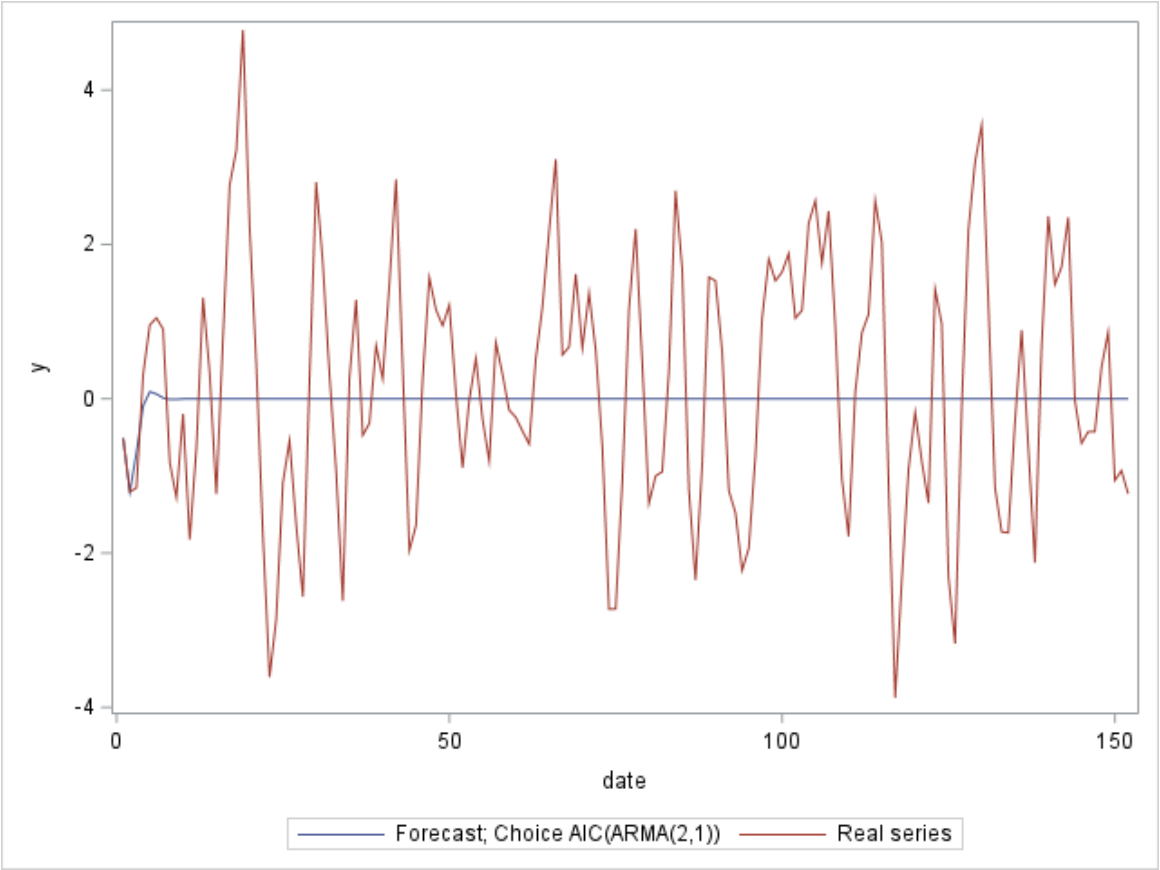
3.1.2 Forecast ARMA(2,1)

