1 ARMA Representations

We work with deviations from the mean. For example, an AR(1) with mean $\mu = E(y_t)$ can be written as

$$y_t - \mu = \phi(y_{t-1} - \mu) + \varepsilon_t.$$

Our examples use $\mu = 0$. The shocks ε_t are white noise: they have mean zero, variance σ^2, and are uncorrelated across dates. We use σ^2 to calculate forecast uncertainty.

1.1 Stationarity and invertibility

Stationarity concerns the statistical properties of y_t. **Invertibility** concerns recovering the shocks ε_t from observations of y_t. Both concepts apply to AR, MA, and ARMA models.

Stationarity: three conditions

A series is **covariance-stationary** if:

1. **Constant mean:** $E(y_t) = \mu$ at every date.
2. **Constant, finite variance:** $\text{Var}(y_t) = \gamma_0 < \infty$ at every date.
3. **Autocovariance depends only on the lag:**

$$\text{Cov}(y_t, y_{t-k}) = \gamma_k.$$

Autocovariance measures how observations of the same series move together. Pairs separated by the same number of periods have the same covariance; for example,

$$\text{Cov}(y_{10}, y_9) = \text{Cov}(y_{20}, y_{19}).$$

Stationarity does not require y_t itself to stay constant.

Example: AR(1) written as an infinite MA

Consider $y_t = \phi y_{t-1} + \varepsilon_t$. When $\boxed{|\phi| < 1}$, use $Ly_t = y_{t-1}$ and the geometric expansion:

$$\begin{aligned}(1 - \phi L)y_t &= \varepsilon_t, \\ y_t &= \frac{1}{1 - \phi L}\varepsilon_t = (1 + \phi L + \phi^2 L^2 + \cdots)\varepsilon_t \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots .\end{aligned}$$

This **pure MA or Wold representation** writes y_t as a weighted sum of current and past shocks.

Using the white-noise properties,

$$\begin{aligned}E(y_t) &= 0, \\ \text{Var}(y_t) &= \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma^2}{1 - \phi^2} < \infty, \\ \text{Cov}(y_t, y_{t-k}) &= \sigma^2 \sum_{j=0}^{\infty} \phi^{j+k}\phi^j = \frac{\phi^k \sigma^2}{1 - \phi^2}, \qquad k \geq 0.\end{aligned}$$

The mean and variance are constant, and the covariance depends only on k. Thus, this solution is stationary.

Intuition: a shock's effect after j periods is ϕ^j. When $|\phi| < 1$, old shocks die out. For $\phi = 0.5$, the effects are $1, 0.5, 0.25, \ldots$.

Invertibility: recovering shocks from observations

We observe $y_t, y_{t-1}, \ldots$, but not the shocks directly. A representation is **invertible** if we can recover ε_t as a stable linear combination of these observations. For an infinite sum, including more past observations makes the weighted sum approach ε_t.

For the MA(1),

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1} = (1 - \theta L)\varepsilon_t,$$

the condition is $\boxed{|\theta| < 1}$. Then

$$\varepsilon_t = \frac{1}{1 - \theta L} y_t = y_t + \theta y_{t-1} + \theta^2 y_{t-2} + \cdots .$$

Rearranging gives the **pure AR representation**:

$$y_t = -\theta y_{t-1} - \theta^2 y_{t-2} - \cdots + \varepsilon_t.$$

For $\theta = 0.5$, the inverse weights are $1, 0.5, 0.25, \ldots$: increasingly old observations receive smaller weights.

Neither property implies the other

Stationary but not invertible:

$$y_t = \varepsilon_t - 2\varepsilon_{t-1}.$$

Here $E(y_t) = 0$, $\mathrm{Var}(y_t) = 5\sigma^2$, $\gamma_1 = -2\sigma^2$, and $\gamma_k = 0$ for $k \geq 2$. These satisfy stationarity. But this shock representation is not invertible: the inverse weights $1, 2, 4, \ldots$ grow instead of decay.

Invertible but not stationary:

$$y_t = y_{t-1} + \varepsilon_t, \qquad y_0 = 0.$$

We recover the shock exactly as $\varepsilon_t = y_t - y_{t-1}$. But $\mathrm{Var}(y_t) = t\sigma^2$ increases over time, so this random walk is not stationary.

