Univariate Timeseries Models I

Lecture 2
ECON5345, Spring 2026

Byoungchan Lee

January 1, 2026

Contents

- Some Definitions
- Wold Decomposition
- ARIMA
  - Basic Properties
  - Model Selection and Estimation
  - Forecasting
- Conditional Heteroskedasticity: ARCH and GARCH

In Lecture 3, we will cover the following:

- Spectral Analysis
  - Estimation of the LRV
  - Filtering (HP, BP, Exponential)
- Unit Root Tests

“Facts” about macroeconomic data

- Typically time series (natural order)
- No strict exogeneity (biased, but hopefully consistent estimates)
- Strong serial correlation (dependence of observations)
- Many series grow over time (trends)
- We see only one history (realization) of events.

If we have only one history of events, we need to assume some structures to make statistical statements applicable for other possible histories.

Stationarity

Consider a sequence of random variables (rv), denoted by x_t.

Strict Stationarity.
The process $\{x_t\}$ is strictly stationary if the joint distribution of $(x_t, x_{t+1}, \ldots, x_{t+k})$ is same for all t. For example, (x_1, x_5) has the same joint distribution as (x_{12}, x_{16}).

ex) $x_t \sim iid\mathcal{N}(0, 1)$.

Stationarity II

(Covariance) Stationarity, Weak Stationarity.
We often use the first two moments to characterize the properties of data. For the t^{th} observation of a time series,

- Mean: $\mu_t = \mathbb{E}(x_t)$;
- Variance: $\gamma_t(0) = var(x_t) = \mathbb{E}(x_t - \mu_t)^2$;
- Autocovariance at lag k:

$$\gamma_t(k) = cov(x_t, x_{t-k}) = \mathbb{E}[(x_t - \mu_t)(x_{t-k} - \mu_{t-k})];$$

- Autocorrelation at lag k:

$$\rho_t(k) = \gamma_t(k)/[\sqrt{\gamma_t(0)}\sqrt{\gamma_{t-k}(0)}].$$

A stochastic process x_t is covariance stationary if the first and second moments do not depend on t and are finite. That is, $\mathbb{E}(x_t) = \mu < \infty$, $\gamma_t(0) = \gamma(0) < \infty$, and $\gamma_t(k) = \gamma(k)$ for all t and k ($\rho_t(k) = \rho(k)$ follows). This means that $cov(x_t, x_{t-k})$ depends only on k and not t.

Ergodicity and Mixing I

Ensemble average vs. Time average

$\mu_t = \mathbb{E}(x_t)$ is an ensemble average (the expectation of x_t at fixed t across the ensemble of all possible realizations of the sequence).

$$\mu_t \approx \frac{1}{M} \sum_{m=1}^{M} x_t^{(m)},$$

for M realizations of the stochastic process, denoted by $\left\{ x_t^{(m)} \right\}$.

However, we have only one realized sample path. Can we say something about the population μ_t (an ensemble average) based on a time average of the data, $\frac{1}{T} \sum_{t=1}^{T} x_t$?

Ergodicity and Mixing II

We assume covariance stationarity, $\mu_t = \mu$ and $\gamma_t(k) = \gamma(k)\ \forall t$. A covariance stationary process is

1. **ergodic** for the mean, if $\frac{1}{T}\sum_{t=1}^{T} x_t \xrightarrow{p} \mu$;
2. **ergodic** for second mements, if $\frac{1}{T-k}\sum_{t=k+1}^{T}(x_t - \mu)(x_{t-k} - \mu) \xrightarrow{p} \gamma(k)$ for all k

ex) Suppose that $x_t = x$ for all t, where $x \sim \mathcal{N}(0,1)$. One can show that this is a covariance stationary process with $\mu = 0$ and $\gamma(k) = 1$ for all k. However, each sample path will look like a constant. We cannot back out $\mathbb{E}[x]$ and $var(x)$ from the data. It is because this process is too persistent. To capture the ensemble averages, the stochastic process should travel all possible realizations on one long sample path.

Ergodicity and Mixing III

To have ergodicity (or LLN, CLT, etc.), the serial correlations should die out with the time gap. If x_t and x_{t+k} tend to "become" independent as $k \to \infty$, one long history will be informative enough for other potential histories. A related mathematical condition is mixing.

(Strong) Mixing. A stationary process x_t is said to be strongly mixing if

$$\alpha(k) \to 0 \text{ as } k \to \infty, \text{ where}$$
$$\alpha(k) = \sup\{|P(A \cap B) - P(A)P(B)| :$$
$$A \in \sigma(\dots, x_{t-1}, x_t), B \in \sigma(x_{t+k}, x_{t+k+1}, \dots), \forall t\}.$$

Types of innovations (or shocks)

Building blocks of a time-series model:

- **White noise (WN)**: $E(e_t) = 0$, $E(e_t e_s) = 0$ for $t \neq s$, $var(e_t) = \sigma^2$.
- **Martingale Difference Sequence (mds)**: $\mathbb{E}_{t-1}[e_t] = 0$. This means that no function of the current information can help predict e_{t+s} in the future.
- **Independent white noise**: $e_t \sim iid(0, \sigma^2)$.
- **Gaussian white noise**: $e_t \sim iid\mathcal{N}(0, \sigma^2)$.

Wold Decomposition

The Wold Decomposition

Any zero-mean covariance stationary process $\{x_t\}$ that is not deterministic can be expressed as a sum $x_t = d_t + y_t$, where $y_t = \sum_{i=0}^{\infty} \psi_i e_{t-i}$, and

1. $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$,
2. $e_t \equiv x_t - P(x_t | x_{t-1}, x_{t-2}, \ldots)$ is $WN(0, \sigma^2)$, where $\sigma > 0$,
3. $\mathbb{E}\left[e_t d_s\right] = 0$ for all t and s,
4. $d_t = P(d_t | x_{t-1}, x_{t-2}, \ldots)$ (predicted by past values of x).

Remarks

1. We can focus on MA(∞) processes and their approximations admitting more parsimonious representations.
2. The "shock" e_t need not have a structural interpretation. For example, if x is a GDP growth rate, e can be a complicated combination of structural shocks to productivity, monetary policy, fiscal policy, uncertainty, risk premium, etc.

Dynamic responses

Consider $y_t = \sum_{i=0}^{\infty} \psi_i e_{t-i}$.

Define the (j period) dynamic multiplier (or the impulse response) as $dy_{t+j}/de_t = \psi_j$. It measures the effect of a shock at time t, j periods later.

The *impact* effect is $dy_t/de_t = \psi_0$.

The j period cumulative dynamic multiplier is the cumulative effect of a change in e_t on y_t over the next j periods, and thus equals $\psi_0 + \psi_1 + \ldots + \psi_j$.

The sum of all individual multipliers is the cumulative long-run dynamic multiplier, $\psi(1) = \sum_{j=0}^{\infty} \psi_j$.

ARMA: General formulation

$$\begin{aligned} x_t &= c + \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + e_t + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} \\ &\Leftrightarrow (1 - \alpha_1 L - \ldots - \alpha_p L^p) x_t = c + (1 + \theta_1 L + \ldots + \theta_q L^q) e_t \\ &\Leftrightarrow \alpha(L) x_t = c + \theta(L) e_t \\ &\Leftrightarrow x_t = c/\alpha(1) + (\theta(L)/\alpha(L)) e_t = \mu + \psi(L) e_t \end{aligned}$$

The building block of ARMA models is e_t, a white noise process.