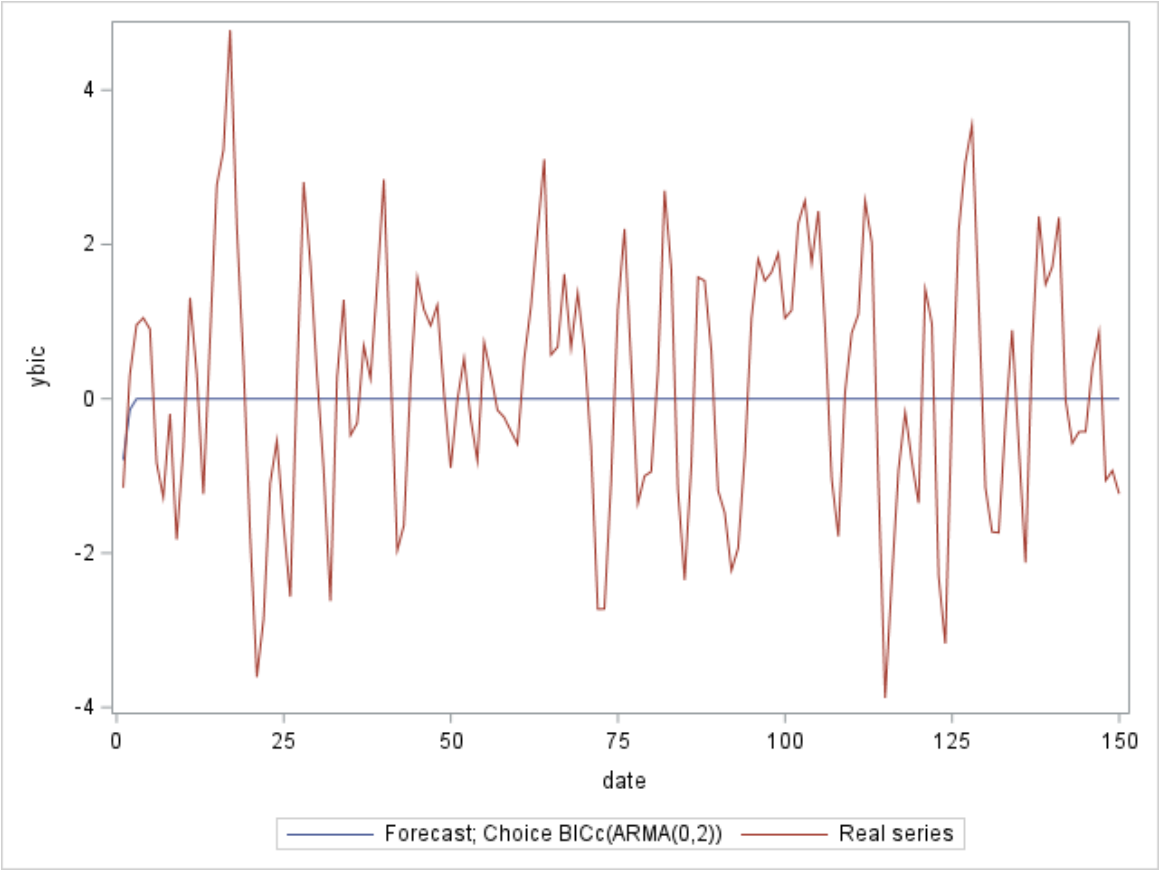
Forecast for ARMA(2,1) simulated process as ARMA(2,1) chosen by AIC criterion and ARMA(0,2) chosen by BICc criterion on the sample with 500 observations.

For this process we have made a forecast of 150 observations(300-450).

```
/* Forecast for an ARMA(2,1) with 100 observations */

Y = j(152,1,0);
Y[2,] = y2;
Y[3,]=y33[1,] * Y[2,]+y33[2,] * Y[1,]+y33[3,] * residu[296,];
print Y;
do i=4 to 152;
    Y[i,]=y33[1,] * Y[i-1,]+y33[2,] * Y[i-2,];
end;
```

```
print Y;

/* Forecast for an ARMA(0,2) with 100 observations */

Ybicc = j(150,1,0);
Ybicc[1,]=y33[2,]residu[295,]+y33[1,] * residu[296,];
Ybicc[2,] = y33[2,] * residu[296,];
print Ybicc;
```





On the graphics we can observe that both forecasts are too similar and it is difficult to determine which forecast is better and which criterion determines better. We also observe that the forecast line converge to 0 fast as the process is stationary.Thus, there is an interest to forecast only a few periods.

