



Università Commerciale  
Luigi Bocconi

# Lecture 3: Autoregressive Moving Average (ARMA) Models and their Practical Applications

Prof. Massimo Guidolin

20192– Financial Econometrics

Winter/Spring 2020

# Overview

---

- Moving average processes
- Autoregressive processes: moments and the Yule-Walker equations
- Wold's decomposition theorem
- Moments, ACFs and PACFs of AR and MA processes
- Mixed ARMA( $p, q$ ) processes
- Model selection: SACF and SPACF vs. information criteria
- Model specification tests
- Forecasting with ARMA models
- A few examples of applications

# Moving Average Process

**Definition** **(Moving average process)** A  $q$ -th order moving average, denoted as  $MA(q)$ , is a process that can be represented as

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

where the process of  $\{\varepsilon_t\}$  is an IID white noise with mean zero and constant variance equal to  $\sigma_\varepsilon^2$ . More compactly, the process is often written as:

$$y_t = \mu + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

- $MA(q)$  models are always stationary as they are finite, **linear** combination of white noise processes
  - Therefore a  $MA(q)$  process has constant mean, variance and autocovariances that differ from zero up to lag  $q$ , but zero afterwards

$$E(y_t) = \mu$$

$$Var(y_t) = \gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2) \sigma^2$$

$$\gamma_h = Cov(y_t, y_{t-h}) = (\theta_h + \theta_{h+1} \theta_1 + \theta_{h+2} \theta_2 + \dots + \theta_q \theta_{q-h}) \sigma^2 \quad \text{for } h = 1, 2, \dots, q$$

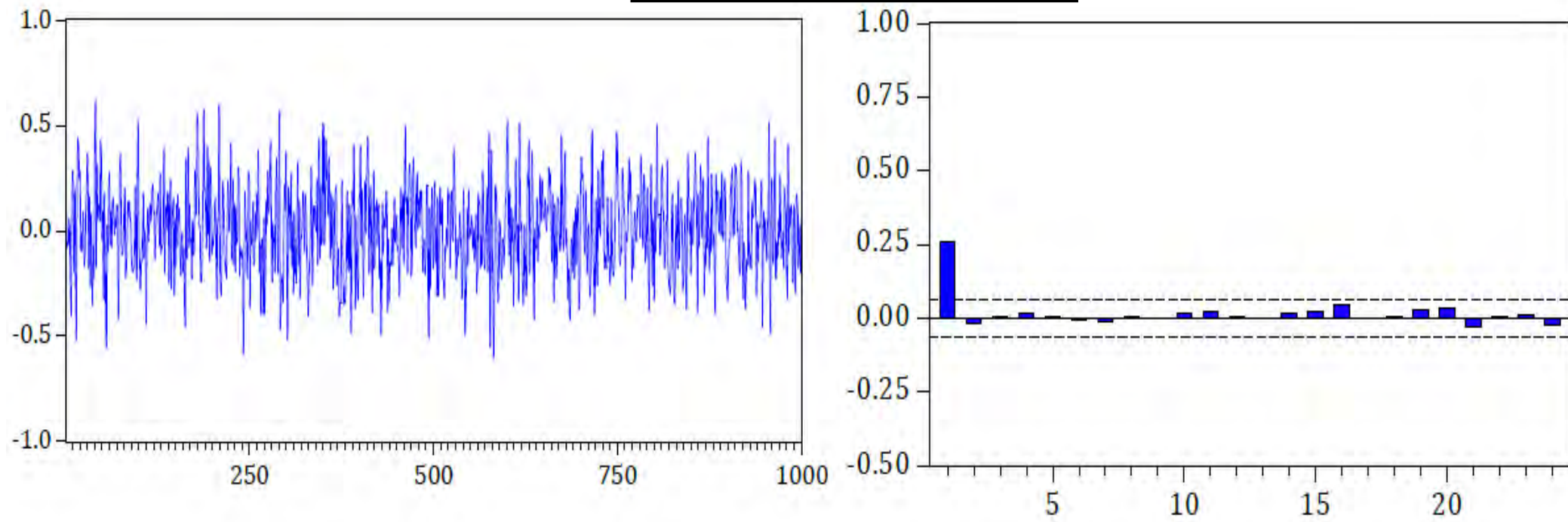
$$\gamma_h = 0$$

$$\text{for } h > q$$

# Moving Average Process: Examples

- MA( $q$ ) models are always stationary as they are finite, linear combination of white noise processes
  - Therefore a MA( $q$ ) process has constant mean, variance and autocovariances that differ from zero up to lag  $q$ , but zero afterwards

MA(1) Simulated Data  
 $R_t = \varepsilon_t + 0.3\varepsilon_{t-1}$

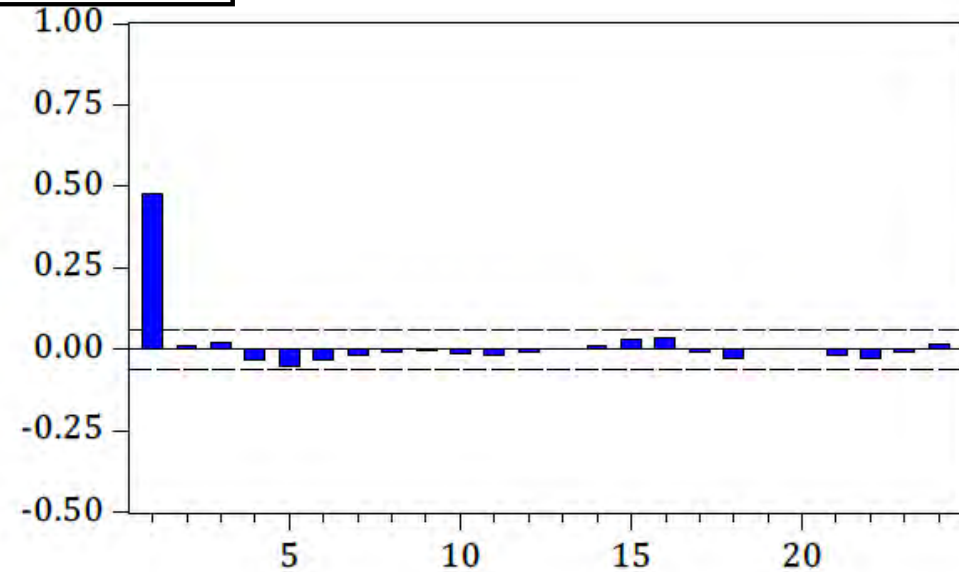
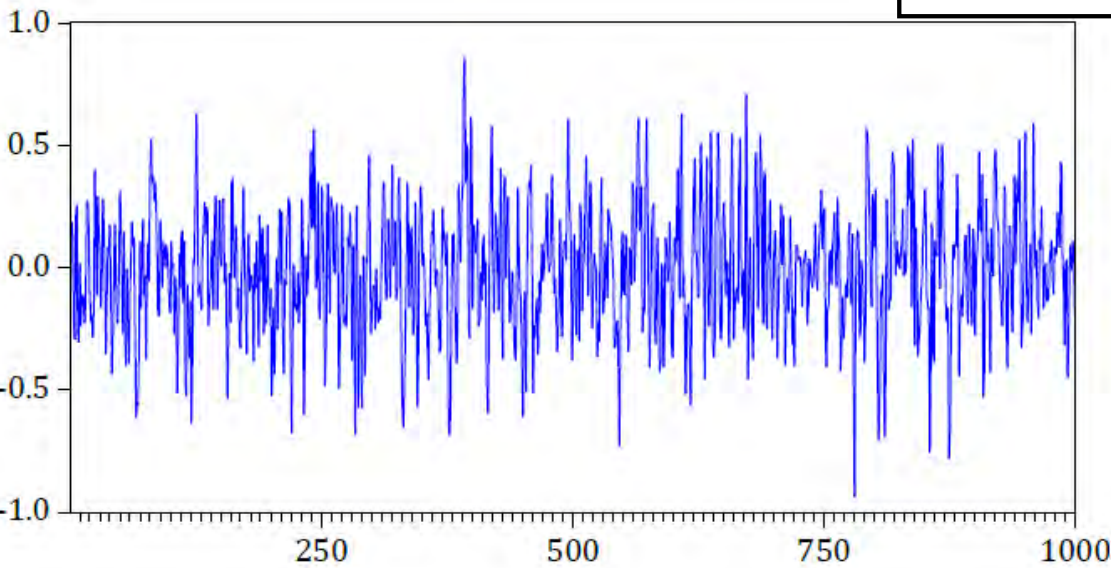