A covariance stationary process ($x_t = d_t + \sum_{i=0}^{\infty} \psi_i e_{t-i}$) can be approximated by a deterministic process + an ARMA(p,q) model with a finite number of parameters that can be estimated.

Great tool:

- Summarize dynamic properties of the data.
- Useful in forecasting (benchmark for structural models).

Moving Average (MA)

q^{th} **order moving average process (MA(q)).**

$$x_t = \mu + e_t + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} = \mu + \theta(L)e_t, e_t \sim WN(0, \sigma^2)$$

MA(q) is a linear, one-sided process because future values of e_t are not included.

- Mean:$E(x_t) = \mu$
- Variance: $\gamma(0) = E[(x_t - \mu)^2] = \sigma^2(1 + \theta_1^2 + \ldots + \theta_q^2)$
- Covariance:$\gamma(j) = \begin{cases} \sigma^2(\theta_j + \theta_{j+1}\theta_1 + \ldots + \theta_q\theta_{q-j}), j = 1, \ldots, q \\ 0, j > q \end{cases}$
- The serial correlation dies out after q lags.

It is known that if a covariance stationary process has $\gamma(j) = 0$ for $j > q$, it admits a MA(q) representation.

Moving Average: Invertibility

MA(q) process is **invertible** if the roots (or the zeros) of $\theta(z) = (1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q) = 0$ lie outside the unit circle.

Invertibility means that we can write $e_t = \sum_{i=0}^{\infty} \pi_i (x_{t-i} - \mu)$ where $\pi(L) = \theta(L)^{-1}$. That is, we can identify the shocks using an autoregression of the observables x_t, justifying the use of AR models. However, this may not be the case for some DSGE models.

Example: MA(1) with $\theta(z) = (1 + \theta_1 z) = 0$. It is 0 at $z = -1/\theta_1$. The invertibility condition is $|\theta_1| < 1$. If $|\theta_1| < 1$,

$$\begin{aligned} x_t &= (1 + \theta_1 L) e_t \\ e_t &= (1 + \theta_1 L)^{-1} x_t = \sum_{i=0}^{\infty} (-\theta_1)^i x_{t-i} \\ &= x_t - \theta_1 x_{t-1} + \theta_1^2 x_{t-2} - \ldots \end{aligned}$$

Autoregressive models

AR(p) model.

$$x_t = c + \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + e_t, e_t \sim WN(0, \sigma^2).$$

Equivalently,

$$\alpha(L)(x_t - \mu) = e_t, \text{ where } \alpha(L) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$$
$$\text{and } \mu = \frac{c}{\alpha(1)} = \frac{c}{1 - \alpha_1 - \alpha_2 - \ldots - \alpha_p}.$$

Example: AR(1) with $|\alpha_1| < 1$

$$\begin{aligned} (1-\alpha_1 L)(x_t-\mu) &= e_t \\ x_t - \mu &= (1-\alpha_1 L)^{-1} e_t = \sum_{i=0}^{\infty} \alpha_1^i e_{t-i} \\ x_t &= \mu + e_t + \alpha_1 e_{t-1} + \alpha_1^2 e_{t-2} + \dots \end{aligned}$$

- Mean:$E(x_t) = \mu$
- Variance:$var(x_t) = \gamma(0) = \sigma^2(1+\alpha^2+\alpha^4+\dots) = \sigma^2/(1-\alpha^2)$
- Autocovariance: $\gamma(j) = \alpha^{|j|}\sigma^2/(1-\alpha^2)$
- Autocorrelation: $\rho(j) = \alpha^{|j|}$
- An AR(1) process is **stable / causal** if the root of $\alpha(z) = 1 - \alpha z = 0$ lies outside the unit circle. This is satisfied if $|\alpha| < 1$.
- For a stable AR(1), $\gamma(j)$ and the dynamic multiplier α^j go to zero gradually, and how fast they go to zero depends on $|\alpha|$.

AR(p): Companion form

An AR(p) model can be rewritten as a p vector AR(1) model using the **companion** matrix form:

$$\begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-p+1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-p} \end{bmatrix} + \begin{bmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We can write this equation as vector autoregression of order 1, or VAR(1):

$$\underbrace{X_t}_{(p \times 1)} = \underbrace{F}_{(p \times p)} \underbrace{X_{t-1}}_{(p \times 1)} + \underbrace{E_t}_{(p \times 1)} .$$

AR(p): Stability / Causality

We say that $\{x_t\}$ is **stable / causal** if

- the roots of $\alpha(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p = 0$ lie outside the unit circle; OR
- the roots of $z^p - \alpha_1 z^{p-1} - \ldots - \alpha_p = 0$ lie inside the unit circle; OR
- the eigenvalues of the companion matrix F lie inside the unit circle.

In this case, we can write $x_t - \mu = \alpha(L)^{-1} e_t = \psi(L) e_t$ for some $\psi(L) = 1 + \psi_1 L + \ldots$ such that $\sum |\psi_j| < \infty$ (absolute summability). Note that $\psi_j \to 0$ as $j \to \infty$.

Example: $x_t = 0.6 x_{t-1} - 0.08 x_{t-2} + e_t$. Roots of $\alpha(z) = 1 - 0.6z + 0.08z^2$ are $z = 2.5$ and 5, both larger than 1.

$$F = \begin{bmatrix} 0.6 & -0.08 \\ 1 & 0 \end{bmatrix}$$

The eigenvalues are the λs that solve $(-0.6 + \lambda)(\lambda) + 0.08 = 0$, giving λ of 0.4 and 0.2. These are the reciprocals of the roots of $\alpha(z) = 0$. $\alpha(L) = (1 - 0.4L)(1 - 0.2L) \Rightarrow \alpha(L)^{-1} = (1 - 0.4L)^{-1}(1 - 0.2L)^{-1} = (1 + 0.4L + \ldots)(1 + 0.2L + \ldots) = 1 + 0.6L + \ldots$

Why stable/causal and invertible ARMA processes?

Let $\{x_t\}$ be an ARMA process such that

$$\alpha(L)x_t = \theta(L)e_t, \quad \{e_t\} \sim WN(0, \sigma^2),$$

where $\alpha(z)$ and $\theta(z)$ do not have a unit root, i.e., $\alpha(z) \neq 0$ and $\theta(z) \neq 0$ for all $|z| = 1$.

What if this process is either

non-causal (i.e., $\exists |z| < 1$ s.t. $\alpha(z) = 0$) or
non-invertible (i.e., $\exists |z| < 1$ s.t. $\theta(z) = 0$)?

Then, there exist polynomials $\tilde{\alpha}(z)$ and $\tilde{\theta}(z)$ of order p and q, and a white noise process $\{\tilde{e}_t\} \sim WN(0, \tilde{\sigma}^2)$ such that $\tilde{\alpha} \neq 0$ and $\tilde{\theta} \neq 0$ for all $|z| \leq 1$ and

$$\tilde{\alpha}(L)x_t = \tilde{\theta}(L)\tilde{e}_t.$$

Ignoring unit root processes, we can focus on causal, invertible ARMA processes.

