

B3 - Time Series Analysis - Fall 2017

Home Work Assignment 1

Lars Forsberg
Uppsala University, Department of Statistics

November 1, 2017

Abstract

The task is twofold: 1) to derive the theoretical Autocorrelation Function (ACF) and 2) simulate the sample ACF (SACF/Correlogram) for a number of stochastic processes.

This exercise will yield a "catalog" of correlograms for different processes with different parameter values, to be used later (in HWA2, and in general) as an identification tool to find out what process could have generated/can approximate the Data Generating Process (DGP) behind real data.

1 General Instructions

For general instructions about the HWA, such as group formation, deadlines etc, see the section Home Work Assignments in the document 'B3 - Time Series - Fall 2017 - Schedule and General Information.pdf.'

2 Introduction

Each stochastic process or model implies a theoretical structure of the autocorrelation function (ACF). This ACF will in a very compact way describe the autocorrelation structure of the process. A "sample version" of this structure can then be seen in the realization from the process (actual data generated/simulated from the process).

A possible analogue here would be that the theoretical ACF is actually the *finger* of the process and the SACF is its "fingerprint". (So, if our process was the criminal at the crime scene, you being Sherlock Holmes, the SACF is the fingerprints (evidence) at that crime scene which then would help you identify the criminal.¹) Note that we will use the word correlogram to mean the plot of ACF versus lag length k *both* for the theoretical ACF and the sample ACF/sample PACF.

So, to identify what process that might have generated the data, or at least what model that can provide a good approximation of the data generating process, we *match the two*, the theoretical ACF and the Sample ACF (correlogram). To make the "match" we need to have a catalogue of ACF so that we can compare the correlogram of the data to possible "suspects". The purpose of this HWA is to create a light version of such a catalog for some stationary(?) ARMA(p,q) processes.

3 Theoretical part

The first task is to derive some statistical properties of a set of models, the models are

1. MA(1)
$$Y_t = e_t - \theta e_{t-1}. \quad (1)$$

2. MA(2)
$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}. \quad (2)$$

3. AR(1)
$$Y_t = \phi Y_{t-1} + e_t \quad (3)$$

4. AR(2)
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \quad (4)$$

5. ARMA(1,1)
$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \quad (5)$$

¹As a side note for the interested student: Scotland Yard was the first Police organisation in the world to systematically use and collect fingerprints of criminals and on crime scenes.

Nowadays genetics have taken much of its place in crime scene investigations. It should be noted however, that the "genetic evidence production" relies heavily on statistical methods.

Note that, for ease of notation and with no loss of generality, we do not have any constants in these models. For all the processes above, we have

$$e_t \sim NID(0, 1). \quad (6)$$

For each of the models, derive and report the following theoretical properties:

1. The Mean function $\mu_t = E(Y_t)$
2. The Variance function $\gamma_0 = V(Y_t)$
3. The First autocovariance $\gamma_1 = Cov(Y_t, Y_{t-1})$
4. The Second autocovariance $\gamma_2 = Cov(Y_t, Y_{t-2})$
5. The First Autocorrelation $\rho_1 = \frac{\gamma_1}{\gamma_0}$
6. The Second Autocorrelation $\rho_2 = \frac{\gamma_2}{\gamma_0}$
7. A general expression for the the Autocorrelation function as a function of the parameters of the process $\rho_k, k \geq 1$

Thus, you need to, sort of, check all the entries in the table

	Statistical property						
Model	μ_t	γ_0	γ_1	γ_2	ρ_1	ρ_2	ρ_k
MA(1)							
MA(2)							
AR(1)							
AR(2)							
ARMA(1,1)							

(Note already: All the *derivations* **must be in an appendix.**)

4 Simulation part

Given a stochastic process, exactly how the autocorrelation function (correlogram) looks (pattern, structure) depends on the actual *parameter values* of the process. It is be possible to plot the (theoretical) ACF vs k for each set of parameters for each process, and this way you would obtain the exact structure of the correlogram.

However, to get a feeling of how the different processes actually look, their *time series plots*, and the matching with the different *correlograms*, you should **simulate/generate realizations the processes** (in R) and (let R) calculate the correlogram of the simulated data. Using $n = 5000$ will ensure that the SACF looks pretty much like the theoretical.

Simulate in R. On studenportalen, in the HWA1 folder there are three programs,

1. MA_Simulate_Plot.R
2. AR_Simulate_Plot.R
3. ARMA_Simulate_Plot.R

These programs are designed to generate realizations of the different processes.

For all processes the parameter values are set to **zero** as default. Thus you need to change them to generate the different processes. (With parameter values zero, which process do you get?)

4.1 How to simulate a realization

To simulate a realization of a process, run the corresponding program. For example, for the MA(1) model.

