Lecture 2 — AR, MA, and ARMA Models

AR(p) roots and Yule–Walker intuition; MA(q) and invertibility; ARMA(p, q), lag polynomials, causality, and invertibility

Jiajing Sun†

School of Economics and Management, University of Chinese Academy of Sciences†

March 2026

Lecture Roadmap

Lecture 2 in the 20-lecture sequence

- **Lecture 1:** dependence, stationarity, ergodicity, mixing, and Wold decomposition.
- **Lecture 2:** AR(p), MA(q), ARMA(p, q), lag polynomials, roots, causality, and invertibility.
- **Lecture 3:** sample ACF/PACF, model identification, estimation, diagnostics, and forecasting.
- **Later blocks:** nonstationarity, VAR / VECM / SVAR, volatility, nonparametrics, robust inference, filtering, and continuous-time finance.

Why this lecture matters

Lecture 1 gave the representation theorem. Lecture 2 turns that representation logic into the classical parametric models that we actually estimate in practice.

Learning goals for this 3-hour lecture

By the end of this lecture, students should be able to:

1. define AR(p), MA(q), and ARMA(p, q) models in both scalar and lag-polynomial form;
2. explain why roots outside the unit circle are the key condition for stable one-sided representations;
3. derive the Yule–Walker equations and understand what information about dependence they encode;
4. explain why MA models are automatically stationary but still need invertibility for identification;
5. interpret ARMA models as parsimonious approximations to Wold-type dynamics.

Practical plan for the three contact hours

- **Hour 1:** AR(p) models, roots, stationarity, and Yule–Walker intuition.
- **Hour 2:** MA(q) models, finite shock memory, and invertibility.
- **Hour 3:** ARMA(p, q), lag polynomials, causality, invertibility, and mixed dynamics.

Teaching emphasis

The algebra is not the destination. The destination is dynamic intuition: how shocks enter, how long they last, what can be predicted, and how the data reveal the underlying dependence structure.

From Wold to ARMA: the conceptual chain

Stationarity $\to$ Wold representation $\to$ low-dimensional approximation $\to$ AR / MA / ARMA models

- Wold says that a covariance-stationary process has a one-sided innovation representation.
- Empirical work needs tractable models with finitely many parameters.
- AR, MA, and ARMA models are precisely the standard finite-dimensional approximations to that general representation.

Why AR, MA, and ARMA are not ad hoc

- An AR model emphasizes dependence through lagged values of the series itself.
- An MA model emphasizes dependence through current and lagged innovations.
- An ARMA model combines both channels in one compact structure.

Key message

The point of ARMA theory is not that every time series is exactly ARMA. The point is that ARMA models give a parsimonious and interpretable approximation to the short-memory dynamics implied by the Wold decomposition.

Three ways to think about a time-series model

1. **Recursion view:** how is Y_t generated from past values and shocks?
2. **Filter view:** how is Y_t written as a linear filter of innovations?
3. **Second-order view:** what autocovariances and autocorrelations does the model imply?

A good time-series economist should be able to move fluently across all three views.

The three model classes in one line

- **AR(p):**
$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t.$$
- **MA(q):**
$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$
- **ARMA(p, q):** combines the two in one equation.

Each class has its own root condition, its own identification logic, and its own dependence pattern.

A quick reminder: what is an innovation?

- The innovation is the genuinely new information arriving at time t.
- In linear forecasting language, it is the one-step-ahead forecast error.
- In model notation, we usually write it as ε_t and assume at least a white-noise structure:

$$\mathrm{E}(\varepsilon_t) = 0, \qquad \mathrm{Var}(\varepsilon_t) = \sigma^2, \qquad \mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0 \text{ for } t \neq s.$$

This innovation language is the common thread linking Wold, ARMA theory, forecasting, and later filtering methods.

Why the lag operator is indispensable

Define the lag operator by

$$LY_t = Y_{t-1}, \qquad L^k Y_t = Y_{t-k}.$$

- The lag operator turns recursions into polynomial algebra.
- It lets us express dynamic models compactly and study them via characteristic polynomials.
- It also scales immediately to multivariate systems later in the course.

Without the lag operator, ARMA theory is much harder to organize and much harder to generalize.

Lag polynomials behave like ordinary polynomials

If

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \qquad \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q,$$

then we can add, multiply, factorize, and sometimes invert these objects just as in ordinary algebra.

- The inversion step is where root conditions become essential.
- Stable inversion gives one-sided infinite representations.