How to compute the impulse responses? I

Let $\{x_t\}$ be a stable/causal invertible ARMA process such that

$$\alpha(L)x_t = \theta(L)e_t, \quad \{e_t\} \sim WN(0, \sigma^2).$$

How to compute the impulse response function $\{\psi_j\}$ such that

$$x_t = \psi(L)e_t = \sum_{j=0}^{\infty} \psi_j e_{t-j} = \psi_0 e_t + \psi_1 e_{t-1} + \dots$$

given $\alpha(L)$ and $\theta(L)$?

Note that

$$\alpha(L)^{-1}\theta(L) = \psi(L) \quad \Rightarrow \quad \theta(L) = \alpha(L)\psi(L).$$

Thus, we can obtain ψ_j by comparing coefficients on L^j on both sides.

How to compute the impulse responses? II

Example: ARMA(1,2).

$$x_t = \alpha_1 x_{t-1} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}.$$

Note that

$$x_t = \alpha(L)^{-1}\theta(L)e_t = \psi(L)e_t.$$

Thus,

$$\begin{aligned}
1 + \theta_1 L + \theta_2 L^2 &= (1 - \alpha_1 L)(\psi_0 + \psi_1 L + \psi_2 L^2 + \ldots) \\
&\Rightarrow 1 = \psi_0 \\
&\quad \theta_1 = \psi_1 - \alpha_1 \Rightarrow \psi_1 = \alpha_1 + \theta_1 \\
&\quad \theta_2 = \psi_2 - \alpha_1\psi_1 \Rightarrow \psi_2 = \alpha_1\psi_1 + \theta_2 \\
&\quad \vdots
\end{aligned}$$

ARMA arithmetic

Consider two AR(1) processes:

$$x_t = \alpha_1 x_{t-1} + u_t,$$

$$y_t = \alpha_2 y_{t-1} + v_t,$$

where u_t and v_t are uncorrelated white noise innovations.

We want to show $z_t = x_t + y_t$ is ARMA(2,1).

$$\begin{aligned} (1-\alpha_1 L)x_t = u_t &\quad\Rightarrow\quad (1-\alpha_1 L)(1-\alpha_2 L)x_t = (1-\alpha_2 L)u_t \\ (1-\alpha_2 L)y_t = v_t &\quad\Rightarrow\quad (1-\alpha_1 L)(1-\alpha_2 L)y_t = (1-\alpha_1 L)v_t \\ \underbrace{(1-\alpha_1 L)(1-\alpha_2 L)}_{AR(2)}\underbrace{(x_t + y_t)}_{z_t} &\quad=\quad \underbrace{(1-\alpha_2 L)u_t + (1-\alpha_1 L)v_t}_{w_t} \end{aligned}$$

ARMA arithmetic II

Next, we need to show $\{w_t = (1 - \alpha_2 L)u_t + (1 - \alpha_1 L)v_t\}$ is MA(1). It is sufficient to show that $\gamma_w(k) = 0$ for all $k > 1$.

$$\begin{array}{rcl} w_t & = & (1 - \alpha_2 L)u_t + (1 - \alpha_1 L)v_t \\ \gamma_w(0) & = & (1 + \alpha_2^2)\sigma_u^2 + (1 + \alpha_1^2)\sigma_v^2 \\ \gamma_w(1) & = & -\alpha_2 \sigma_u^2 - \alpha_1 \sigma_v^2 \\ \gamma_w(k) & = & 0, k > 1. \end{array}$$

Then, how can we actually compute θ for $w_t = e_t + \theta e_{t-1}$ and $var(e_t)$? In principle, we can use the autocovariances.

$$\frac{\theta}{1 + \theta^2} = \rho_w(1) = \frac{\gamma_w(1)}{\gamma_w(0)} = \frac{-\alpha_2 \sigma_u^2 - \alpha_1 \sigma_v^2}{(1 + \alpha_2^2)\sigma_u^2 + (1 + \alpha_1^2)\sigma_v^2}.$$

For a more general situation, see Hamilton (1994, pp. 391-4).

Autocovariance (Generating) Function (ACGF)

Definition: $G(z) = \sum_{k=-\infty}^{\infty} \gamma(k) z^k$ for $z \in \mathbb{C}$.

Suppose that $x_t = \psi(L) e_t$, $e_t \sim WN(0, \sigma^2)$.

$$\sum |\psi_j| < \infty \quad \Rightarrow \quad \begin{cases} \sum |\gamma(k)| < \infty \\ G(z) \text{ is well-defined for } r^{-1} < |z| < r \\ \qquad \text{for some } r > 1 \\ G(z) = \sigma^2 \psi(z) \psi(z^{-1}) \end{cases}$$

Example: Consider MA(1) process

$$x_t = (1+\theta L)e_t \Rightarrow G(z) = \sigma^2(1+\theta z)(1+\theta z^{-1}) = \sigma^2(\theta z + 1 + \theta^2 + \theta z^{-1}),$$

implying that $\gamma(0) = (1+\theta^2)\sigma^2$, $\gamma(1) = \theta\sigma^2$, and $\gamma(k) = 0$ for all $k > 1$.

(Linear Time-invariant) Filters

Suppose that h_j is a sequence of real numbers that is absolutely summable, $\sum_{j=-\infty}^{\infty} |h_j| < \infty$.

Original series: $\{x_t\}$
Filtered series: $\{y_t\}$, where

$$y_t = h(L)x_t = \sum_{j=-\infty}^{\infty} h_j x_{t-j}.$$

ACGF of the filtered series:

$$G_y(z) = h(z)h(z^{-1})G_x(z).$$

Example: Consider $\alpha(L)y_t = \theta(L)e_t$, $x_t = e_t$, and $h(z) = \alpha^{-1}(z)\theta(z)$. Then, the ACGF of an ARMA process $\{y_t\}$ is

$$G_y(z) = \sigma^2 \frac{\theta(z)\theta(z^{-1})}{\alpha(z)\alpha(z^{-1})}.$$

Estimation of ARIMA(p,d,q): Overview

Suppose that you want to describe the dynamics of real GDP using an ARMA model. What would you do?

1. (d) It is evident that log(RGDP), y_t, is not stationary. Thus, to use ARMA models, you need to difference the series until it becomes stationary. Usually, differencing once is enough ($d = 1$):

$$x_t = \Delta y_t = y_t - y_{t-1}.$$

2. (p and q) You need to choose p and q. For the purpose, you may (1) investigate autocorrelations and partial autocorrelations, (2) rely on information criteria, and (3) use statistical tests.
3. (Estimation of ARMA) Once a stationary time series is ready and p and q are selected, you need to estimate an ARMA(p,q) model. You can use (1) OLS (for AR models), (2) MoM or GMM, (3) Hannan-Rissanen algorithm, and (4) MLE (Kalman filter).
4. After fitting your model to the data, various diagnostic checks can be done. ex) Does $\hat{e}_t$ look like a white noise?

Model Selection I

Principle of parsimony (Box-Jenkins): Use the smallest number of parameters for adequate representation.

Method 1: We can investigate the ACF and PACF to determine appropriate p and q.