- Simulations are based on  $\varepsilon_t$  IID  $N(0,0.04)$

# Moving Average Process : Examples

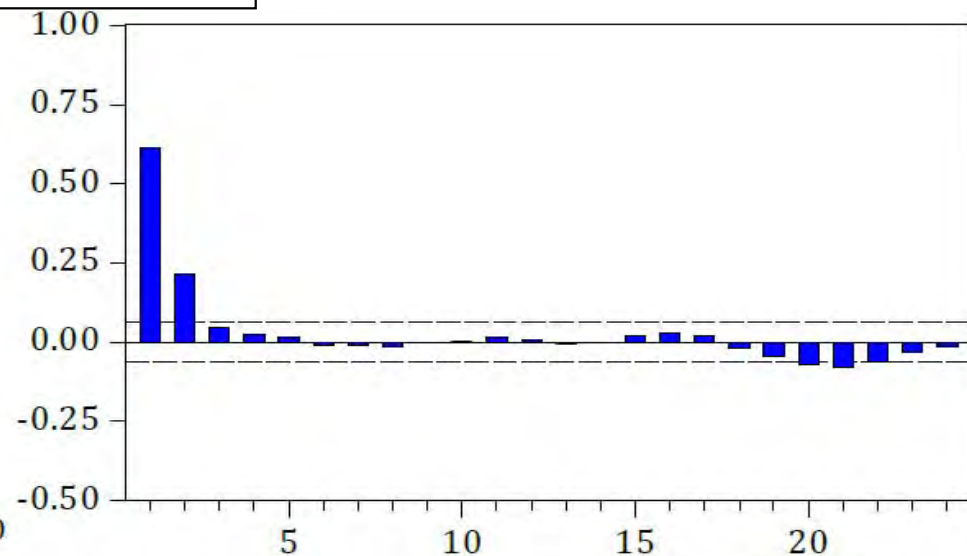
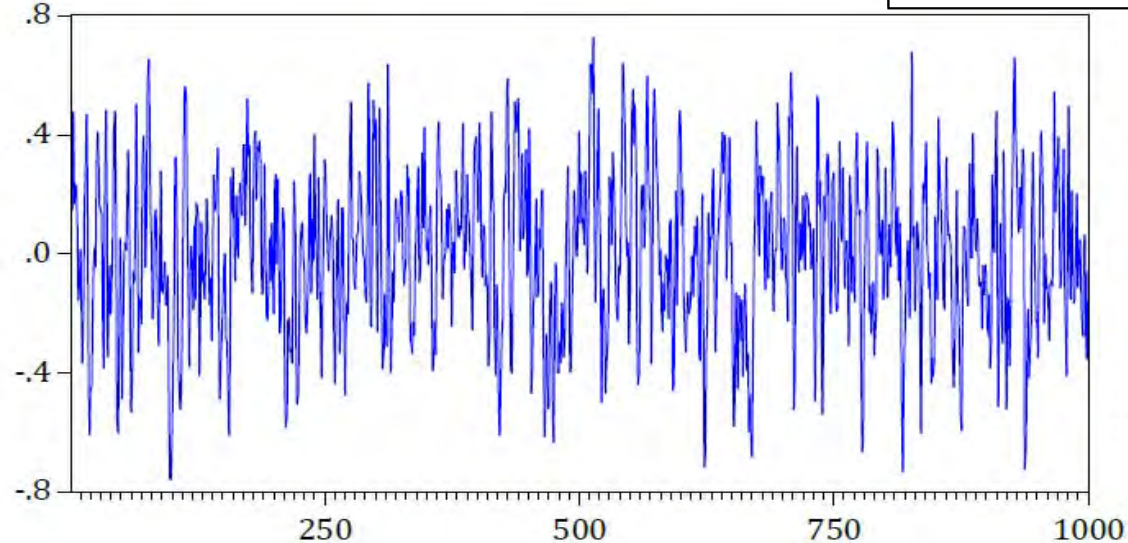
MA(1) Simulated Data

$$R_t = \varepsilon_t + 0.8\varepsilon_{t-1}$$



MA(2) Simulated Data

$$R_t = \varepsilon_t + 0.8\varepsilon_{t-1} + 0.3\varepsilon_{t-2}$$





# Autoregressive Process

- An autoregressive (henceforth AR) process of order  $p$  is a process in which the series  $\{y_t\}$  is a weighted sum of  $p$  past variables in the series  $(y_{t-1}, y_{t-2}, \dots, y_{t-p})$  plus a white noise error term,  $\epsilon_t$ 
  - AR( $p$ ) models are simple univariate devices to capture the observed **Markovian nature** of financial and macroeconomic data, i.e., the fact that the series tends to be influenced at most by a finite number of past values of the same series, which is often also described as the series only having a **finite memory**

**Definition**      **(Autoregressive process)** A  $p$ -th order autoregressive process, denoted as AR( $p$ ), is a process that can be represented by the  $p$ -th order stochastic **difference equation**

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t,$$

where the process of  $\{\epsilon_t\}$  is an IID white noise with mean zero and constant variance  $\sigma_\epsilon^2$ . More compactly, we can write:

$$y_t = \phi_0 + \sum_{j=1}^p \phi_j y_{t-j} + \epsilon_t.$$

# The Lag and Difference Operators

---

- The **lag operator**, generally denoted by  $L$ , shifts the time index of a variable regularly sampled over time backward by one unit
  - Therefore, applying the lag operator to a generic variable  $y_t$ , we obtain the value of the variable at time  $t-1$ , i.e.,  $Ly_t = y_{t-1}$
  - Equivalently, applying  $L^k$  means lagging the variable  $k > 1$  times, i.e.,  
$$L^k y_t = L^{k-1}(Ly_t) = L^{k-1}y_{t-1} = L^{k-2}(Ly_{t-1}) = \cdots = y_{t-k}$$
- The **difference operator**,  $\Delta$ , is used to express the difference between consecutive realizations of a time series,  $\Delta y_t = y_t - y_{t-1}$ 
  - With  $\Delta$  we denote the first difference, with  $\Delta^2$  we denote the second-order difference, i.e.,  $\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$  and so on
  - Note that  $\Delta^2 y_t \neq y_t - y_{t-2}$
  - $\Delta y_t$  can also be rewritten using the lag operator, i.e.,  $\Delta y_t = (1 - L)y_t$
  - More generally, we can write a difference equation of any order,  $\Delta^k y_t$  as  $\Delta^k y_t = (1 - L)^k y_t$ ,  $k \geq 1$

# Stability and Stationarity of AR( $p$ ) Processes

---

- In case of an AR( $p$ ), because it is a stochastic difference equation, it can be rewritten as  $(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) y_t = \phi_0 + \varepsilon_t$  or, more compactly, as  $\phi(L) y_t = \phi_0 + \varepsilon_t$ , where  $\phi(L)$  is a polynomial of order  $p$ ,  $(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)$
- Replacing in the polynomial  $\phi(L)$  the lag operator by a variable  $\lambda$  and setting it equal to 0, i.e.,  $\phi(\lambda) = 0$ , we obtain the **characteristic equation** associated with the difference equation  $\phi(L) y_t = \phi_0 + \varepsilon_t$ 
  - A value of  $\lambda$  which satisfies the polynomial equation is called a **root**
  - A polynomial of degree  $p$  has  $p$  roots, often complex numbers
- If the absolute value of all the roots of the characteristic equations is higher than one the process is said to be **stable**
- **A stable process is always weakly stationary**
  - Even if stability and stationarity are conceptually different, stability conditions are commonly referred to as stationarity conditions



# Wold's Decomposition Theorem

**Result**      **(Wold's decomposition)** Every weakly stationary, purely non-deterministic, stochastic process  $(y_t - \mu)$  can be written as an infinite, linear combination of a sequence of white noise components:

$$y_t - \mu = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \dots = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