The geometric-series prototype

For a scalar constant a,

$$(1 - aL)^{-1} = 1 + aL + a^2 L^2 + \cdots$$

provided $|a| < 1$.

- This is the prototype for AR causality.
- The same idea, with a different polynomial, is also the prototype for MA invertibility.
- The entire lecture can be viewed as a sophisticated extension of this simple geometric-series identity.

What roots mean economically

- A root of the AR polynomial tells us whether the recursion can be written as a stable function of current and past shocks.
- A root of the MA polynomial tells us whether shocks can be recovered uniquely from observed current and past values.
- Root locations therefore have direct probabilistic and economic meaning: they tell us whether dependence is stable, whether shocks are identifiable, and whether the model can be used operationally for forecasting.

Lecture Roadmap

1. Motivation and setup
2. AR(p) models: definitions and basic intuition
3. MA(q) models and invertibility
4. Lag polynomials and ARMA(p, q) models
5. R illustrations
6. Summary and wrap-up

Hour 1 begins here

AR(p) models, roots, stationarity, and Yule–Walker intuition

- We begin with autoregressions because they are the cleanest parametric models of persistent stationary dependence.

Definition of an AR(p) model

A process follows an autoregressive model of order p if

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is usually taken to be white noise with

$$\mathrm{E}(\varepsilon_t) = 0, \qquad \mathrm{Var}(\varepsilon_t) = \sigma^2.$$

- The model is recursive: current outcomes inherit information from past outcomes.
- The innovation refreshes the system with genuinely new information.

How to interpret the AR recursion

- The AR part captures persistence in the *state* of the process.
- If the coefficients are small, shocks die out quickly.
- If the coefficients are large in magnitude, shocks propagate for a long time.
- Because the model depends on lagged values of the series itself, an AR model is often a natural first model for macroeconomic and financial data with visible persistence.

White noise is the benchmark shock sequence

- White noise means zero mean, constant finite variance, and zero autocovariances at nonzero lags.
- White noise is weaker than i.i.d.; nonlinear dependence may still exist unless independence is imposed separately.
- In linear time-series theory, white noise is the minimal benchmark needed to interpret ε_t as an innovation sequence.

$$\varepsilon_t \sim WN(0, \sigma^2)$$

is therefore the standard baseline notation.

Why we usually center the process

If the AR(p) process is stationary with mean μ, then taking expectations gives

$$\mu = c + \phi_1\mu + \cdots + \phi_p\mu,$$

so

$$\mu = \frac{c}{1 - \phi_1 - \cdots - \phi_p},$$

provided the denominator is nonzero.

It is therefore convenient to define $X_t = Y_t - \mu$ and work with the mean-zero representation.

Centered AR(p) form

With $X_t = Y_t - \mu$, the model becomes

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t.$$

- The constant term determines the unconditional mean.
- The root condition determines whether the dependence structure is stable.
- Most theoretical derivations—ACF formulas, Yule–Walker, and infinite representations—are cleaner in centered form.

AR(1): the basic benchmark

The simplest autoregression is

$$X_t = \phi X_{t-1} + \varepsilon_t.$$

This model already illustrates almost all of the core themes of Lecture 2:

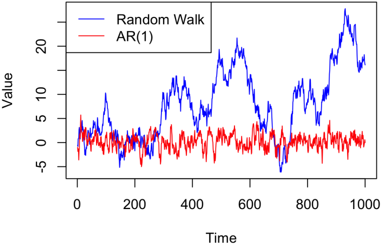
- stable versus unstable persistence;
- characteristic roots;
- infinite moving-average representation;
- geometric decay of impulse effects;
- autocorrelation patterns and forecast convergence.

AR(1): three regimes

- $|\phi| < 1$**:** stationary and causal. Shocks decay over time.
- $|\phi| = 1$**:** unit-root boundary. Shocks do not die out.
- $|\phi| > 1$**:** explosive. The recursion magnifies disturbances instead of damping them.

These are not minor variations. They are fundamentally different probabilistic regimes with different asymptotic theory and different forecasting behavior.

A stable AR(1) versus a random walk



A stationary AR(1) fluctuates around a stable center. A random walk accumulates shocks permanently. This is exactly why the unit-root boundary is such a major dividing line in time-series econometrics.