Autocorrelation Function (ACF).

$$\rho(k) = corr(x_{t+k}, x_t) = \gamma(k)/\gamma(0).$$

Partial Autocorrelation Function (PACF).

$$\phi(k) = \begin{cases} corr(x_{t+k} - \hat{x}_{t+k}, x_t - \hat{x}_t) & k > 1 \\ corr(x_{t+1}, x_t) & k = 1 \end{cases},$$

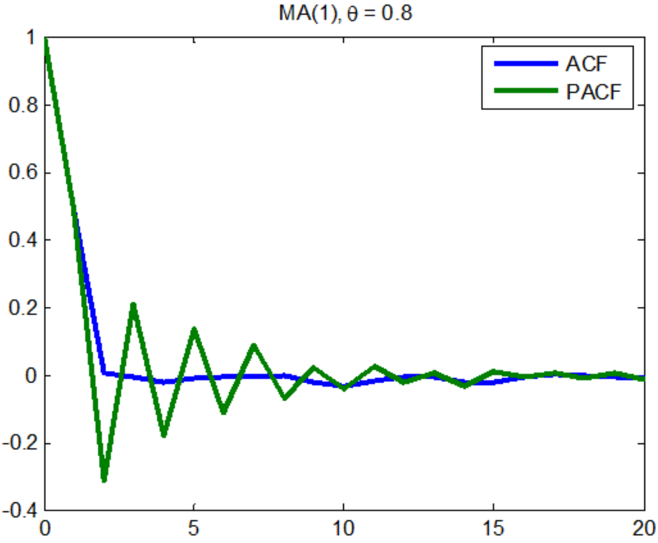
where $\hat{x}_{t+k} = P(x_{t+k} | x_{t+k-1}, \dots, x_{t+1})$ and $\hat{x}_t = P(x_t | x_{t+k-1}, \dots, x_{t+1})$

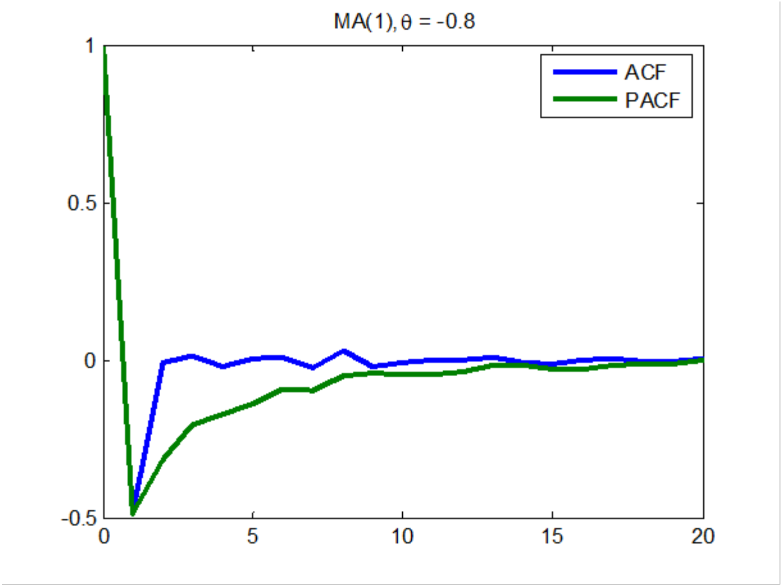
It is known that

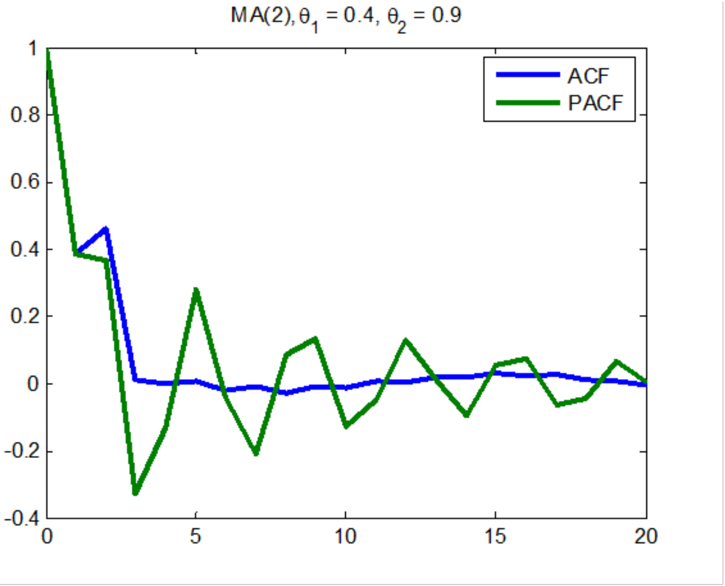
$$x_{t+k} = c + \beta_{1,k} x_{t+k-1} + \cdots + \beta_{k-1,k} x_{t+1} + \phi(k) x_t + error.$$

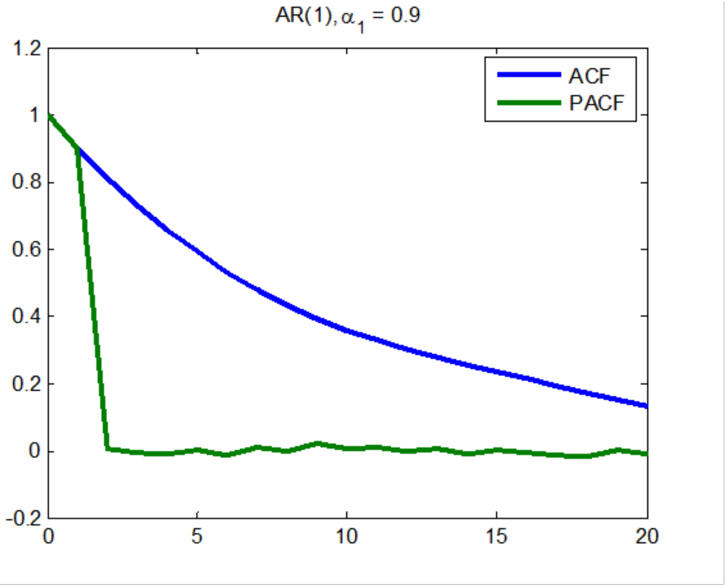
Model Selection II

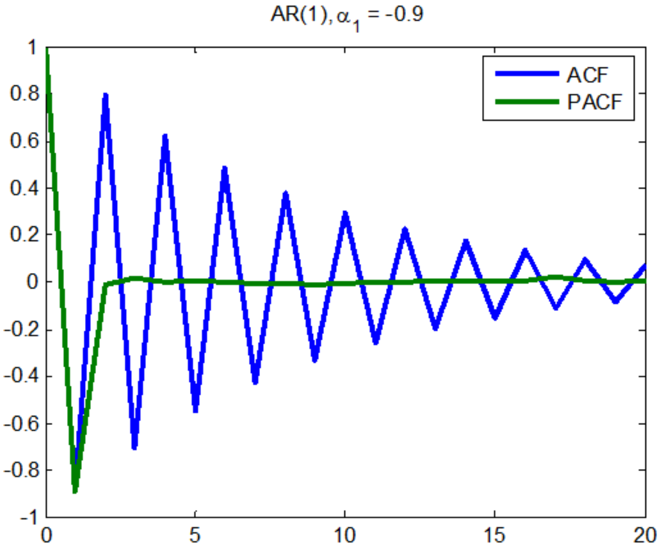
	$AR(p)$	$MA(q)$
$\rho(k)$	Dies off slowly	Zero at lag $k \geqslant q+1$
$\phi(k)$	Zero at lag $k \geqslant p+1$	Dies off slowly