- An autoregressive process of order  $p$  with no constant and no other predetermined, fixed terms can be expressed as an infinite order moving average process,  $MA(\infty)$ , and it is therefore **linear**
- If the process is stationary, the sum  $\sum_{i=0}^{\infty} \psi_j \epsilon_{t-j}$  will converge
- The **(unconditional) mean** of an  $AR(p)$  model is

$$\mu = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \dots - \phi_p}$$

- The sufficient condition for the mean of an  $AR(p)$  process to exist and be finite is that the sum of the AR coefficients is less than one in absolute value,  $|\phi_1 + \phi_2 + \dots + \phi_p| < 1$ , see next

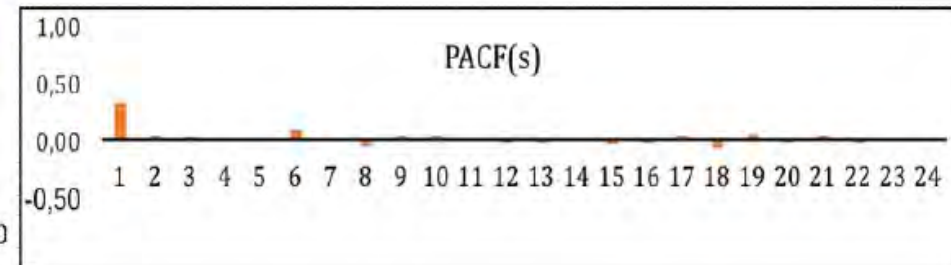
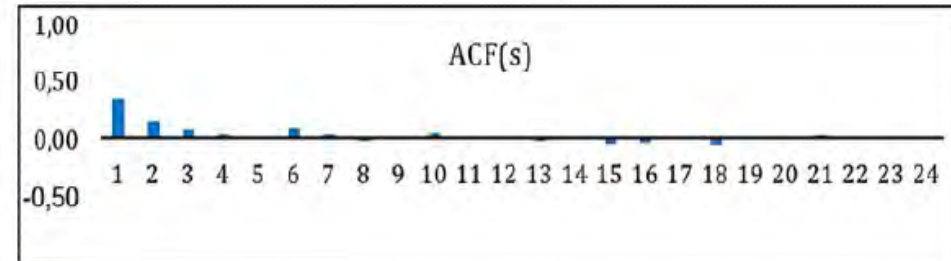
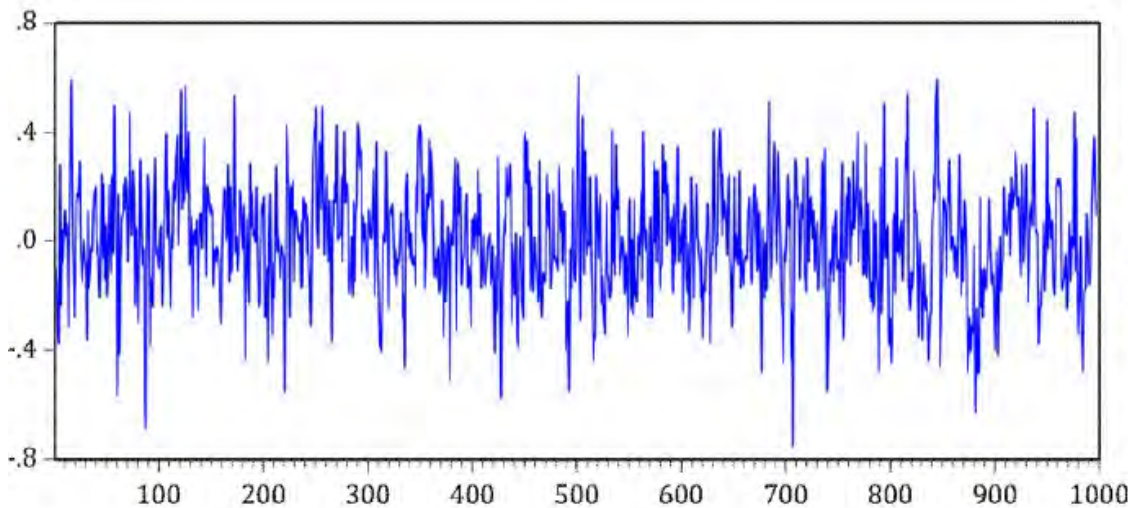
# Moments and ACFs of an AR( $p$ ) Process

- The **(unconditional) variance** of an AR( $p$ ) process is computed from Yule-Walker equations written in recursive form (see below)
  - In the AR(2) case, for instance, we have
$$\text{Var}[y_t] = \frac{(1 - \phi_2)\sigma_\epsilon^2}{(1 + \phi_2)(1 - \phi_1 - \phi_2)(1 + \phi_1 - \phi_2)}$$
  - For AR( $p$ ) models, the characteristic polynomials are rather convoluted – it is infeasible to define simple restrictions on the AR coefficients that ensure covariance stationarity
  - E.g., for AR(2), the conditions are  $\phi_1 + \phi_2 < 1$ ,  $\phi_1 - \phi_2 < 1$ ,  $|\phi_2| < 1$
- The **autocovariances and autocorrelations functions** of AR( $p$ ) processes can be computed by solving a set of simultaneous equations known as **Yule-Walker equations**
  - It is a system of  $K$  equations that we recursively solve to determine the ACF of the process, i.e.,  $\rho_h$  for  $h = 1, 2, \dots$
  - See example concerning AR(2) process given in the lectures and/or in the textbook
- For a stationary AR( $p$ ), **the ACF will decay geometrically to zero**

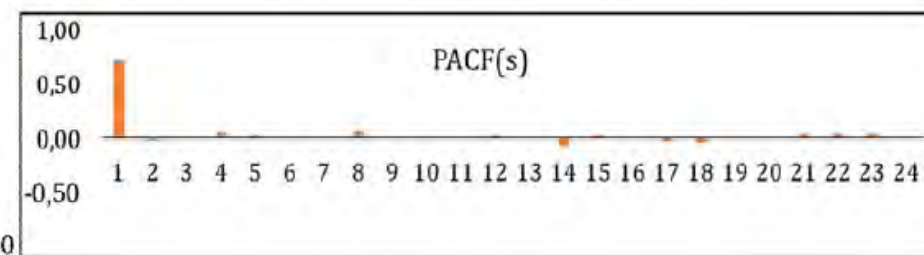
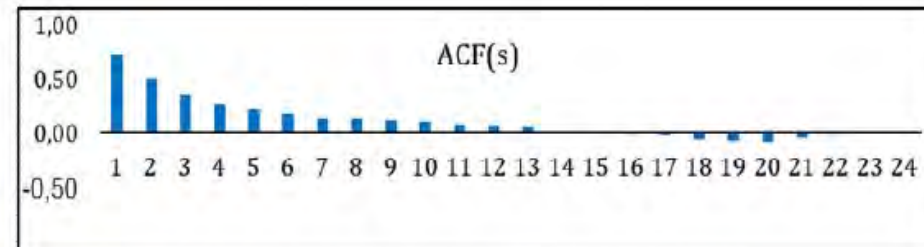
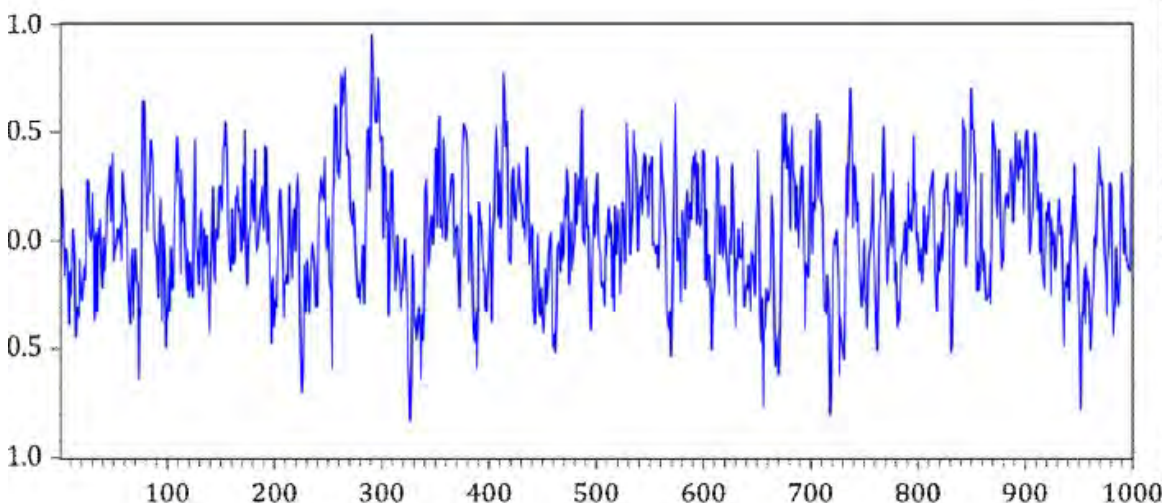
# ACF and PACF of AR( $p$ ) Process