Why I prefer geometric-decay language here

For a stable AR(1),

$$X_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots .$$

- The coefficient ϕ^h is the effect of a shock h periods later.
- This geometric-decay interpretation is always clean and always correct for AR(1).
- The popular "half-life" shortcut is really just a special summary for positive AR(1) persistence. The more general and more robust language is simply: shocks decay geometrically when $|\phi| < 1$.

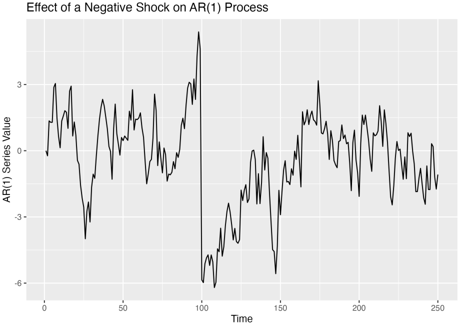
AR(1) as an MA(∞) process

If $|\phi| < 1$, repeated substitution gives

$$X_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}.$$

- This is the simplest concrete example of a causal one-sided representation.
- A stable autoregression can always be rewritten as an infinite moving average.
- The weights ϕ^j show exactly how fast old shocks are forgotten.

Shock propagation in a stable AR(1)



The effect of a one-time innovation is largest on impact and then decays geometrically. Persistence is not binary; it is a quantitative statement about how quickly the impulse-response weights shrink.

Moments of a stationary AR(1)

If

$$X_t = \phi X_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2), \qquad |\phi| < 1,$$

then

$$\mathrm{E}(X_t) = 0, \qquad \mathrm{Var}(X_t) = \frac{\sigma^2}{1 - \phi^2}.$$

- The variance is finite only when $|\phi| < 1$.
- As $|\phi|$ gets close to one, the unconditional variance becomes very large.
- This is another way to see why near-unit-root processes look highly persistent in finite samples.

Autocorrelation of a stationary AR(1)

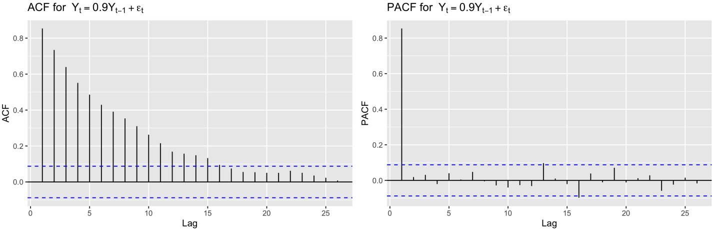
For the same AR(1) process,

$$\rho_k = \phi^k, \qquad k = 0, 1, 2, \dots .$$

- Positive ϕ gives monotone decay.
- Negative ϕ gives alternating signs.
- Larger $|\phi|$ means stronger persistence at all horizons.

The AR(1) therefore links one coefficient to the full second-order dependence pattern.

AR(1) ACF and PACF pattern



For a pure autoregression, the ACF tails off while the PACF cuts off at the true lag order. In Lecture 3 we will formalize how this helps with identification.

Near-unit-root AR(1) is still not a unit root

- If $\phi = 0.98$, the process is still stationary, and old shocks eventually die out.
- If $\phi = 1$, shocks never die out at all.
- In finite samples the two may look deceptively similar, but their probability limits, test statistics, and forecast behavior are fundamentally different.

This distinction is one of the most important cautions in empirical macroeconometrics.

AR(2): why order two is already much richer

An AR(2) model is

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t.$$

Compared with AR(1), AR(2) can generate:

- monotone mean reversion,
- damped cycles,
- alternating dependence,
- more flexible short-run autocorrelation patterns.

The second lag is enough to make root geometry and dynamic interpretation genuinely richer.

Real and complex roots in AR dynamics

- **Real roots** tend to produce non-oscillatory adjustment.
- **Complex roots** typically produce cyclical or oscillatory behavior.
- The modulus of the root controls the decay rate.
- The angle of a complex root controls the oscillation frequency.

This is why root analysis is not just a mathematical trick. It tells us whether the series returns to equilibrium smoothly or through damped cycles.

AR(p) in lag-polynomial form

The centered AR(p) model can be written as

$$\phi(L)X_t = \varepsilon_t,$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p.$$

- The recursion is now condensed into a single polynomial equation.
- The key question becomes whether $\phi(L)$ can be inverted stably.

The AR characteristic polynomial

The associated characteristic polynomial is

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p.$$

Its zeros are the characteristic roots.

Stationarity / causality criterion

A causal covariance-stationary AR(p) solution exists if and only if every root of $\phi(z) = 0$ lies outside the unit circle.