For better comparison we have calculated a RMSE and Theil Statistic. The results for ARMA(2,1) for RMSE and for Theil Statistic are following:

	RMSE
ARMA(2,1)	1,5953415
ARMA(0,2)	1,6065701

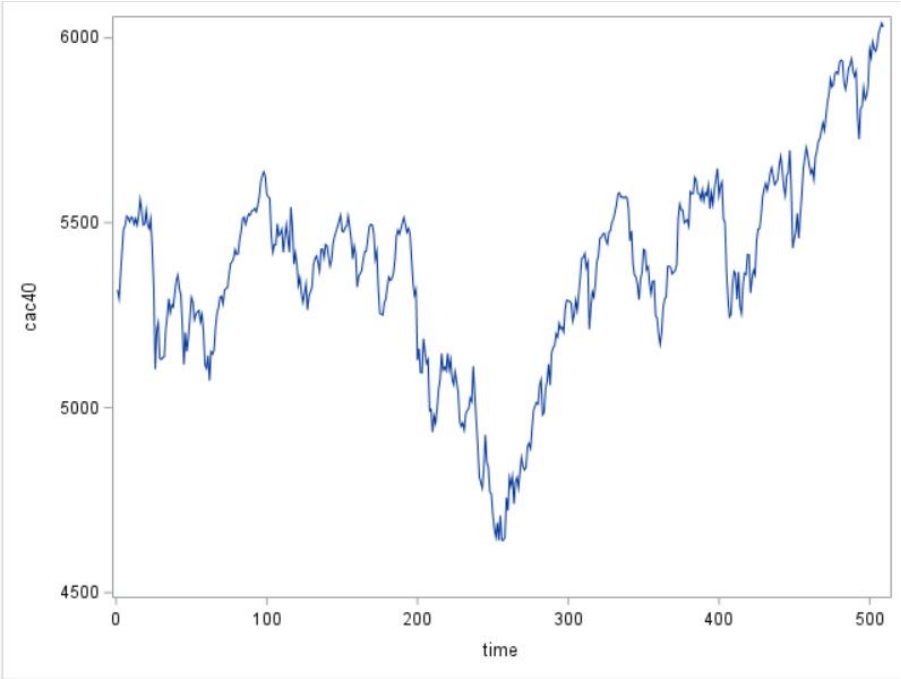
Observing the table, we can result that RMSE is less for ARMA(2,1) than for ARMA(0,2) . Even though they are almost equal, the better forecast is the forecast for ARMA (2,1).

Knowing that the original series is an ARMA(2,1), RMSE proves that as well as AIC criterion. BICc criterion made a mistake and chose a wrong model and forecast is worse in this case.

This result proves that AIC performs better than BICc.

3.2 Forecasting true series

In this section, we are going to use a real series and apply the techniques describes above. We are going to observe if information criteria allow to make a coherent forecast.
The base used for the forecast represent CAC 40 Index at the opening 1st January 2018 - 9 April 2020. Initially we are going to use the information in period 1st January 2018-1st January 2020.