Common-factor cancellation

An ARMA equation can simplify to a lower-order model. For example,

$$\begin{aligned} y_t &= 0.5y_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}, \\ (1 - 0.5L)y_t &= (1 - 0.5L)\varepsilon_t. \end{aligned}$$

For the stationary solution, the common factor cancels, leaving $y_t = \varepsilon_t$: white noise.

1.2 Wold representation and forecast errors

For our stationary, invertible ARMA models, write the Wold representation as

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad \psi_0 = 1.$$

The **Wold weight** ψ_j is the effect of a one-unit shock after j periods. To find the weights, solve for y_t in lag notation and expand.

For example, the ARMA(1,1)

$$y_t = 0.6y_{t-1} + \varepsilon_t - 0.2\varepsilon_{t-1}$$

has

$$\begin{aligned}(1-0.6L)y_t &= (1-0.2L)\varepsilon_t,\\ y_t &= \frac{1-0.2L}{1-0.6L}\varepsilon_t\\ &= (1+0.6L+0.36L^2+\cdots)(1-0.2L)\varepsilon_t\\ &= \varepsilon_t + 0.4\varepsilon_{t-1} + 0.24\varepsilon_{t-2} + \cdots .\end{aligned}$$

Thus $\psi_1 = 0.6 - 0.2 = 0.4$ and $\psi_2 = 0.36 - 0.2(0.6) = 0.24$.

Forecast errors contain only future shocks

Let T be the last observed period and h the forecast horizon. Treat parameters and past shocks as known, and assume future shocks have conditional mean zero.

Separate the Wold representation into

$$y_{T+h} = \underbrace{\sum_{j=0}^{h-1}\psi_j\varepsilon_{T+h-j}}_{\text{future shocks: unknown at } T} + \underbrace{\sum_{j=h}^{\infty}\psi_j\varepsilon_{T+h-j}}_{\text{past and current shocks: known at } T}.$$

The forecast keeps the known terms:

$$\widehat{y}_{T+h|T} = \sum_{j=h}^{\infty}\psi_j\varepsilon_{T+h-j}.$$

Subtracting gives

$$\begin{aligned}e_{T+h|T} &= y_{T+h} - \widehat{y}_{T+h|T}\\ &= \varepsilon_{T+h} + \psi_1\varepsilon_{T+h-1} + \cdots + \psi_{h-1}\varepsilon_{T+1}.\end{aligned}$$

Since shocks are uncorrelated,

$$\boxed{v_h^2 = \mathrm{Var}(e_{T+h|T}) = \sigma^2\sum_{j=0}^{h-1}\psi_j^2.}$$

Use the **first h weights, starting with ψ_0**. Negative weights still add variance because they are squared. For example,

$$e_{T+2|T} = \varepsilon_{T+2} + \psi_1\varepsilon_{T+1}, \qquad v_2^2 = \sigma^2(1+\psi_1^2).$$

2 ACF, PACF, and Model Selection

We need to choose the model type and orders p, q. ACF/PACF suggest candidates; AIC/BIC compare fitted models.

ACF: are observations k periods apart correlated? For a covariance-stationary series,

$$\rho_k = \text{Corr}(y_t, y_{t-k}) = \frac{\gamma_k}{\gamma_0}.$$

Here γ_k is the lag-k autocovariance and γ_0 is the variance of y_t. An ACF plot shows ρ_k against the lag k.

PACF: does an older observation add information?

The lag-k PACF measures the correlation between y_t and y_{t-k} after controlling for $y_{t-1}, \ldots, y_{t-k+1}$.

For lag 2, use the regression

$$y_t = m + b_1 y_{t-1} + \beta_2 y_{t-2} + u_t.$$

Here m is the intercept and u_t is the regression error. The coefficient β_2 is the population lag-2 PACF: does y_{t-2} add linear predictive information after including y_{t-1}?

For lag k, include the first k lags and look at the coefficient on the last lag.