Why roots outside the unit circle?

If all roots satisfy $|z| > 1$, then the inverse filter exists as a convergent one-sided power series:

$$\phi(L)^{-1} = \sum_{j=0}^{\infty} \psi_j L^j, \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Hence

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

- The current value depends only on current and past shocks.
- The dependence weights are summable, so old shocks are damped rather than amplified.

Root condition in reciprocal form

Sometimes we factor the AR polynomial as

$$\phi(L) = \prod_{j=1}^{p} (1 - \lambda_j L).$$

Then the roots of $\phi(z) = 0$ are $z_j = 1/\lambda_j$.

So the condition $|z_j| > 1$ is equivalent to $|\lambda_j| < 1$.

This reciprocal notation is often useful when we want to read the decay rates directly from the factorization.

The geometric-series inverse revisited

For AR(1), the inversion is especially transparent:

$$(1-\phi L)^{-1} = 1 + \phi L + \phi^2 L^2 + \cdots$$

whenever $|\phi| < 1$.

- This is exactly the algebra behind $X_t = \sum_{j\geq 0} \phi^j \varepsilon_{t-j}$.
- Higher-order AR models use the same idea, only with more complicated polynomials and possibly multiple factors.

A caution on the unit-root boundary

If a root lies on the unit circle, the inverse filter is no longer absolutely summable.

- The classic example is $(1 - L)X_t = \varepsilon_t$, or $X_t = X_{t-1} + \varepsilon_t$.
- Here shocks cumulate permanently.
- Variance grows over time rather than staying fixed.

Later in the course this boundary case will lead us to unit-root tests, stochastic trends, and cointegration.

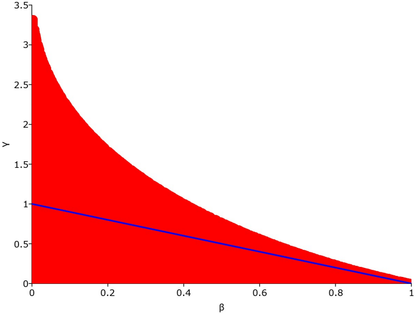
The AR(2) stationarity region

For AR(2), stationarity is not described by a single interval but by a region in the (ϕ_1, ϕ_2) plane.

- Different coefficient pairs can imply stable monotone adjustment.
- Others imply damped oscillation.
- Outside the region, the recursion becomes unstable.

So already at order two, root geometry becomes more informative than simply staring at coefficient magnitudes.

A visual stationarity region for AR(2)



The admissible region is the set of coefficient pairs for which both roots of the AR polynomial lie outside the unit circle. The boundary is where one root touches the unit circle.

Equivalent inequalities for a stationary AR(2)

For the AR(2) model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t,$$

the stationarity region can be written as

$$\phi_2 + \phi_1 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad -1 < \phi_2 < 1.$$

- These inequalities are algebraically equivalent to the root-outside-the-unit-circle condition.
- They are convenient for quick checks, but the root formulation is conceptually more general.

Dynamic interpretation of AR roots

- Roots far outside the unit circle correspond to rapid decay of dependence.
- Roots close to the unit circle correspond to strong persistence.
- A repeated or nearly repeated root can generate especially persistent dynamics.
- Complex roots create oscillatory mean reversion.

This is why economists often read roots as a map of shock persistence.

Causality is the operative notion for AR models

Definition

A process is **causal** if it can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

- Causality means the present depends on current and past shocks only.
- For AR(p), the causal condition is exactly the root-outside-the-unit-circle condition.

Why causality matters

- It gives an operational data-generating mechanism: shocks arrive, then propagate forward.
- It produces stable forecasting formulas.
- It links AR models directly to the Wold representation.
- It rules out dependence on future innovations, which would be economically and probabilistically awkward.

So when applied time-series textbooks say a stationary AR model, they usually mean the causal stationary solution.

The impulse-response sequence for AR(p)

If $\phi(L)^{-1} = \sum_{j=0}^{\infty} \psi_j L^j$, then

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

- $\psi_0 = 1$ is the impact effect.
- ψ_j for $j \geq 1$ measures the effect of a shock j periods later.
- The full sequence $\{\psi_j\}$ is the dynamic propagation profile of the model.

For AR(1), $\psi_j = \phi^j$. For higher-order AR models, the sequence is determined recursively by the polynomial inverse.