- The SACF and SPACF are of primary importance to identify the lag order  $p$  of a process

AR(1) Simulated Data  
 $R_t = 0.3R_{t-1} + \varepsilon_t$

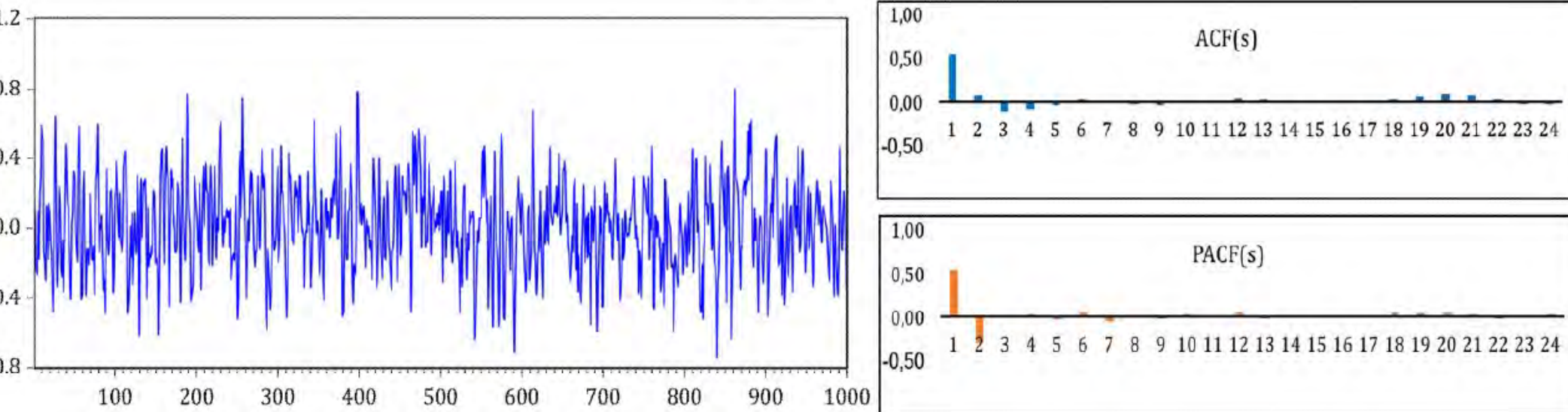


AR(1) Simulated Data  
 $R_t = 0.7R_{t-1} + \varepsilon_t$



# ACF and PACF of AR( $p$ ) and MA( $q$ ) Processes

AR(2) Simulated Data  
 $R_t = 0.7R_{t-1} - 0.3R_{t-2} + \varepsilon_t$



- An AR( $p$ ) process is described by an ACF that may slowly tail off at infinity and a PACF that is zero for lags larger than  $p$
- Conversely, the ACF of a MA( $q$ ) process cuts off after lag  $q$ , while the PACF of the process may slowly tail off at infinity

	ACF	PACF
AR( $p$ )	Decays towards zero.	Cuts off after lag $p$ .
MA( $q$ )	Cuts off after lag $q$ .	Decays towards zero.
ARMA( $p, q$ )	Decays towards zero starting at lag $q$ .	Decays towards zero starting at lag $p$ .



# ARMA( $p, q$ ) Processes

- In some applications, the empirical description of the dynamic structure of the data require us to specify high-order AR or MA models, with many parameters
- To overcome this problem, the literature has introduced the class of autoregressive moving-average (AR-MA) models, combinations of AR and MA models

**Definition**      **(ARMA process)** A time series is said to follow an ARMA( $p, q$ ) if it satisfies

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where the process of  $\{\varepsilon_t\}$  is an IID white noise with mean zero and constant variance equal to  $\sigma_\varepsilon^2$ . More compactly,

$$y_t = \phi_0 + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=0}^q \theta_i \varepsilon_{t-i},$$

with  $\theta_0 = 1$ .



# ARMA( $p, q$ ) Processes

---

- We can also write the ARMA( $p, q$ ) process using the lag operator:

$$(1 - \sum_{i=1}^p \phi_i L^i) y_t = \phi_0 + \sum_{i=0}^q \theta_i \varepsilon_{t-i}$$

- The ARMA( $p, q$ ) model will have a stable solution (seen as a deterministic difference equation) and will be co-variance stationary if the roots of the polynomial  $(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)$  lie outside the unit circle
- The statistical properties of an ARMA process will be a combination of those its AR and MA components
- The unconditional expectation of an ARMA( $p, q$ ) is

$$\mu = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \dots - \phi_p}$$

- An ARMA( $p, q$ ) process gives the same mean as the corresponding ARMA( $p, 0$ ) or AR( $p$ )
- The general variances and autocovariances can be found solving the Yule-Walker equation, see the book

# ARMA( $p, q$ ) Processes

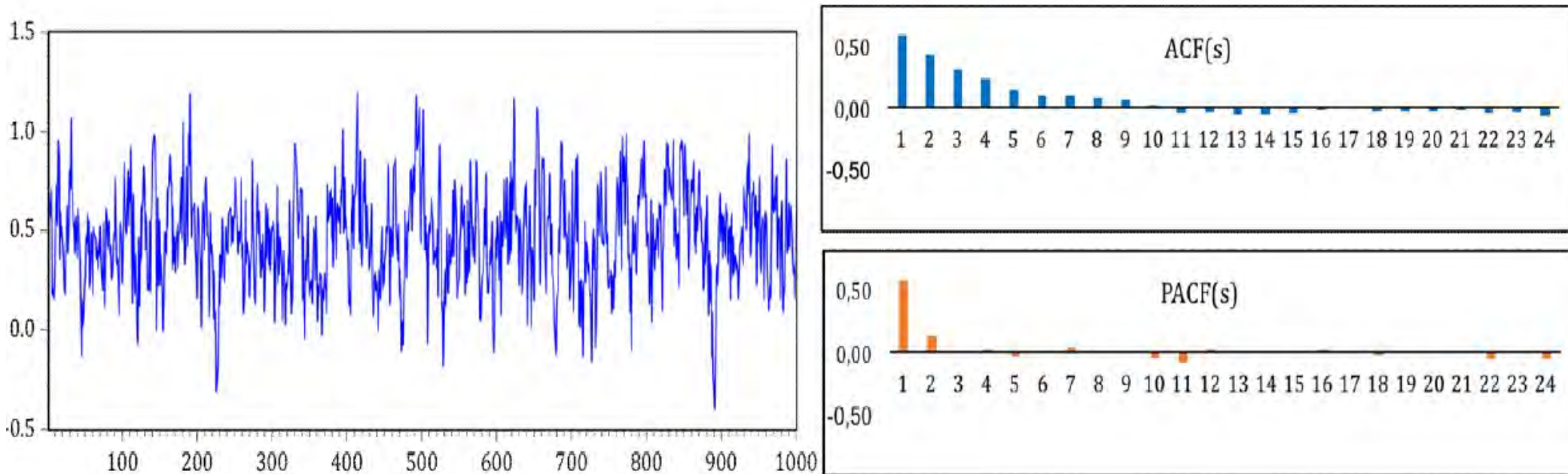
- For a general ARMA( $p, q$ ) model, beginning with lag  $q$ , the values of  $\rho_s$  will satisfy:
$$\rho_s = \frac{\gamma_s}{\gamma_0} = \phi_1 \rho_{s-1} + \phi_2 \rho_{s-2} + \dots + \phi_p \rho_{s-p}$$
  - After the  $q$ th lag, the ACF of an ARMA model is geometrically declining, similarly to a pure AR( $p$ ) model
- The PACF is useful for distinguishing between an AR( $p$ ) process and an ARMA( $p, q$ ) process
  - While both have geometrically declining autocorrelation functions, the former will have a partial autocorrelation function which cuts off to zero after  $p$  lags, while the latter will have a partial autocorrelation function which declines geometrically

	ACF	PACF
AR( $p$ )	Decays towards zero.	Cuts off after lag $p$ .
MA( $q$ )	Cuts off after lag $q$ .	Decays towards zero.
ARMA( $p, q$ )	Decays towards zero starting at lag $q$ .	Decays towards zero starting at lag $p$ .