6000
5500
5000
4500
cac40
0
100
200
300
400
500
time

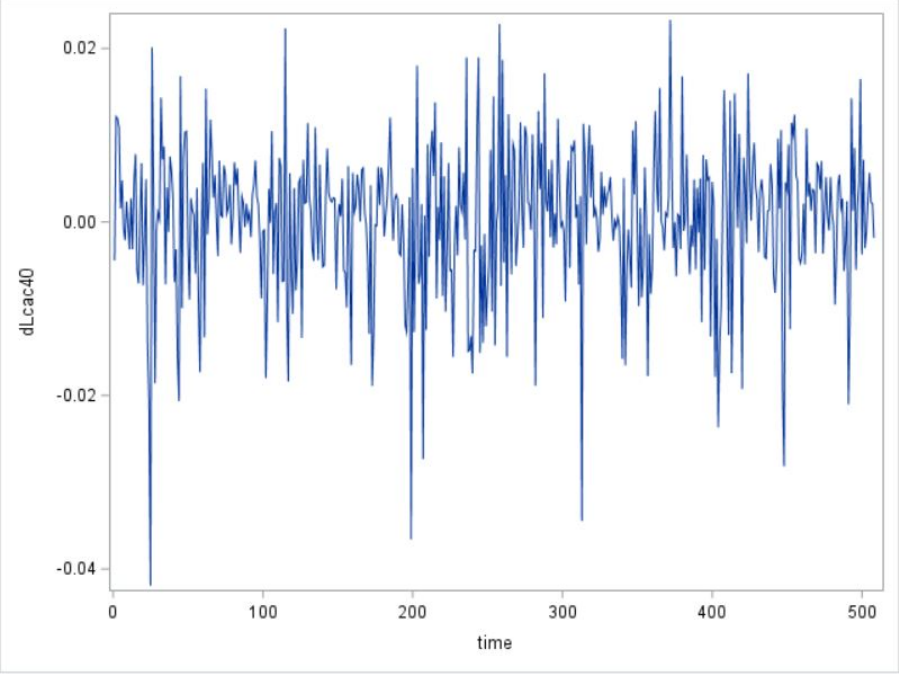
3.2.1 Order of Integration

On the plot we can observe a non stationarity of the time series.

We can check that using tests:

Tests de racine unitaire de Dickey-Fuller augmentés							
Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Moyenne zéro	1	0.1253	0.7121	0.66	0.8572		
	2	0.1113	0.7088	0.56	0.8360		
	3	0.0948	0.7049	0.45	0.8100		
	4	0.0854	0.7026	0.42	0.8035		
	5	0.0902	0.7038	0.48	0.8189		
	6	0.0861	0.7028	0.46	0.8127		
Moyenne simple	1	-4.3296	0.5024	-1.13	0.7044	0.90	0.8418
	2	-5.0404	0.4304	-1.24	0.6570	0.97	0.8237
	3	-6.2386	0.3276	-1.42	0.5717	1.15	0.7770
	4	-5.6394	0.3761	-1.33	0.6180	1.00	0.8147
	5	-4.3266	0.5027	-1.11	0.7152	0.76	0.8780
	6	-4.5521	0.4790	-1.14	0.7022	0.78	0.8718
Tendance	1	-6.6500	0.6908	-1.62	0.7859	1.78	0.8212
	2	-7.6689	0.6083	-1.76	0.7229	2.09	0.7577
	3	-9.2946	0.4836	-1.98	0.6138	2.55	0.6643
	4	-8.7824	0.5215	-1.93	0.6389	2.63	0.6496
	5	-7.2158	0.6448	-1.72	0.7423	2.35	0.7059
	6	-7.5533	0.6175	-1.76	0.7214	2.47	0.6821

For theoretical reasons and to facilitate our task, we will use the first difference of the logarithm of Cac 40: $\Delta \ln(Y_t)$ où Y_t =Cac40$_t$



Before starting to use our series, we are going to check that it is I(0) and the residuals of the model are white. The table above shows that $\ln(Y_t)$ is I(1)