1. MA_Simulate_Plot.R
2. Make sure the *parameters* are set to the correct value(s).
3. Set the random number generator seed to the birthdate of one of the students in the group e.g. `set.seed(960231)`.
4. Simulate $n = 5\,000$ observations (make sure NumObsSim is set to the correct value)
5. Plot the first, say 500 observations (set numObsToPlot=500)
6. See below on how to report these two graphs.

For your learning: try other sample sizes (no need to report this) such as $n = 400$ or even smaller sample, to see how this affects the correlogram, the intuition would be that the smaller the sample size the "noisier" the correlogram would be, that is, the pattern will not be as clear. In practice it is reasonable that you would have samples around $n = 100$ to 400, depending on the application, so it is instructive to "play around" with these smaller samples". When using $n = 5000$ we would expect the that autocorrelations for the different lags converge to, or at least is very close to the true values, thus getting a "cleaner" picture of how the autocorrelation function actually looks for this process and these parameter value(s).

4.2 Parameter values in the simulations

Below are the parameter values or combinations of parameter values you should use when simulating the realizations of the different stochastic processes. These parameter values you need to change in the program code for each set of parameters.

For the MA(1) process, simulate data using the following parameter values:

		θ			
-1.0	-0.50	0.0	0.50	1.0	2

(7)

For the MA(2) process, simulate data using the following parameter combinations:

		$\theta_1 =$		
		-0.4	0.0	0.4
θ_2	0.0			
	0.7			
	1.0			

(8)

of course, this means that in each cell, we have the combination (θ_1, θ_2) so that

		$\theta_1 =$		
		-0.4	0.0	0.4
θ_2	0.0	$\theta_1 = -0.4, \theta_2 = 0.0$	$\theta_1 = 0.0, \theta_2 = 0.0$	etc
	0.7	etc		
	1.0			

For the AR(1) process, simulate series with the following parameter values:

		$\phi_1 =$				
-1.00	-0.90	-0.75	0	0.75	0.90	1.00

(9)

For the AR(2) process simulate, data using the following combinations of parameter values:

		$\phi_1 =$		
		-0.9	0.0	0.8
$\phi_2 =$	0.1			
	0.2			
	0.8			

(10)

For the ARMA(1,1) process simulate, data using the following combinations of parameter values:

		$\theta =$	
		-0.4	0.4
$\phi =$	-0.90		
	0.80		
	0.90		

(11)

5 How to present the results/graphs

Report the result **model by model**. That is, first the results for the MA(1) theory and comments on the simulation. Then report the results for MA(2), etc.

For each model, first you report the model and its theoretical properties. All derivations of the theoretical properties must be allocated to an Appendix, preferably Appendix A. In the main text report only the **result**, that is, the actual (final) formula, such as, for the MA(1) model

$$\begin{aligned}\mu &= 0 \\ \gamma_0 &= \text{some formula} \\ \gamma_1 &= \text{some other formula} \\ &\text{etc.}\end{aligned}\tag{12}$$

Comment on stationarity, for what values of the parameter (or combination of parameters) is the process (covariance) stationary?

Then you comment on the time series plots and corresponding correlogram for each set of parameters. When analysing the plots, for **each model** and **each parameter combination**, use the following questions as a guide to your analysis:

- Given these parameter value is the process stationary? If not, why not?
- Given these exact parameter values, does the model become some kind of special case? If so, what is it and why does it become that?

- Studying the time series plot:
 - How is the times series behaving? Is it choppy? Is is smooth?
 - Comment on *stationarity*, what do these parameter value(s) imply concerning stationarity? How is that being manifested in the time series plot?
 - Does it drift or not? If so, does it drift far?Why?
 - Is there a trend?
 - How often does it cross its mean?
 - Does the variance look stable? Should it?

- Studying the corresponding correlogram ACF:
 - What is its characteristics?
 - Does is have a spike and then nothing? Two spikes? If so, why?
 - Does it decrease/increase gradually? If so, how "fast"?

- Does it oscillate between positive and negative values? If so, why? Refer to the theoretical ACF.
 - Does the correlogram of this process resemble the correlogram for some other process? If so, why?
 - Comment on stationarity, what does the parameter values imply concerning stationarity? How is that being manifested in the correlogram?
- Studying the corresponding Sample **Partial** Autocorrelation SPACF:
 - What is its characteristics?
 - Does it have a spike and then nothing? If so, why?
 - Does it decrease/increase gradually? If so, how "fast"?
 - Does it oscillate between positive and negative values? If so, why?
 - Does the SPACF of this process resemble the SPACF for some other process? If so, why?