# ARMA( $p,q$ ) Processes

ARMA(1,1) Simulated Data

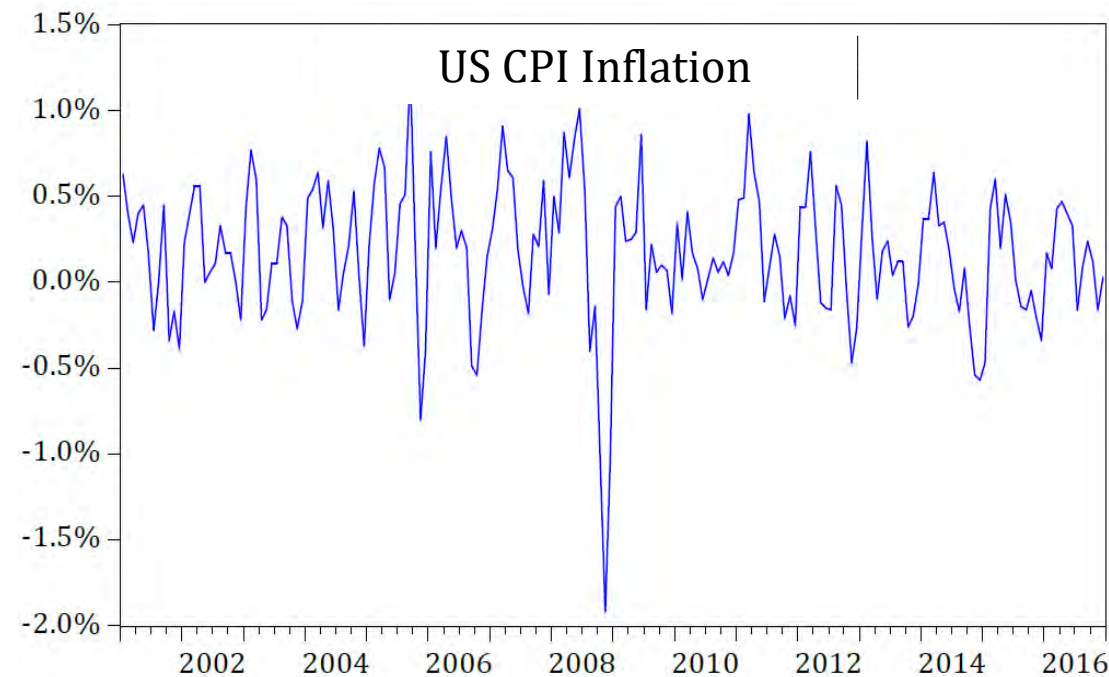
$$R_t = 0.7R_{t-1} - 0.3\varepsilon_{t-1} + \varepsilon_t$$



- As one would expect of an ARMA process, both the ACF and the PACF decline geometrically: the ACF as a result of the AR part and the PACF as a result of the MA part
- However, as the coefficient of the MA part is quite small the PACF becomes insignificant after only two lags. Instead, the AR coefficient is higher (0.7) and thus the ACF dies away after 9 lags and rather slowly

# Model Selection: SACF and SPACF

- A first strategy, compares the sample ACF and PACF with the theoretical, population ACF and PACF and uses them to identify the order of the ARMA( $p, q$ ) model



- Process of some ARMA type, but it remains quite difficult to determine its precise order (especially the MA)

	Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1	0.510	0.510	50.792	0.000		
2	0.069	-0.259	51.714	0.000		
3	-0.147	-0.085	55.976	0.000		
4	-0.156	-0.011	60.806	0.000		
5	-0.181	-0.152	67.340	0.000		
6	-0.190	-0.085	74.574	0.000		
7	-0.193	-0.114	82.082	0.000		
8	-0.221	-0.183	92.008	0.000		
9	-0.149	-0.027	96.518	0.000		
10	0.070	0.123	97.530	0.000		
11	0.275	0.120	113.09	0.000		
12	0.301	0.049	131.88	0.000		
13	0.171	-0.022	137.95	0.000		
14	0.066	0.035	138.87	0.000		
15	-0.008	-0.014	138.89	0.000		
16	-0.059	-0.018	139.63	0.000		
17	-0.099	-0.014	141.70	0.000		
18	-0.122	-0.009	144.90	0.000		
19	-0.116	0.035	147.82	0.000		
20	-0.163	-0.107	153.58	0.000		
21	-0.110	-0.005	156.23	0.000		
22	0.019	0.015	156.30	0.000		
23	0.148	0.040	161.16	0.000		
24	0.326	0.271	184.78	0.000		

# Model Selection: Information Criteria

---

- The alternative is to use information criteria (often shortened to IC)
- They essentially **trade off the goodness of (in-sample) fit and the parsimony of the model** and provide a (cardinal, even if specific to an estimation sample) summary measure
  - We are interested in forecasting out-of-sample: using too many parameters we will end up fitting noise and not the dependence structure in the data, reducing the predictive power of the model (**overfitting**)
- Information criteria include in rather simple mathematical formulations two terms: one which is a function of the sum of squared residual (SSR), supplemented by a penalty for the loss of degrees of freedom from the number of parameters of the model
  - Adding a new variable (or a lag of a shock or of the series itself) will have two opposite effects on the information criteria: it will reduce the residual sum of squares but increase the value of the penalty term
- **The best performing** (promising in out-of-sample terms) **model will be the one that minimizes the information criteria**



# Model Selection: Information Criteria

$$\begin{aligned}
 AIC &= \ln(\hat{\sigma}^2) + \frac{2k}{T}, \\
 HQIC &= \ln(\hat{\sigma}^2) + \frac{2k}{T} \ln(\ln(T)), \\
 SBIC &= \ln(\hat{\sigma}^2) + \frac{2k}{T} \ln(T)
 \end{aligned}$$

Number of parameters  $k$  and Sample size  $T$  are indicated by arrows pointing to the respective terms in the formulas.

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^2$$

- The SBIC is the one IC that imposes the strongest penalty ( $\ln T$ ) for each additional parameter that is included in the model.
- The HQIC embodies a penalty that is somewhere in between the one typical of AIC and the SBIC

- SBIC is a consistent criterion, i.e., it determinates the true model asymptotically
- AIC asymptotically overestimates the order/complexity of a model with positive probability

	Log-likelihood	AIC	SBIC	HQIC
AR(1)	817.289	-8.4822	-8.4313	-8.4616
AR(2)	823.9792	-8.5415	<b>-8.474</b>	-8.514
AR(3)	824.6745	-8.5383	-8.4534	-8.5039
AR(4)	824.6798	-8.5279	-8.4261	-8.4867
MA(1)	818.3171	-8.4929	-8.442	-8.4723
MA(2)	823.7956	-8.5395	-8.4717	-8.5121
MA(3)	824.3971	-8.5354	-8.4506	-8.501
MA(4)	824.8733	-8.5299	-8.4281	-8.4887
ARMA(1,1)	821.745	-8.5182	-8.4503	-8.4907
ARMA(2,1)	826.4089	<b>-8.556</b>	-8.4715	<b>-8.522</b>
ARMA(2,2)	826.9914	-8.552	-8.4502	-8.5108
ARMA(3,2)	827.0005	-8.5417	-8.4229	-8.4936
ARMA(3,3)	827.6836	-8.5384	-8.4026	-8.4834