Example: an AR(1)

For $y_t = 0.6y_{t-1} + \varepsilon_t$,

$$\begin{aligned} \text{ACF:} \quad & \rho_1 = 0.6, \quad \rho_2 = 0.36, \quad \rho_3 = 0.216, \ldots, \\ \text{PACF:} \quad & \beta_1 = 0.6, \quad \beta_2 = \beta_3 = \cdots = 0. \end{aligned}$$

Today is correlated with earlier periods, but once yesterday's value is included, older observations add no further linear predictive information. A nonzero lag-2 ACF therefore does not by itself imply AR(2).

Patterns to look for

Model	Theoretical ACF	Theoretical PACF
AR(p)	Tails off	Zero after lag p
MA(q)	Zero after lag q	Tails off
ARMA(p, q)	Usually tails off	Usually tails off

Cutoff: zero after a particular lag. **Tailing off**: gradually approaches zero, possibly with oscillations. For MA(q), observations more than q periods apart share no shocks, explaining the ACF cutoff.

These are population patterns. Sample plots are noisy, so use the overall pattern to suggest models rather than expecting an exact cutoff.

AIC and BIC: is an extra lag worth adding?

On the same estimation sample, adding AR lags can improve fit but may also fit noise. AIC and BIC penalize the extra parameters. For an AR(p) with an intercept,

$$\begin{aligned} \text{AIC} &= \underbrace{\ln \widehat{\sigma}^2}_{\text{fit}} + \underbrace{\frac{2(p+1)}{T}}_{\text{penalty}}, \\ \text{BIC} &= \underbrace{\ln \widehat{\sigma}^2}_{\text{fit}} + \underbrace{\frac{(p+1)\ln T}{T}}_{\text{penalty}}. \end{aligned}$$

Here T is the sample size, $p+1$ counts the lag coefficients plus the intercept, and

$$\widehat{\sigma}^2 = \frac{1}{T}\sum_t \widehat{\varepsilon}_t^2$$

is the average squared residual, without a degrees-of-freedom adjustment.

Lower is better. Compare models using the same series and estimation sample. When $\ln T > 2$, BIC penalizes extra parameters more heavily than AIC.

For example, with $T = 100$:

Model	$\widehat{\sigma}^2$	AIC	BIC
AR(1)	1.00	0.0400	**0.0921**
AR(2)	0.97	**0.0295**	0.1077
AR(3)	0.96	0.0392	0.1434

For AR(2), AIC $= \ln(0.97) + 2(3)/100 = 0.0295$. AIC selects AR(2); BIC selects AR(1). AR(3) fits best in-sample, but neither criterion considers the improvement worth its extra parameters.

Check the residuals

The ACF of y_t helps identify a model. The ACF of the fitted residuals $\widehat{\varepsilon}_t$ checks what the model missed. Clear residual autocorrelation suggests that some linear predictive patterns remain unexplained.

3 Missing Data

For a few isolated gaps, the lecture recommends interpolation.

1. **Preserve the dates.** If February is missing, January and March are still two months apart. Do not compress the timeline.

2. **Fill an internal gap using its neighbors.**

   $$\widetilde{y}_t = \frac{y_{t-1} + y_{t+1}}{2}.$$

   Here $\widetilde{y}_t$ is the filled-in value. For example, neighboring unemployment rates of 3.1% and 3.2% give 3.15%.

   If the series will be logged, interpolate the logs:

   $$\widetilde{\ln P_t} = \frac{\ln P_{t-1} + \ln P_{t+1}}{2}.$$

   Then difference if needed to construct growth or inflation.

3. **Use only information available at the forecast date.** Both neighbors must already be available. A missing value at the end of the sample has no observed right-hand neighbor, so the same formula cannot be used.

For the project, record which values were filled and how. This simple approach is intended for a small number of gaps.

Exercises

Exercise 1: Representations and forecasts

For parts 1–2, consider

$$y_t = 0.8y_{t-2} + \varepsilon_t.$$

1. State the AR order and write the model using the lag operator. Expand into an infinite MA and identify $\psi_0, \psi_1, \psi_2, \psi_3$. Is the usual solution stationary? Can we recover ε_t from observations?