Note that the simulation programs **should work**, there is not a problem with the programs. If you get some kind of error message, of missing data or some other error message, there must be something strange going on with the process, with the sample size or mistyped parameter values. If you are sure you have typed in the correct values and it still does not work. Find out what could be "wrong" with the process, and comment on that, if you need to take some special measure to be able to report a plot and a correlogram, do that, and explain why.

Recall that you should later use these setups of "fingerprints" to identify processes that might have generated observed data. These are the tools for *identification*.

5.1 Presentation of plots and correlograms in Appendix A

All the graphs (and tables) must be in **an appedix**, preferably **Appendix B**, but the comments and analysis of them should be in the main text.

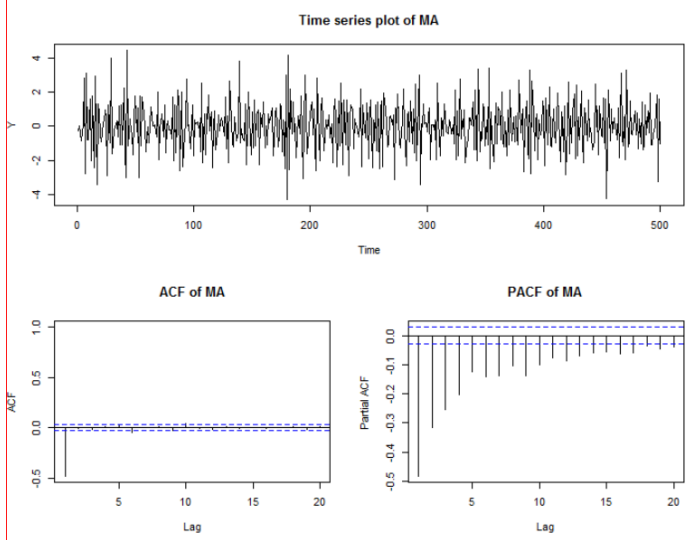
1. Again: All the graphs (and tables) must be in **an appendix**.
2. All graphs and tables must have **name, number and caption** so that you can easily refer to them.
3. The comments must be in the text, no comments in the appendix. Refer to the graphs by number and comment on them that way.

4. There must be no un-clarity about what plot is a realization of what stochastic process with what parameter values.
5. If there a set of parameter combinations that you have already reported, (some parameter combination being a special case of some other model special case, or some model/parameter combination that are recurring you may exclude the reporting of these (additional) visualizations. However, if so, you need to write out that you excluded them and motivate why.

In Appendix A, you collect all the graphs for all the models. To present the graphs you do the following:

- First, you create a **table** that includes a number of parameter/parameter combinations
- inside that table you paste the graphs.

For example for the MA(1) process, the presentation of the graphs should look like the following:

θ	Visualizations of MA(1)
-1	<div style="display: flex; align-items: center;">  (13) </div>
-0.5	Plot here
0	Plot here
0.5	Plot here
1	
2	

Of course you must be careful with the layout. NO PAGE BREAKS IN THE MIDDEL OF A TABLE. Split the table in two if nessecary.

For the MA(2) and AR(2) processes you must split up the graphs in several "tables". To present the graphs, you need to create a table for each of the values of θ_1 and then plot all the combinations together with θ_2 . Then a new table for next value of θ_1 , that is that is, the first table should be

Visualisation of MA(2) (1 of 3)		
		$\theta_1 = -0.4$
$\theta_2 =$	0	Plot here
	0.7	Plot here
	1.0	Plot here

(14)

and

Visualisation of MA(2) (1 of 3)		
		$\theta_1 = 0$
$\theta_2 =$	0	Plot here
	0.7	Plot here
	1.0	Plot here

(15)

finally

Visualisation of MA(2) (1 of 3)		
		$\theta_1 = 0.4$
$\theta_2 =$	0	Plot here
	0.7	Plot here
	1.0	Plot here

(16)

Further:

- Report the results from the AR(1) simulation the same way as for the MA(1)
- Report the results from the AR(2) simulation the same way as for the MA(2)
- Report the results from the ARMA(1,1) simulation the same way as for the MA(2) and AR(2).

6 Random numbers

About the simulations and what to save, here are some bullet points:

1. In the R program file it is possible to set the seed for the random number generator. This enables the user to use the same (psuedo) random numbers when replicating the simulation.
2. As random number generator seed you must use one of the group members date of birth on the format: YYMMDD.

3. That is, if you are born 19-Jan-1995, you set the seed to 950119
4. You *must* use your birthdate as the seed, using a specific seed makes it possible to replicate the results exactly. Also, unfortunately, it has happened in the past that students do not hand in original work. Setting the seed as the birthdate makes it easy to make sure that the results are simulated by that particular group. (Yes, there is the possibility that two persons have the same birthdate...)

End of document.