Information criteria for alternative models used to describe CPI Inflation

# Estimation Methods: OLS vs MLE

---

- It is not uncommon that different criteria lead to different models
- Using the guidance derived from the inspection of the correlogram, we believe that an ARMA model is more likely, given that the ACF does not show signs of geometric decay
- Could be inclined to conclude in favor of a ARMA(2,1) for the US monthly CPI inflation rate
- The estimation of an  $AR(p)$  model because it can be performed simply by (conditional) **OLS**
  - Conditional on  $p$  starting values for the series
- When an  $MA(q)$  component is included, the estimation becomes more complicated and requires **Maximum Likelihood**
  - Please review Statistics prep-course + see the textbook
- However, this opposition is only apparent: conditional on the  $p$  starting values, under the assumptions of a classical regression model, OLS and MLE are identical for an  $AR(p)$ 
  - See 20191 for the classical linear regression model

# Estimation Methods: MLE

---

- The first step in deriving the MLE consists of defining the joint probability distribution of the observed data
- The joint density of the random variables in the sample may be written as a **product of conditional densities** so that the log-likelihood function of ARMA( $p, q$ ) process  $\phi(L)y_t = \theta(L)\varepsilon_t$  has the form

$$\ln L(\phi_1, \dots, \phi_p, \dots, \theta_1, \dots, \theta_q) = \sum_{t=1}^T \ell_t(\cdot)$$

- For instance, if  $\{y_t\}$  has a joint and marginal normal pdf (which must derive from the fact that  $\{\varepsilon_t\}$  has it), then

$$\ell_t(\cdot) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma_\varepsilon^2 - \frac{(\theta(L)^{-1} \phi(L) y_t)^2}{2\sigma_\varepsilon^2}$$

- MLE can be applied to any parametric distribution even when different from the normal
- Under general conditions, the resulting estimators will then be **consistent** and have an **asymptotic normal distribution**, which may be used for inference



# Example: ARMA(2,1) Model of US Inflation

$$CPIInfl_t = \underset{(0.000)}{0.002} + \underset{(0.000)}{1.285}CPIInfl_{t-1} - \underset{(0.000)}{0.555}CPIInfl_{t-2} - \underset{(0.000)}{0.731}\varepsilon_{t-1} + \varepsilon_t$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001717	0.000257	6.681979	0.0000
AR(1)	1.285087	0.082165	15.64032	0.0000
AR(2)	-0.555030	0.055236	-10.04829	0.0000
MA(1)	-0.730921	0.107332	-6.809925	0.0000
SIGMASQ	1.07E-05	9.63E-07	11.06491	0.0000
R-squared	0.330274	Mean dependent var		0.001716
Adjusted R-squared	0.315948	S.D. dependent var		0.003999
S.E. of regression	0.003308	Akaike info criterion		-8.556343
Sum squared resid	0.002046	Schwarz criterion		-8.471513
Log likelihood	826.4089	Hannan-Quinn criter.		-8.521986
F-statistic	23.05464	Durbin-Watson stat		1.919884
Prob(F-statistic)	0.000000			
Inverted AR Roots	.64-.38i	.64+.38i		
Inverted MA Roots	.73			

# Model Specification Tests

---

- If the model has been specified correctly, all the structure in the (mean of the) data ought to be captured and the residuals shall not exhibit any predictable patterns
- Most diagnostic checks involve the analysis of the residuals
- ① An intuitive way to identify potential problems with a ARMA model is to **plot the residuals** or, better, the standardized residuals, i.e.,  $\hat{\varepsilon}_t^s = (\hat{\varepsilon}_t - \hat{\bar{\varepsilon}}) / \hat{\sigma}_{\varepsilon}$ 
  - If the residuals are normally distributed with zero mean and unit variance, then approximately 95% of the standardized residuals should fall in an interval of  $\pm 2$  around zero
  - Also useful to plot the squared (standardized) residuals: if the model is correctly specified, such a plot of squared residuals should not display any clusters, i.e., the tendency of high (low) squared residuals to be followed by other high (low) squared standardized residuals
- ② A more formal way to test for normality of the residuals is the **Jarque-Bera test**



# Model Specification Tests: Jarque-Bera Test

- Because the normal distribution is symmetric, the third central moment, denoted by  $\mu_3$ , should be zero; and the fourth central moment,  $\mu_4$ , should satisfy  $\mu_4 = 3\sigma_\epsilon^4$
- A typical index of asymmetry based on the third moment (**skewness**), that we denote by  $\hat{S}$ , of the distribution of the residuals is

$$\hat{S} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{\epsilon}_t^3}{\hat{\sigma}_\epsilon^3}$$

- The most commonly employed index of tail thickness based on the fourth moment (**excess kurtosis**), denoted by  $\hat{K}$ , is

$$\hat{K} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{\epsilon}_t^4}{\hat{\sigma}_\epsilon^4} - 3$$

- If the residuals were normal,  $\hat{S}$  and  $\hat{K}$  would have a zero-mean asymptotic distribution, with variances  $6/T$  and  $24/T$ , respectively
- The Jarque-Bera test concerns the composite null hypothesis:

$$H_0 : \frac{\mu_3}{\sigma^3} = 0 \quad \text{and} \quad H_0 : \frac{\mu_4}{\sigma^4} - 3 = 0$$

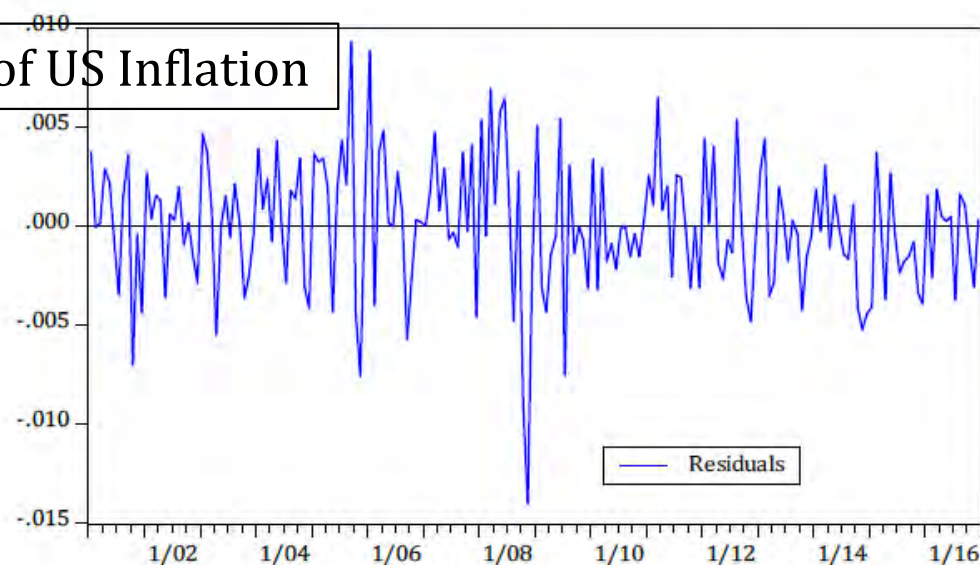
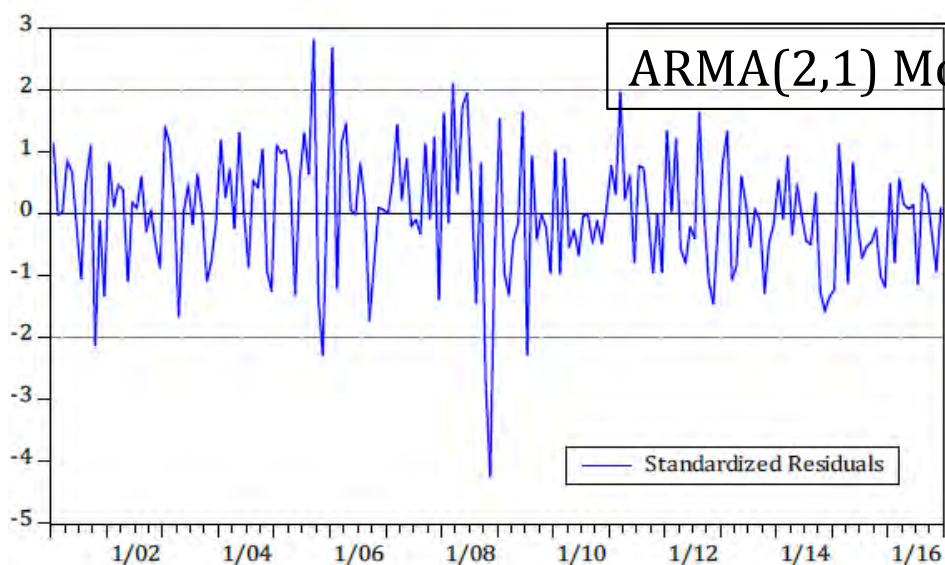
# Model Specification Tests: Jarque-Bera Test