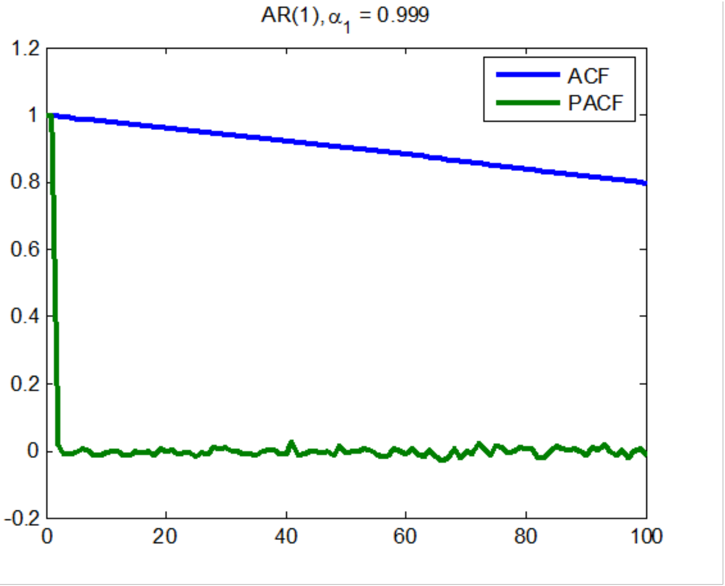

MA(1), θ = 0.8
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20

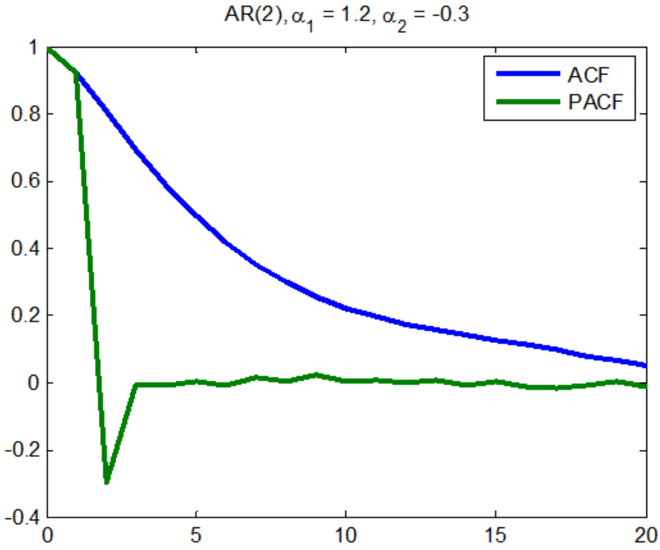

MA(1), θ = -0.8
1
0.5
0
-0.5
0
5
10
15
20
ACF
PACF

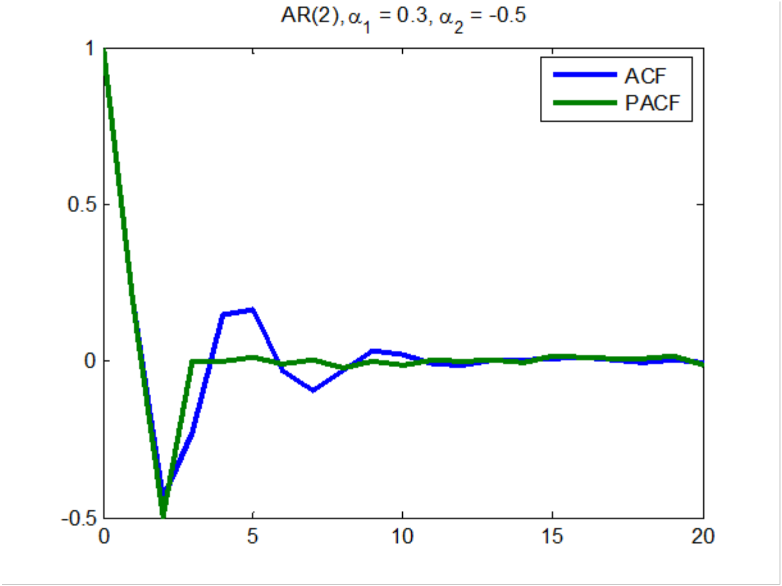

MA(2), θ1 = 0.4, θ2 = 0.9
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20

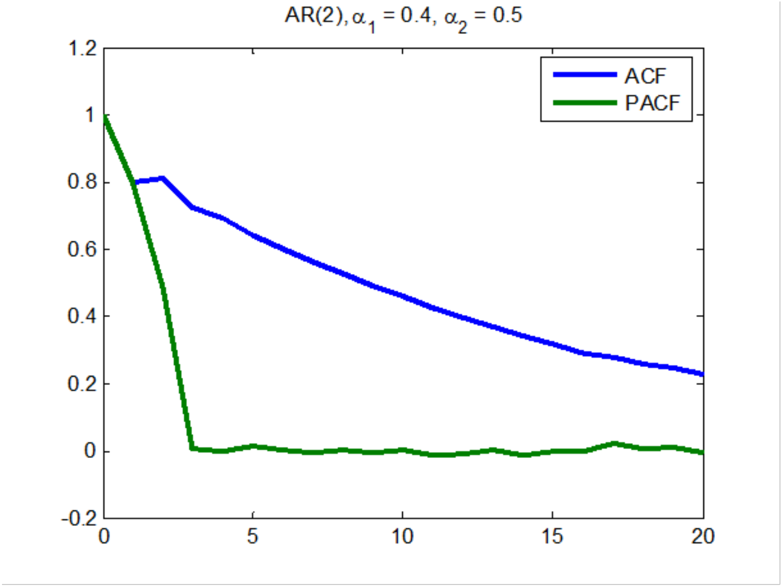

AR(1), α_1 = 0.9
ACF
PACF
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
5
10
15
20

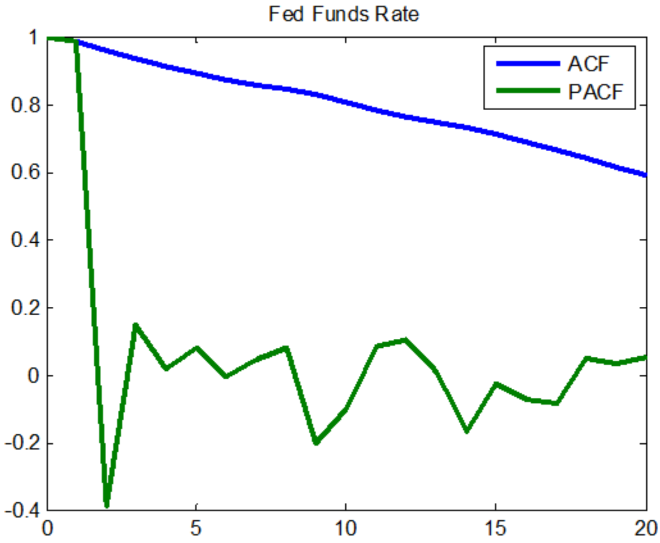

AR(1), α_1 = -0.9
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
5
10
15
20