Tests de racine unitaire de Dickey-Fuller augmentés							
Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Moyenne zéro	1	-454.719	0.0001	-15.07	<.0001		
	2	-381.531	0.0001	-11.79	<.0001		
	3	-432.709	0.0001	-10.85	<.0001		
	4	-705.688	0.0001	-10.69	<.0001		
	5	-671.971	0.0001	-9.57	<.0001		
	6	-815.631	0.0001	-8.96	<.0001		
Moyenne simple	1	-455.770	0.0001	-15.07	<.0001	113.55	0.0010
	2	-382.870	0.0001	-11.79	<.0001	69.54	0.0010
	3	-434.888	0.0001	-10.85	<.0001	58.87	0.0010
	4	-712.882	0.0001	-10.69	<.0001	57.16	0.0010
	5	-680.949	0.0001	-9.58	<.0001	45.88	0.0010
	6	-833.035	0.0001	-8.97	<.0001	40.23	0.0010
Tendance	1	-458.406	0.0001	-15.10	<.0001	114.03	0.0010
	2	-387.284	0.0001	-11.84	<.0001	70.10	0.0010
	3	-443.929	0.0001	-10.92	<.0001	59.61	0.0010
	4	-743.234	0.0001	-10.77	<.0001	58.02	0.0010
	5	-725.124	0.0001	-9.67	<.0001	46.77	0.0010
	6	-926.774	0.0001	-9.07	<.0001	41.15	0.0010

Vérification de l'autocorrélation pour le bruit blanc									
Jusqu'au retard	Khi-2	DDL	Pr > khi-2	Autocorrélations					
6	7.65	6	0.2647	-0.000	0.051	0.061	-0.032	-0.086	0.008
12	10.24	12	0.5951	-0.035	0.022	-0.021	0.018	-0.017	0.047
18	16.30	18	0.5718	-0.057	-0.015	0.033	-0.006	0.038	0.074
24	21.69	24	0.5979	-0.044	-0.044	-0.004	-0.070	-0.026	-0.026

We can conclude that both hypothesis are respected.

3.2.2 DGP Identification

We apply the Arma and Armasim methods by integrating different processes on the series: AR(1),MA(1),ARMA(1,1),ARMA(2,1),ARMA(1,2)

```
resid=y[a+1:nrow(X4)]-X4[a+1:nrow(X4),]*beta;
sigma2=(t(resid)*resid)/(nrow(resid)-ncol(X4));
AIC=log(sigma2)+2*(p+q)/nrow(resid);
BIC=log(sigma2)+2*(p+q)*log(nrow(resid))/nrow(resid);
AICc=AIC+(2*(p+q)*(p+q+1))/(nrow(resid)-p-q-1);
BICc=BIC+(2*(p+q)*log(nrow(resid)))/(nrow(resid)-p-q-1);
AICu=log(sigma2)+(2*(p+q+1))/(nrow(resid)-p-q-2);
Critere = AIC // BIC // AICc // BICc // AICu ;
return(Critere);
Finish Arma;
Critere10 = ARMA(Diff,1,0);
Critere01 = ARMA(Diff,0,1);
Critere11 = ARMA(Diff,1,1);
Critere21 = ARMA(Diff,2,1);
Critere12 = ARMA(Diff,1,2);
Critere = Critere10 || Critere01 || Critere11 || Critere21 || Critere12 ;
print Critere;
Critere_AIC = Critere[1,];
Critere_BIC = Critere[2,];
Critere_AICc = Critere[3,];
Critere_BICc = Critere[4,];
Critere_AICu = Critere[5,];
Min_AIC=Min(Critere_AIC);
Min_BIC=Min(Critere_BIC);
Min_AICc=Min(Critere_AICc);
Min_BICc=Min(Critere_BICc);
Min_AICu=Min(Critere_AICu);
print Min_AIC Min_BIC Min_AICc Min_BICc Min_AICu;
```

The information criteria give:

Critere				
-9.522362	-9.522363	-9.52035	-9.516909	-9.517084
-9.501704	-9.501705	-9.479033	-9.454935	-9.455109
-9.514426	-9.514427	-9.496493	-9.469101	-9.469275
-9.476996	-9.476997	-9.429518	-9.380514	-9.380689
-9.518363	-9.518364	-9.516303	-9.512799	-9.512974

Min_AIC	Min_BIC	Min_AICc	Min_BICc	Min_AICu
-9.522363	-9.501705	-9.514427	-9.476997	-9.518364

The criteria choose unanimously the 2nd column which corresponds to an MA (1).