- **Jarque and Bera** prove that because the square roots of the sample statistics

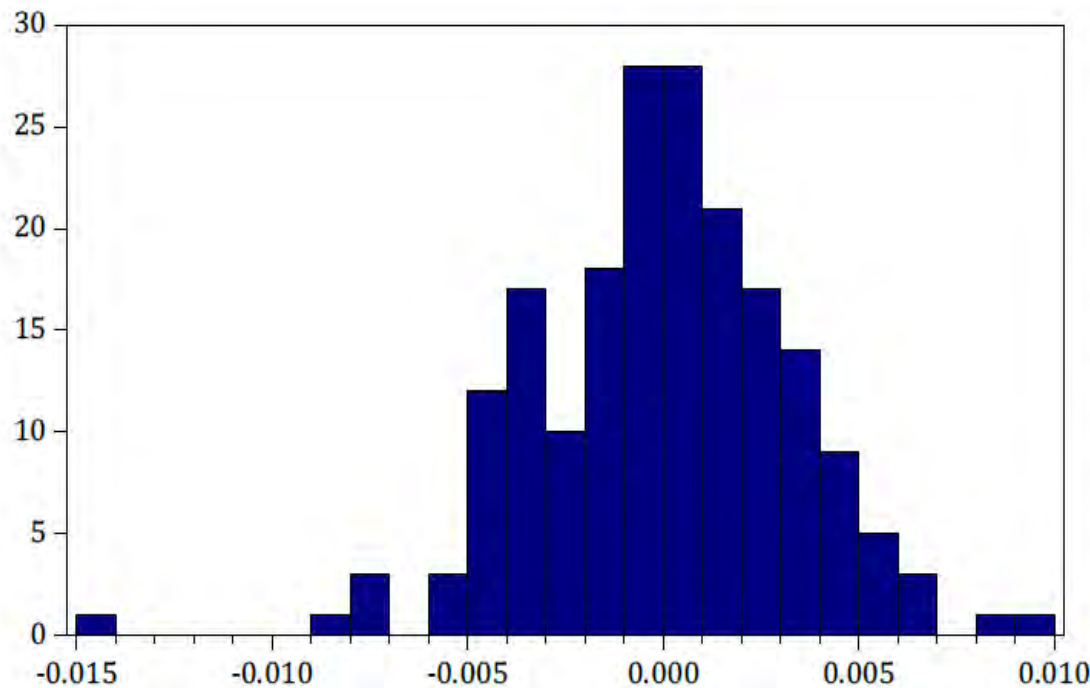
$$\lambda_1 = \frac{1}{\sqrt{6T}} \sum_{t=1}^T \left( \frac{\hat{\varepsilon}_t^3}{\hat{\sigma}_\varepsilon^3} \right) \quad \lambda_2 = \frac{1}{\sqrt{24T}} \left( \sum_{t=1}^T \frac{\hat{\varepsilon}_t^4}{\hat{\sigma}_\varepsilon^4} - 3 \right)$$

are  $N(0,1)$  distributed, the null consists of a joint test that  $\lambda_1$  and  $\lambda_2$  are zero tested as  $H_0: \lambda_1^2 + \lambda_2^2 = 0$ , where  $\lambda_1^2 + \lambda_2^2 \sim \chi_2^2$  as  $T \rightarrow \infty$

- ③ Compute **sample autocorrelations of residuals** and perform tests of hypotheses to assess whether there is any linear dependence
  - Same portmanteau tests based on the Q-statistic can be applied to test the null hypothesis that there is no autocorrelation at orders up to  $h$



# Example: ARMA(2,1) Model of US Inflation



Series: Residuals  
Sample 2001M01 2016M12  
Observations 192

Mean 8.16e-06  
Median 6.05e-05  
Maximum 0.009310  
Minimum -0.014030  
Std. Dev. 0.003273  
Skewness -0.333498  
Kurtosis 4.360373

Jarque-Bera 18.36399  
Probability 0.000103

**Residuals**

**Squared Residuals**

Residuals							Squared Residuals						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob		Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1 0.037	0.037	0.2610	0.609				1 0.182	0.182	6.4907	0.011	
		2 -0.062	-0.064	1.0145	0.602				2 0.188	0.160	13.450	0.001	
		3 -0.059	-0.054	1.6911	0.639				3 0.078	0.022	14.661	0.002	
		4 0.100	0.101	3.6570	0.454				4 0.094	0.052	16.425	0.002	
		5 0.014	-0.001	3.6943	0.594				5 0.112	0.080	18.945	0.002	
		6 -0.004	0.004	3.6982	0.717				6 0.053	0.001	19.508	0.003	
		7 0.020	0.034	3.7819	0.805				7 0.059	0.018	20.215	0.005	
		8 -0.088	-0.102	5.3392	0.721				8 0.244	0.230	32.264	0.000	
		9 -0.120	-0.114	8.2665	0.508				9 0.012	-0.087	32.294	0.000	
		10 0.019	0.022	8.3397	0.596				10 0.040	-0.038	32.625	0.000	
		11 0.164	0.139	13.826	0.243				11 -0.025	-0.032	32.754	0.001	
		12 0.148	0.154	18.319	0.106				12 -0.000	-0.022	32.754	0.001	

# Forecasting with ARMA

---

- **In-sample forecasts** are those generated with reference to the same data that were used to estimate the parameters of the model
  - The R-square of the model is a measure of in-sample goodness of fit
  - Yet, ARMA are time series models in which the past of a series is used to explain the behavior of the series, so that using the R-square to quantify the quality of a model faces limitations
- More interested in how well the model performs when it is used to **forecast out-of-sample**, i.e., to predict the value of observations that were not used to specify and estimate the model
- Forecasts can be one-step-ahead,  $\hat{y}_t(1)$ , or multi-step-ahead,  $\hat{y}_t(h)$
- In order to evaluate the usefulness of a forecast we need to specify a **loss function** that defines how concerned we are if our forecast were to be off relative to the realized value, by a certain amount.
- Convenient results obtain if one assumes a quadratic loss function, i.e., the minimization of:  $MSFE[\hat{y}_t(h)] \equiv E[(y_{t+h} - \hat{y}_t(h))^2]$

# Forecasting with AR( $p$ )

- This is known as the **mean square forecast error** (MSFE)
- It is possible to prove that MSFE is minimized when  $\hat{y}_t(h)$  is equal to  $E[y_{t+h} | \mathfrak{I}_t]$  where  $\mathfrak{I}_t$  is the information set available
- In words, **the conditional mean of  $y_{t+h}$  given its past observations is the best estimator of  $\hat{y}_t(h)$  in terms of MSFE**
- In the case of an AR( $p$ ) model, we have:

$$\hat{y}_t(h) = E[y_{t+h} | y_t, y_{t-1}, \dots] = \phi_0 + \sum_{i=1}^p \phi_i \hat{y}_t(h-i)$$

where  $\hat{y}_t(i) = y_{t+i}$  if  $i \leq 1$

- For instance,  $\hat{y}_t(1) = E[y_{t+1} | y_t, y_{t-1}, y_{t-2}, \dots] = \phi_0 + \sum_{i=1}^p \phi_i y_{t+1-i}$
- The forecast error is  $u_t(1) = y_{t+1} - \hat{y}_t(1) = \varepsilon_{t+1}$
- The  $h$ -step forecast can be computed recursively, see the textbook/class notes
- For a stationary AR( $p$ ) model,  $\hat{y}_t(h)$  converges to the mean  $E[y_t]$  as  $h$  grows, the **mean reversion property**