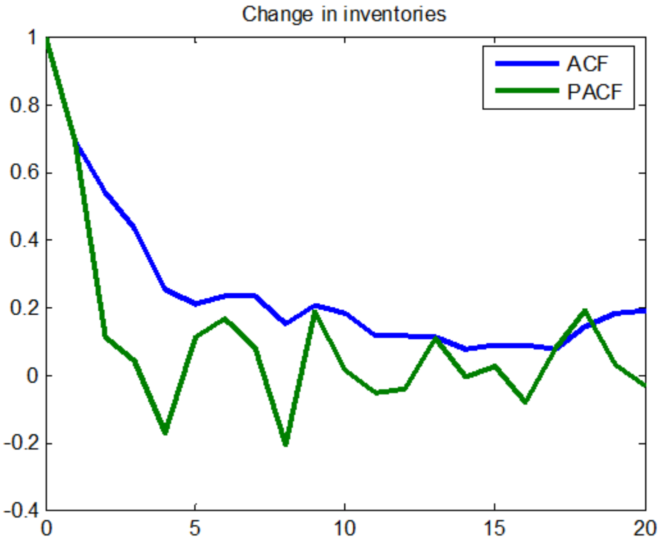

AR(1), α_1 = 0.999
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
ACF
PACF


AR(2), α_1 = 1.2, α_2 = -0.3
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20


AR(2), α_1 = 0.3, α_2 = -0.5
ACF
PACF
1
0.5
0
-0.5
0
5
10
15
20


AR(2), $\alpha_1 = 0.4$, $\alpha_2 = 0.5$
ACF
PACF
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
5
10
15
20


Fed Funds Rate
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20


Change in inventories
ACF
PACF
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20

Model Selection III

Information criteria

Let n be the number of parameters. ex) ARMA(p,q) $\Rightarrow n = p + q$.

$$\underset{n}{\operatorname{argmax}} \quad \log(\hat{L}) - C(n, T),$$

where $\hat{L}$ is the maximized likelihood at the MLE.

The first term captures the fit of the model.
The second term penalizes over-parametrization.
We estimate different models, compare the IC, and select one that maximizes the IC.

Widely used penalty terms: $C(n, T) = \begin{cases} n & AIC \\ n\log(\log(T)) & HQIC \\ \frac{1}{2}n\log(T) & BIC \end{cases}$

Model Selection IV

When $e_t \sim iid\mathcal{N}(0, \sigma^2)$, we can equivalently select a model by minimizing the following:

$$\begin{aligned} AIC &= \log \hat{\sigma}^2 + 2n/T \\ HQIC &= \log \hat{\sigma}^2 + 2n \log(\log(T))/T \\ BIC &= \log \hat{\sigma}^2 + n \log(T)/T \end{aligned}$$

In theory, asymptotically, the BIC and HQIC consistently select the true model (if it is included in the set of candidate models).

In practice, with reasonable sample sizes, the AIC tends to over-parametrize, while the BIC tends to under-parametrize.

Other approaches based on sequential testing:

- General-to-specific testing.
- Specific-to-general testing.

IC: Remarks

Suppose that we observe $\{x_1, \ldots, x_{50}\}$. We want to select an AR(p) model subject to $1 \leq p \leq 5$.
Consider the case where $p = 5$ ($T = 45$, $n = 5$).

$$x_6 = c + \alpha_1 x_5 + \ldots \alpha_5 x_1 + e_6,$$

$$\vdots$$

$$x_{50} = c + \alpha_1 x_{49} + \cdots + \alpha_5 x_{45} + e_{50}.$$

We effectively have 45 observations due to initialization.
In contrast, when $p = 1$, we have 49 observations.

$$x_2 = c + \alpha_1 x_1 + e_2,$$

$$\vdots$$

$$x_{50} = c + \alpha_1 x_{49} + e_{50}.$$

To make the comparison fair, when evaluating the IC, we need to use only $T = 45$ observations. That is, we should start from $x_6 = c + \alpha_1 x_5 + e_6$.

Once p is selected (and $p < 5$), we can re-estimate the model by using all observations.

Estimation of AR Models

Yule-Walker Equations (MoM).
For an AR(p) process, $\sum_{k=1}^{p} \gamma(k-i)\alpha_k = \gamma(i)$, $i = 1, \ldots, p$.
In matrix form:

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \vdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \vdots & \gamma(p-2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma(p-1) & \gamma(p-2) & \gamma(p-3) & \vdots & \gamma(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \cdots \\ \gamma(p) \end{bmatrix}$$

Conditional MLE = OLS.
Given the first p observations, both the conditional MLE and OLS minimize

$$S(\alpha) = \sum_{t=p+1}^{T} (x_t - \alpha_1 x_{t-1} - \ldots - \alpha_p x_{t-p})^2$$

when $e_t \sim iidN(0, \sigma^2)$. In both cases, use standard asymptotic inference.

Estimation of MA Models

Consider MA(1) process: $x_t = e_t + \theta e_{t-1}$. Note that neither e_t nor e_{t-1} is directly observable.

- **MoM**: match moments: $\gamma_0 = (1+\theta^2)\sigma^2$ and $\gamma_1 = \theta\sigma^2$
- **MLE**: Assuming e_t is normal, the conditional (ignoring initial conditions, i.e., e_0) log likelihood is

  $$\log L \propto -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T} e_t^2$$

  Given e_0 and θ, we can back out the series of e_t from $\{x_t\}$. Then, we can compute $L(\theta)$. We can (1) assume that $e_0 = 0$ or (2) consider an unconditional likelihood as a function of both e_0 and θ.
- **Kalman filter**: to be studied.

Practical Issues

- Starting values matter (a lot!). Need to check for a global optimum (for MLE).
- MA component requires use of non-linear estimators (notoriously difficult for high order MA processes).
- Keep only invertible MA solutions (constrain admissible parameter space to $(-1, 1)$ for MA(1) processes).
- Common factor problem: Suppose $(1 - \alpha_1 L)(1 - \alpha_2 L)x_t = (1 - \theta L)e_t$ and $\alpha_2 \approx \theta$. Then the model is over-parameterized because it could be equally described by a more parsimonious process, $(1 - \alpha_1 L)x_t = e_t$. In this case, the parameter estimates for less parsimonious model are highly unstable (sensitive to starting values).

Hannan-Rissanen procedure for ARMA(p,q)

Want: estimate $\alpha(L)x_t = \theta(L)e_t$.
Assumption: $\theta(L)$ is invertible. Thus, $\pi(L)x_t = e_t$, where $\pi(L) = \alpha(L)\theta(L)^{-1}$.