$$\Delta \ln(Y_t) = \tilde{Y}_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

We estimate parameters and we check the significance of delay of the residue.

```
beta=inv(t(X4[a+1:nrow(X4),]) * X4[a+1:nrow(X4),])*t(X4[a+1:nrow(X4),]) *y[a+1:nrow(X4),];
resid=y[a+1:nrow(X4)]-X4[a+1:nrow(X4),]*beta;
sigma2=(t(resid)*resid)/(nrow(resid)-ncol(X4));
Varcovar=vecdiag(sigma2* (inv(t(X4[a+1:nrow(X4),])*X4[a+1:nrow(X4),])));
dfe=nrow(X4)-ncol(4);
tstat=beta/sqrt(varcovar);
pvalue=(1-probT(abs(tstat),dfe))*2;
beta_pvalue = beta // pvalue;
return(beta_pvalue);
Finish Arma;
```

signif
-0.001781
0.9680809

We pbserve that the parameter isn't significant and we can say that our explanatory variable is similar to white noise.

3.2.3 Forecast and approximation

Knowing that we have a white noise, the forecasts are always equal to zero because the series varies around 0. If we calculate the Theil U and the RMSE for our 500 observations, we get this:

```
MSE = DIFF[##,]/nrow(Diff);
RMSE = sqrt(MSE);

M= j(nrow(Diff)-1,1,0);
do i=1 to nrow(M);
        M[i] = Diff[i+1]-Diff[i];
end;
U = sqrt(MSE *nrow(Diff)/M[##,]);
print RMSE U;
```

RMSE	U
0.0085322	0.7072096

The Theil Statistic is lower than 1 and it means our forecast is better than naive predictions.

Regarding approximation, it would be interesting to see if it is possible to approximate $\Delta \ln(Y_t)$ by the return rate.

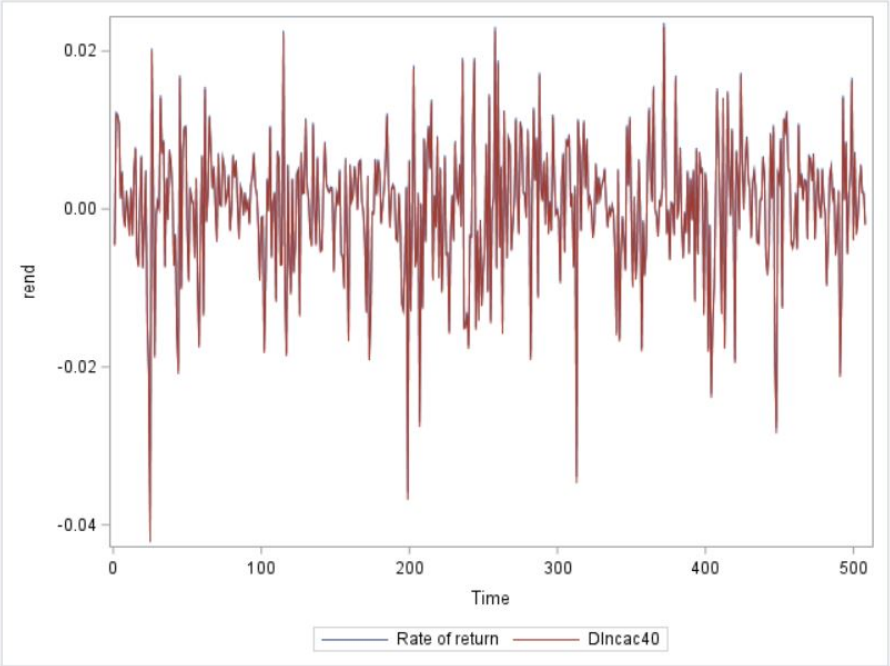
```
rend = j(nrow(DIFF),1,0);
do i=1 to nrow(Diff);
        rend[i] = (serie[i+1]-serie[i])/serie[i];
end;
Max_rend = Max(abs(rend));
print Max_rend;
```