2. Suppose $y_T = 10$ and $y_{T-1} = 12$. Compute $\widehat{y}_{T+1|T}$ and $\widehat{y}_{T+2|T}$. Show which observed value enters each forecast.

3. Now consider a **different process**, whose first four Wold weights are

$$\psi_0 = 1, \qquad \psi_1 = 0.8, \qquad \psi_2 = 0.1, \qquad \psi_3 = -0.3, \qquad \sigma^2 = 10.$$

Write the four-step forecast error $e_{T+4|T}$ in terms of future shocks. Calculate $v_1^2, v_2^2, v_3^2, v_4^2$. Could this process be an MA(2)? Explain briefly.

Exercise 2: Choosing and checking a model

1. Sample ACF/PACF plots show the following patterns. Suggest a simple AR or MA model for each series. Explain which cutoff supports your choice.

Series	Sample ACF	Sample PACF
A	Gradually decays	Near zero after lag 2
B	Near zero after lag 1	Gradually decays

2. For series A, three AR models with intercepts are fitted using the same $T = 100$ observations. Complete the table and select a model under each criterion. Explain why the choices can differ.

   Use $\ln(100) \approx 4.6052$ and the supplied log values.

Model	$\widehat{\sigma}^2$	$\ln \widehat{\sigma}^2$	AIC	BIC
AR(1)	1.00	0.00000		
AR(2)	0.96	−0.04082		
AR(3)	0.95	−0.05129		

3. Suppose the model selected by BIC still has clear serial correlation in its residuals. Is having the lowest BIC enough to accept the model? Explain briefly.

Exercise 3: One missing observation

At the end of May, a monthly sales series used in levels has

$$y_{\text{March}} = 100, \qquad y_{\text{April}} = \text{missing}, \qquad y_{\text{May}} = 120.$$

Fill the April value using linear interpolation. Would the same fill be valid for a forecast made at the end of April? Explain in one sentence.

Answers

Exercise 1 Representations and forecasts.

For parts 1–2, consider

$$y_t = 0.8y_{t-2} + \varepsilon_t.$$

1. State the AR order and write the model using the lag operator. Expand into an infinite MA and identify $\psi_0, \psi_1, \psi_2, \psi_3$. Is the usual solution stationary? Can we recover ε_t from observations?

   This is an **AR(2)**: the largest lag is 2, even though the coefficient on y_{t-1} is zero.

   Using the lag operator,

$$\begin{aligned}(1 - 0.8L^2)y_t &= \varepsilon_t,\\ y_t &= \frac{1}{1 - 0.8L^2}\varepsilon_t\\ &= (1 + 0.8L^2 + 0.64L^4 + \cdots)\varepsilon_t\\ &= \varepsilon_t + 0.8\varepsilon_{t-2} + 0.64\varepsilon_{t-4} + \cdots .\end{aligned}$$

   Therefore,

$$\boxed{\psi_0 = 1, \qquad \psi_1 = 0, \qquad \psi_2 = 0.8, \qquad \psi_3 = 0.}$$

   The infinite MA is stationary: its coefficients decay geometrically, with

$$\sum_{j=0}^{\infty} \psi_j^2 = 1 + 0.8^2 + 0.8^4 + \cdots = \frac{1}{1 - 0.8^2} < \infty.$$

   The representation is also invertible because

$$\varepsilon_t = y_t - 0.8y_{t-2}.$$

2. Suppose $y_T = 10$ and $y_{T-1} = 12$. Compute $\widehat{y}_{T+1|T}$ and $\widehat{y}_{T+2|T}$. Show which observed value enters each forecast.

   Write the model at each target date:

$$y_{T+1} = 0.8y_{T-1} + \varepsilon_{T+1}, \qquad y_{T+2} = 0.8y_T + \varepsilon_{T+2}.$$

   Future shocks have conditional mean zero, so

$$\boxed{\widehat{y}_{T+1|T} = 0.8(12) = 9.6}, \qquad \boxed{\widehat{y}_{T+2|T} = 0.8(10) = 8}.$$

   The one-step forecast uses y_{T-1}, while the two-step forecast uses y_T.