1. Fit a long autoregression of order M,
   $x_t = \hat{\pi}_1 x_{t-1} + \ldots + \hat{\pi}_M x_{t-M} + \hat{e}_t$.
2. Construct the regression:
   $$x_t - \hat{e}_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + \theta_1 \hat{e}_{t-1} + \ldots + \theta_q \hat{e}_{t-q} + u_t.$$
   Use OLS. The estimates are consistent and asymptotically normal for appropriate choice of M, p and q.

Main advantage: computationally fast and simple. One may want to use this algorithm to evaluate performance of some ARMA models before trying more complicated, time-consuming estimators.

Forecasting

Want: forecast y_{t+s} using x_t available at time t.
Forecast: $y_{t+s|t}$
Need: a loss function to evaluate the quality of the forecast

The **MSE criterion** (a quadratic loss function) is widely used:

$$mse = \mathbb{E}(y_{t+s} - y_{t+s|t})^2 = \text{ forecast error variance.}$$

Two approaches:

1. Regress y_{t+s} on x_t for each horizon s. Then, use x_T to forecast y_{T+s}.
2. Assume a model, e.g., AR(p). Estimate the model. Conditioning on x_T, x_{T-1}, $\ldots$, iterate the estimated model forward.

Example of the First Approach

Suppose $e_t \sim WN(0, \sigma^2)$ and $y_t - \mu = \psi(L)e_t$. Then,

$$y_{t+s} - \mu = e_{t+s} + \psi_1 e_{t+s-1} + \ldots + \psi_{s-1} e_{t+1} + \psi_s e_t + \psi_{s+1} e_{t-1} + \ldots$$

If we know the model parameters and can observe $\{e_k\}_{k=-\infty}^{t}$,

$$\hat{y}_{t+s} | e_t, e_{t-1}, \ldots = \mu + \psi_s e_t + \psi_{s+1} e_{t-1} + \ldots.$$

The forecast error is

$$y_{t+s} - \hat{y}_{t+s|t} = e_{t+s} + \psi_1 e_{t+s-1} + \ldots + \psi_{s-1} e_{t+1}.$$

Note that the forecast error is mean zero and uncorrelated with e_t, e_{t-1}, $\ldots$. Furthermore, the forecast error variance is given by

$$var(y_{t+s} - \hat{y}_{t+s|t}) = \mathbb{E}(y_{t+s} - \hat{y}_{t+s|t})^2 = (1 + \psi_1^2 + \ldots + \psi_{s-1}^2)\sigma_e^2$$

Example of the Second Approach

The univariate AR(p) process $x_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + e_t$ admits the following companion form:

$$X_t = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-p+1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_p \\ 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-p} \end{bmatrix} + \begin{bmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} = FX_{t-1} + E_t.$$

Note that

$$\begin{aligned} X_{t+s} &= FX_{t+s-1} + E_{t+s} = F^2 X_{t+s-2} + E_{t+s} + FE_{t+s-1} \\ &= F^s X_t + \sum_{j=0}^{s-1} F^j E_{t+s-j}. \end{aligned}$$

Clearly, $X_{t+s|t} = F^s X_t$ and the forecast error variance $var(X_{t+s} - X_{t+s|t}) = var(\sum_{j=0}^{s-1} F^j E_{t+s-j}) = \sum_{j=0}^{s-1} F^j \Sigma (F^j)'$, where $\Sigma = diag(\sigma_e^2, 0, 0, \ldots, 0)$.

Forecast Comparison

Question. Which model is better at forecasting?
Example: Can a structural model beat a random walk model in forecasting exchange rate?

Diebold and Mariano (1995) proposed a popular test for equality in forecast precision. We have two *non-nested* models: A and B.

- Let g be the loss function. E.g., if $g(\hat{e}_t) = \hat{e}_t^2$, then $\frac{1}{T}\sum_{t=1}^{T} g(\hat{e}_t)$ is the sample MSE.
- Define the loss differential $d_t = g(\hat{e}_t^A) - g(\hat{e}_t^B)$. Form

$$H0 : \mathbb{E}(d_t) = 0 \text{ and } H_1 : \mathbb{E}(d_t) \neq 0$$

- Let $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$. Then $\sqrt{T}(\bar{d} - \mu) \xrightarrow{d} N(0, LRV(d_t))$ where $LRV(d_t) = T \lim var(\sqrt{T}\bar{d}) = \sum_{j=-\infty}^{\infty} \gamma_d(j)$.
- We can consider the test statistic $S_T = \bar{d}/\sqrt{\frac{1}{T} avar(\bar{d})}$ which distributed as a t-statistic with $T - 1$ degrees of freedom.

Mark (AER, 1995)