Max_rend
0.0410788

We see that the maximum return, in absolute value, is very small comparing to 1 so the approximation is relevant.

$$\Delta \ln(Y_t) \approx \frac{Y_t - Y_{t-1}}{Y_{t-1}} \tag{3.1}$$

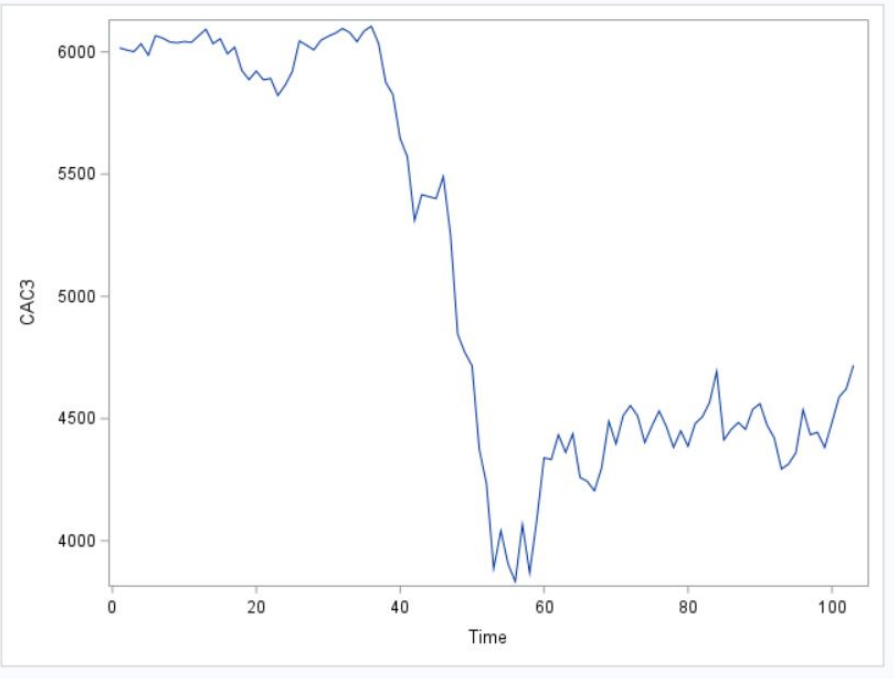
We can also see it graphically.



The values are so close that it is not possible to distinguish the 2 series.

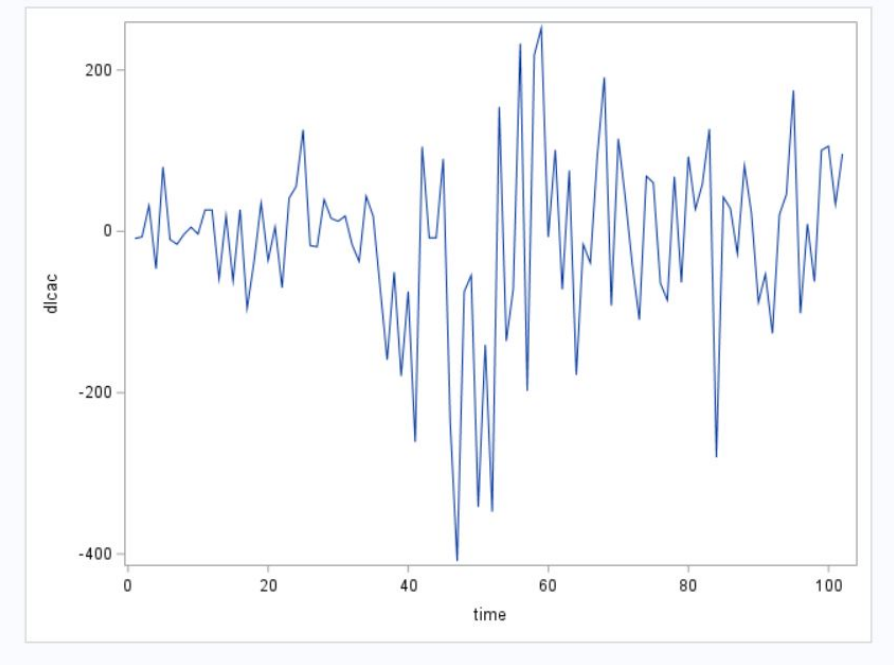
After having shown that the DGP of the return rate is white noise, the aim is to see if it is still the case during the Coronavirus crisis. The idea is to say that, if the forecasts are terrible, there is a structural shock and this would demonstrate that the information criteria are not all-powerful indicators. The criteria should be used when it is certain that there is no structural break. To spot these breaks,we have to make some tests and we don't have to forget about economic theory and about the reality that can point to structural breaks.There is indeed a lot of chance, without the need for any test, that the Covid has created a structural break on the initial series which means that this series no longer has the same properties as before. But is that the case for return rate?