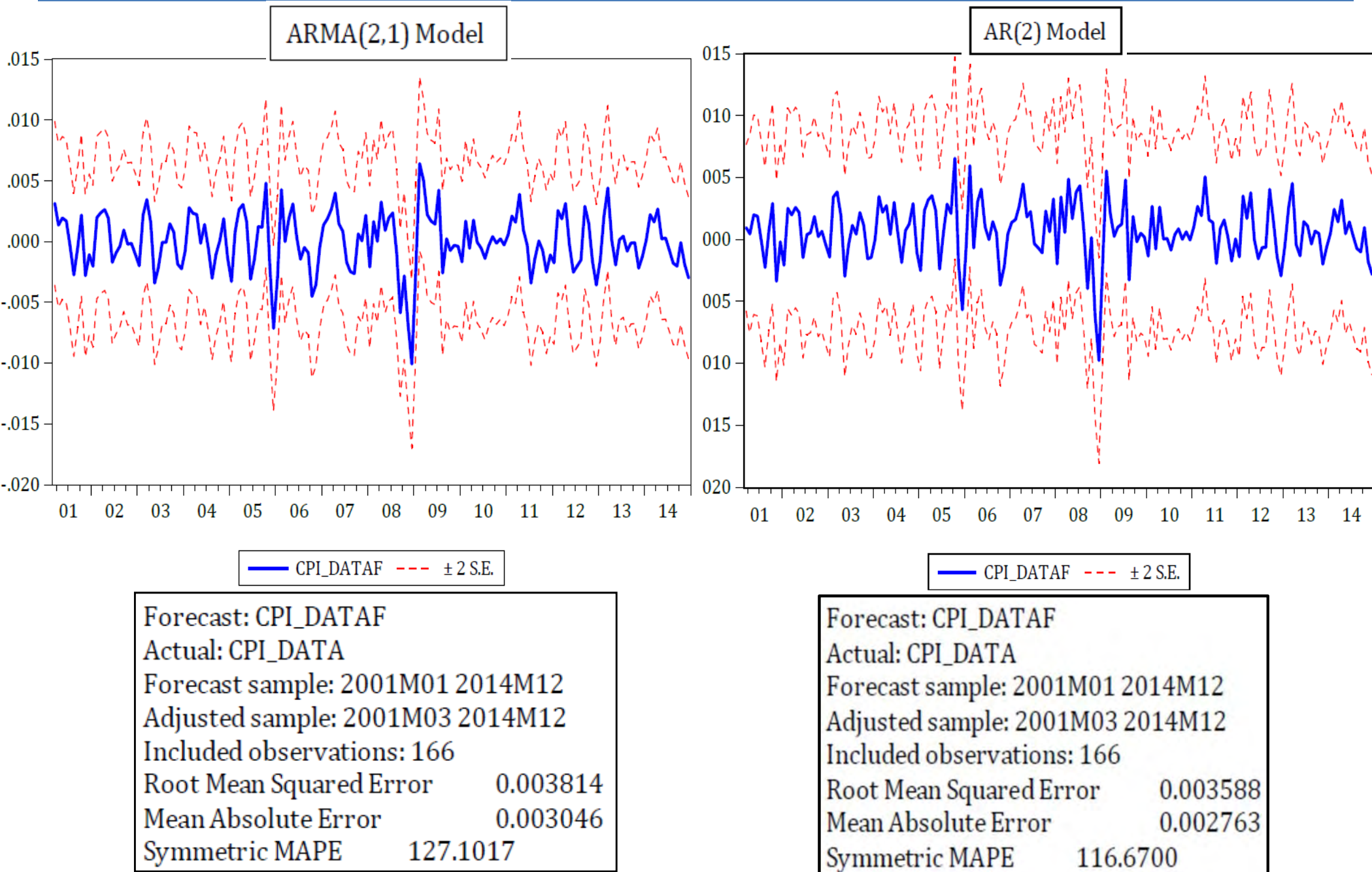
# Forecasting with MA( $q$ )

- Because the model has a memory limited to  $q$  periods only, the point forecasts converge to the mean quickly and they are forced to do so when the forecast horizon exceeds  $q$  periods
  - E.g., for a MA(2),  $\hat{y}_t(1) = E[y_{t+1} | y_t, y_{t-1}, \dots] = \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1}$  because both shocks have been observed and are therefore known
  - Because  $\varepsilon_t$  has not yet been observed at time  $t$ , and its expectation at time  $t$  is zero, then  $\hat{y}_t(2) = E[y_{t+2} | y_t, y_{t-1}, \dots] = \mu + \theta_2 \varepsilon_t$
  - By the same principle,  $\hat{y}_t(3) = E[y_{t+3} | y_t, y_{t-1}, \dots] = \mu$  because  $\varepsilon_{t+3}$ ,  $\varepsilon_{t+2}$ , and  $\varepsilon_{t+1}$  are not known at time  $t$
- By induction, the forecasts of an ARMA( $p, q$ ) model can be obtained from
 

$$\hat{y}_t(h) = \phi_0 + \sum_{i=1}^p \phi_i \hat{y}_t(h-i) + \sum_{j=1}^q \theta_j E_t[\varepsilon_{t+h-j}]$$
- How do we assess the forecasting accuracy of a model?

$$MSFE(h) = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^T (y_{t+h} - \hat{y}_t(h))^2 \quad MAPE(h) = \frac{100}{T - (T_1 - 1)} \sum_{t=T_1}^T \left| \frac{y_{t+h} - \hat{y}_t(h)}{y_{t+h}} \right|$$

# Forecasting US CPI Inflation with ARMA Models



# A Few Examples of Potential Applications

---

- To fit the time series dynamics of the **earnings** of company  $i = 1, 2, \dots, N$  with  $\text{ARMA}(p_i, q_i)$  models, to compare the resulting forecasts with earnings forecasts published by analysts
  - Possibly, also compare the 90% forecast confidence intervals from  $\text{ARMA}(p_i, q_i)$  with the dispersion over time of analysts forecasts
  - Possibly also price stocks of each company using some DCF model and compare it with the target prices published by the analysts
- To fit the time series dynamics of **commodity future returns** (for a range of underlying assets) using  $\text{ARMA}(p, q)$  models to forecast
  - Possibly, compare such forecasts with those produced by predictive regressions that just use (or also use) commodity-specific information
  - A predictive regression is a linear model to predict the conditional mean with structure  $\hat{y}_t(h) = \alpha + \sum_{i=1}^K \hat{\beta}_i x_{i,t}$  where the  $x_{1,t}, x_{2,t}, \dots, x_{K,t}$  are the commodity-specific variables
  - Possibly, to try and understand why and when only the past of a series helps to predict future returns or not (i.e., for which commodities)

# A Few Examples of Potential Applications

- Given the mkt. ptf., use **mean-variance portfolio theory** (see 20135, part 1) and ARMA models to forecast the (conditional risk premium) and decide the optimal weight to be assigned to risky vs. riskless assets
  - Also called **strategic asset allocation** problem
  - As you will surely recall,  $\hat{\omega}_t = (1/\lambda\sigma^2) \times \text{Forecast of } (r_{t+1}^{ARMA} - r_{t+1}^f)$
  - Similar/partly identical to a question of Homework 2 in 20135!?
  - Possibly compare with the performance results (say, Sharpe ratio) produced by the strategy  $\hat{\omega}_t = (1/\lambda\sigma^2) \times \text{Hist. Mean of } (\hat{r}_{t+1} - r_{t+1}^f)$  which results from ARMA(0,0) processes (== white noise returns)
- After measuring which portion of a given policy or company announcement represents news/unexpected information, measure **how long it takes for the news to be incorporated in the price**
  - Equivalent to test the number of lags  $q$  in a ARMA(0,  $q$ ) model
  - Unclear what the finding of  $p > 0$  could mean in a ARMA( $p$ ,  $q$ ) case
  - Related to standard and (too) popular **event studies**