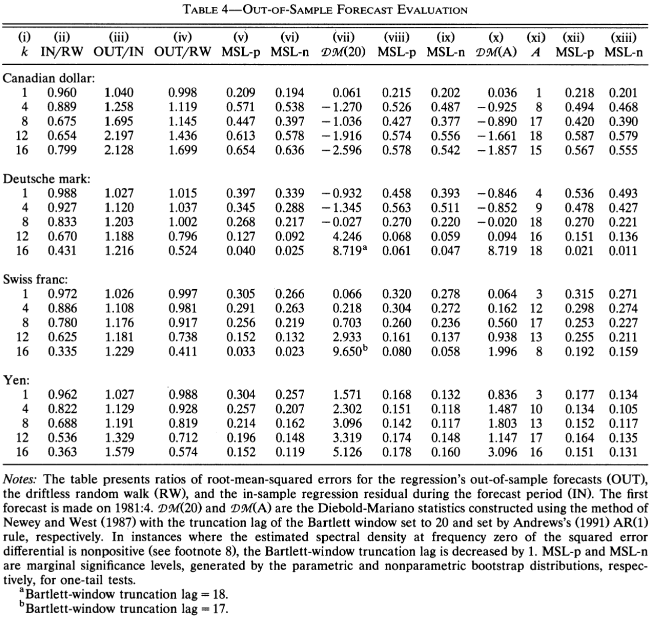


TABLE 4—OUT-OF-SAMPLE FORECAST EVALUATION

(i) k	(ii) IN/RW	(iii) OUT/IN	(iv) OUT/RW	(v) MSL-p	(vi) MSL-n	(vii) $\mathcal{DM}(20)$	(viii) MSL-p	(ix) MSL-n	(x) $\mathcal{DM}(A)$	(xi) A	(xii) MSL-p	(xiii) MSL-n
Canadian dollar:												
1	0.960	1.040	0.998	0.209	0.194	0.061	0.215	0.202	0.036	1	0.218	0.201
4	0.889	1.258	1.119	0.571	0.538	−1.270	0.526	0.487	−0.925	8	0.494	0.468
8	0.675	1.695	1.145	0.447	0.397	−1.036	0.427	0.377	−0.890	17	0.420	0.390
12	0.654	2.197	1.436	0.613	0.578	−1.916	0.574	0.556	−1.661	18	0.587	0.579
16	0.799	2.128	1.699	0.654	0.636	−2.596	0.578	0.542	−1.857	15	0.567	0.555
Deutsche mark:												
1	0.988	1.027	1.015	0.397	0.339	−0.932	0.458	0.393	−0.846	4	0.536	0.493
4	0.927	1.120	1.037	0.345	0.288	−1.345	0.563	0.511	−0.852	9	0.478	0.427
8	0.833	1.203	1.002	0.268	0.217	−0.027	0.270	0.220	−0.020	18	0.270	0.221
12	0.670	1.188	0.796	0.127	0.092	4.246	0.068	0.059	0.094	16	0.151	0.136
16	0.431	1.216	0.524	0.040	0.025	8.719[a]	0.061	0.047	8.719	18	0.021	0.011
Swiss franc:												
1	0.972	1.026	0.997	0.305	0.266	0.066	0.320	0.278	0.064	3	0.315	0.271
4	0.886	1.108	0.981	0.291	0.263	0.218	0.304	0.272	0.162	12	0.298	0.274
8	0.780	1.176	0.917	0.256	0.219	0.703	0.260	0.236	0.560	17	0.253	0.227
12	0.625	1.181	0.738	0.152	0.132	2.933	0.161	0.137	0.938	13	0.255	0.211
16	0.335	1.229	0.411	0.033	0.023	9.650[b]	0.080	0.058	1.996	8	0.192	0.159
Yen:												
1	0.962	1.027	0.988	0.304	0.257	1.571	0.168	0.132	0.836	3	0.177	0.134
4	0.822	1.129	0.928	0.257	0.207	2.302	0.151	0.118	1.487	10	0.134	0.105
8	0.688	1.191	0.819	0.214	0.162	3.096	0.142	0.117	1.803	13	0.152	0.117
12	0.536	1.329	0.712	0.196	0.148	3.319	0.174	0.148	1.147	17	0.164	0.135
16	0.363	1.579	0.574	0.152	0.119	5.126	0.178	0.160	3.096	16	0.151	0.131

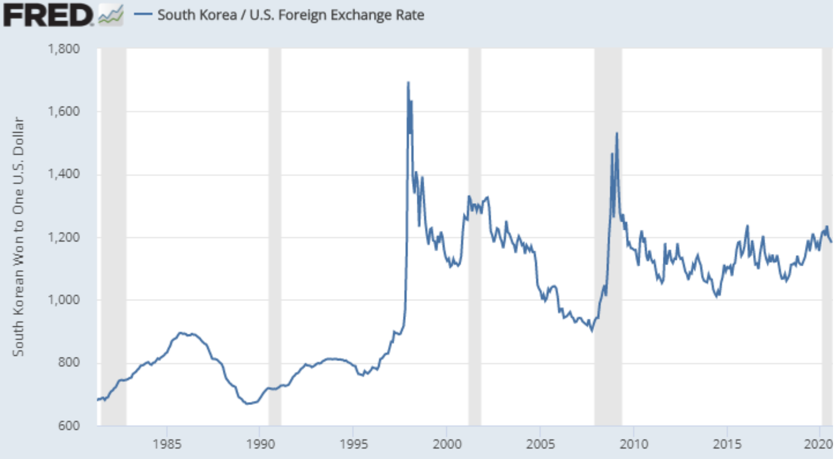
Notes: The table presents ratios of root-mean-squared errors for the regression's out-of-sample forecasts (OUT), the driftless random walk (RW), and the in-sample regression residual during the forecast period (IN). The first forecast is made on 1981:4. $\mathcal{DM}(20)$ and $\mathcal{DM}(A)$ are the Diebold-Mariano statistics constructed using the method of Newey and West (1987) with the truncation lag of the Bartlett window set to 20 and set by Andrews's (1991) AR(1) rule, respectively. In instances where the estimated spectral density at frequency zero of the squared error differential is nonpositive (see footnote 8), the Bartlett-window truncation lag is decreased by 1. MSL-p and MSL-n are marginal significance levels, generated by the parametric and nonparametric bootstrap distributions, respectively, for one-tail tests.

[a] Bartlett-window truncation lag = 18.

[b] Bartlett-window truncation lag = 17.

Conditional Heteroskedasticity

A Motivating Example



The variance of the innovations might be time-varying.

Autoregressive Conditional Heteroskedasticity (ARCH)

ARMA model:

$$\alpha(L)x_t = \theta(L)e_t, \quad e_t \sim WN(0, \sigma^2)$$

Covariance stationarity: $\mathbb{E}\left[e_t^2\right] = \sigma^2$, which is constant.
How about $\mathbb{E}_{t-1}\left[e_t^2\right]$?

ARCH: $e_t \sim ARCH(m)$

$$e_t^2 = \zeta + \xi_1 e_{t-1}^2 + \cdots + \xi_m e_{t-m}^2 + w_t, \quad w_t \sim WN(0, \lambda^2)$$

Restrictions needed: $\zeta > 0$, $\xi_i \geq 0$, $w_t > -\zeta$, $\xi_1 + \cdots + \xi_m < 1$.

A Convenient Alternative Formulation:

$$e_t = \sqrt{h_t} v_t, \text{ where}$$
$$h_t = \zeta + \xi_1 e_{t-1}^2 + \cdots + \xi_m e_{t-m}^2 \quad \text{and} \quad v_t \sim WN(0, 1)$$

Restrictions needed: $\zeta > 0$, $\xi_i \geq 0$, $\xi_1 + \cdots + \xi_m < 1$.

Generalized ARCH (GARCH)

ARMA model:

$$\alpha(L)x_t = \theta(L)e_t, \quad e_t \sim WN(0, \sigma^2)$$

GARCH: $e_t \sim GARCH(r, m)$

$$e_t = \sqrt{h_t}v_t, \text{ where } \quad v_t \sim WN(0,1) \quad \text{and}$$
$$h_t = \zeta + \delta_1 h_{t-1} + \cdots + \delta_r h_{t-r} + \xi_1 e_{t-1}^2 + \cdots + \xi_m e_{t-m}^2$$

Restrictions needed: $\zeta > 0$, $\delta_i \geq 0$, $\xi_i \geq 0$,
$\delta_1 + \cdots + \delta_r + \xi_1 + \cdots + \xi_m < 1$.

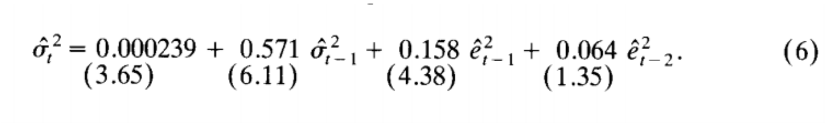
Many variants: Integrated GARCH, ARCH-in-Mean, EGARCH, TGARCH, etc.

At quarterly or lower frequencies, conditional heteroskedasticity is not that salient. It matters more at high frequencies (e.g., daily financial data).

Pagan and Schwert (JoEconometrics, 1990)

$$y_t = x_t'\beta + u_t$$

- y_t: monthly stock return (1835 - 1925)
- x_t: monthly dummy
- u_t: $u_t = e_t + \theta e_{t-1}$
- $\hat{e}_t$ is obtained as a residual for $\hat{u}_t \sim AR(10)$.



$$\hat{\sigma}_t^2 = \underset{(3.65)}{0.000239} + \underset{(6.11)}{0.571}\,\hat{\sigma}_{t-1}^2 + \underset{(4.38)}{0.158}\,\hat{e}_{t-1}^2 + \underset{(1.35)}{0.064}\,\hat{e}_{t-2}^2. \tag{6}$$

- t-stat in parentheses.
- They find $\hat{\theta}$ is insignificant.