Graphically, the (initial) series during the Covid looks like this:



We see that the series does not resemble at all what it was before. There is a high chance to have a structural break.

Regarding the return rate, we observe that graphically it looks like this:



We observe that the series does not seem stationary so we can think that there is a structural break during the Covid. However it strongly resembles a cluster of heteroscedasticity which was also perceptible in the initial series. It seems indeed that a more adequate DGP would be a SARIMA rather than an ARIMA but we omit this type of modeling here.

We are going to look at what Theil's U and RMSE give us, to see if a null forecast remains relevant in this case.

RMSE2	U2
114.60667	0.7229641

We see that even in this case, making a null forecast is better than a static forecast, however the RMSE is quite high.This may be due to the fact that we did not use more adequate modeling and not because of a wrong choice of information criteria. Despite all this, the criteria seem quite relevant because the return rate remains stationary on average (around 0).

Conclusion

Information criteria are relevant tools insofar as their expression allows us to choose statistical models. Being asymptotically equivalent, they return the same information on large samples and the problem does not arise on what criteria to take to choose a model. We observe that the criteria find the right DGP or arrive at a DGP which is close with a model which gives estimates close to the initial ones. The problem arose concerning small samples insofar as groups of criteria were imposed. Indeed, it was observed that there were sometimes 2 groups of criteria which did not choose the same model. We can see that in this case, the AIC and the BIC do not return the same information and that in general the AIC more often finds the right DGP than the BIC. In this sense, the AIC remains more reliable than the BIC. The other problem that arose was the fact that the BIC, which did not choose the right DGP, gave less forecast errors than the AIC when using RMSE or Theil's U. The other drawback of the information criteria is that the model chosen depends on the models that one wishes to test. To find the right DGP, you need to know the context of the series and have a good knowledge of it. This is why it is important to submit several models to be able to be sure of having the right DGP or being close to it.

Bibliography

[1] Allan D R McQuarrie, Chih-Ling Tsai *Regression and Time Series Model Selection* North Dakota State University, University of California, Davis,1998.

[2] Catherine Bruneau *Cours Econométrie financière* Paris 1 Panthéon-Sorbonne,2020.

[3] Didier Delignières *Séries temporelles – Modèles ARIMA.* Séminaire EA "Sport – Performance – Santé",Mars 2000.

[4] GESINE REINERT *Lecture Time Series* Oxford University,2010.

[5] Gilbert Colletaz *Présentation et utilisation des critères de sélection* 12 septembre 2019.

[6] Philippe De Peretti *Cours Économétrie appliquée(SAS)* Paris 1 Panthéon-Sorbonne,2020.

[7] SAS *SAS/ETS®14.1 User's Guide The ARIMA Procedure* 2015, SAS Institute Inc., Cary, NC, USA.

[8] Arthur Charpentier *Cours de séries temporelles, théorie et applications* ENSAE,Université Paris-Dauphine