3. Now consider a **different process**, whose first four Wold weights are

$$\psi_0 = 1, \qquad \psi_1 = 0.8, \qquad \psi_2 = 0.1, \qquad \psi_3 = -0.3, \qquad \sigma^2 = 10.$$

   Write the four-step forecast error $e_{T+4|T}$ in terms of future shocks. Calculate $v_1^2, v_2^2, v_3^2, v_4^2$. Could this process be an MA(2)? Explain briefly.

The four-step forecast error is

$$e_{T+4|T} = \varepsilon_{T+4} + 0.8\varepsilon_{T+3} + 0.1\varepsilon_{T+2} - 0.3\varepsilon_{T+1}.$$

Using $v_h^2 = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2$,

$$\begin{aligned} v_1^2 &= 10(1) = \boxed{10}, \\ v_2^2 &= 10(1 + 0.8^2) = \boxed{16.4}, \\ v_3^2 &= 10(1 + 0.8^2 + 0.1^2) = \boxed{16.5}, \\ v_4^2 &= 10(1 + 0.8^2 + 0.1^2 + (-0.3)^2) = \boxed{17.4}. \end{aligned}$$

The negative weight still adds $10(-0.3)^2 = 0.9$ to the variance.

This cannot be an MA(2), because an MA(2) has $\psi_j = 0$ for every $j > 2$, but here $\psi_3 = -0.3 \neq 0$.

Exercise 2 Choosing and checking a model.

1. Sample ACF/PACF plots show the following patterns. Suggest a simple AR or MA model for each series. Explain which cutoff supports your choice.

Series	Sample ACF	Sample PACF
A	Gradually decays	Near zero after lag 2
B	Near zero after lag 1	Gradually decays

- **Series A: AR(2).** The PACF is near zero after lag 2, while the ACF gradually decays.
- **Series B: MA(1).** The ACF is near zero after lag 1, while the PACF gradually decays.

These are candidate models suggested by noisy sample patterns, not proof of the true model.

2. For series A, three AR models with intercepts are fitted using the same $T = 100$ observations. Complete the table and select a model under each criterion. Explain why the choices can differ.

   Use $\ln(100) \approx 4.6052$ and the supplied log values.

Model	$\widehat{\sigma}^2$	$\ln \widehat{\sigma}^2$	AIC	BIC
AR(1)	1.00	0.00000		
AR(2)	0.96	−0.04082		
AR(3)	0.95	−0.05129		

Use

$$\text{AIC} = \ln \widehat{\sigma}^2 + \frac{2(p+1)}{100}, \qquad \text{BIC} = \ln \widehat{\sigma}^2 + \frac{(p+1)\ln 100}{100}.$$

For example, for AR(2),

$$\begin{aligned} \text{AIC} &= \ln(0.96) + \frac{2(3)}{100} \approx 0.0192, \\ \text{BIC} &= \ln(0.96) + \frac{3\ln 100}{100} \approx 0.0973. \end{aligned}$$

Model	AIC	BIC
AR(1)	0.0400	**0.0921**
AR(2)	**0.0192**	0.0973
AR(3)	0.0287	0.1329

AIC selects AR(2); BIC selects AR(1). Because $\ln 100 > 2$, BIC penalizes additional parameters more heavily. The improvement from AR(1) to AR(2) is enough under AIC, but not under BIC.

3. Suppose the model selected by BIC still has clear serial correlation in its residuals. Is having the lowest BIC enough to accept the model? Explain briefly.

   No. Clear residual autocorrelation suggests that the model has left some linear predictive patterns unexplained.

   The lowest BIC identifies the best score among the models considered; it does not guarantee that the selected model is adequate. Revisit the candidate models and check the residuals again.

Exercise 3 One missing observation.

At the end of May, a monthly sales series used in levels has

$$y_{\text{March}} = 100, \qquad y_{\text{April}} = \text{missing}, \qquad y_{\text{May}} = 120.$$

Fill the April value using linear interpolation. Would the same fill be valid for a forecast made at the end of April? Explain in one sentence.

At the end of May, both neighbors are available:

$$\boxed{\widetilde{y}_{\text{April}} = \frac{100 + 120}{2} = 110.}$$

This fill is **not valid for a forecast made at the end of April**, because it uses the May observation, which was not yet available.