A recursion for the AR(p) MA(∞) coefficients

If

$$\phi(L)^{-1} = \sum_{j=0}^{\infty} \psi_j L^j,$$

then multiplying both sides by $\phi(L)$ yields the coefficient recursion

$$\psi_0 = 1, \qquad \psi_j = \phi_1 \psi_{j-1} + \cdots + \phi_p \psi_{j-p} \quad (j \geq 1),$$

where $\psi_j = 0$ for $j < 0$.

This recursion is useful both theoretically and computationally.

The Yule–Walker idea in one sentence

The Yule–Walker equations come from a simple trick:

take the AR equation and covary both sides with lagged values of the process.

This converts model coefficients into restrictions on autocovariances.

Yule–Walker derivation for AR(1)

Start from

$$X_t = \phi X_{t-1} + \varepsilon_t.$$

Take covariance with X_{t-k} for $k \geq 1$:

$$\mathrm{Cov}(X_t, X_{t-k}) = \phi \mathrm{Cov}(X_{t-1}, X_{t-k}) + \mathrm{Cov}(\varepsilon_t, X_{t-k}).$$

Because ε_t is orthogonal to the past under the causal representation,

$$\mathrm{Cov}(\varepsilon_t, X_{t-k}) = 0, \qquad k \geq 1,$$

so

$$\gamma_k = \phi \gamma_{k-1}, \qquad k \geq 1.$$

What Yule–Walker immediately implies for AR(1)

From

$$\gamma_k = \phi \gamma_{k-1},$$

we obtain

$$\rho_k = \phi^k, \qquad k \geq 0.$$

- One parameter determines the whole ACF.
- The ACF decays geometrically.
- The sign of ϕ determines whether the correlations all have the same sign or alternate.

The variance equation for AR(1)

Set $k = 0$ in the covariance logic:

$$\gamma_0 = \mathrm{Cov}(X_t, X_t) = \phi \mathrm{Cov}(X_{t-1}, X_t) + \mathrm{Cov}(\varepsilon_t, X_t).$$

Using $\gamma_1 = \phi\gamma_0$ and $\mathrm{Cov}(\varepsilon_t, X_t) = \sigma^2$, we get

$$\gamma_0 = \phi^2\gamma_0 + \sigma^2.$$

Therefore

$$\gamma_0 = \frac{\sigma^2}{1-\phi^2}.$$

This is the cleanest moment derivation for a stable AR(1).

Yule–Walker for AR(2)

For

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t,$$

the same covariance logic gives

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}, \qquad k \geq 1.$$

So the autocovariances satisfy the same homogeneous recursion as the data process itself, except without the innovation term at positive lags.

The first Yule–Walker equations for AR(2)

For AR(2),

$$\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1,$$

$$\gamma_2 = \phi_1\gamma_1 + \phi_2\gamma_0.$$

In autocorrelation form this becomes

$$\rho_1 = \frac{\phi_1}{1-\phi_2}, \qquad \rho_2 = \phi_1\rho_1 + \phi_2.$$

These equations show how AR coefficients map into the first few dependence measures.

General Yule–Walker equations for AR(p)

For the centered AR(p) model

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

the Yule–Walker equations are

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p}, \qquad k \geq 1.$$

In autocorrelation form,

$$\rho_k = \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p}.$$

The ACF of a pure AR process therefore never truly cuts off; it obeys a recursive decay law.

Matrix form of Yule–Walker

The first p Yule–Walker equations can be written as

$$\begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}.$$

This Toeplitz structure is one reason autoregressions are so convenient computationally.

Why Yule–Walker is useful

- It links model parameters to the observed second-order structure.
- It gives a closed-form estimation route once we replace population autocovariances by sample ones.
- It clarifies why AR models have a tailing-off ACF but a finite-order PACF.

In short, Yule–Walker is the bridge between recursion parameters and correlation patterns.

Yule–Walker estimation intuition

Replace γ_k by sample autocovariances $\hat{\gamma}_k$ and solve the linear system.

- This gives the classical Yule–Walker estimator of the AR coefficients.
- The method is fast and often useful pedagogically.
- Later we will compare it with least squares and maximum likelihood.

The point for today is conceptual: sample dependence measures can be used to recover autoregressive parameters.

Sample ACF versus model ACF

- In population, the AR ACF satisfies the Yule–Walker recursion exactly.
- In finite samples, the sample ACF only approximates that pattern.
- Therefore identification is always probabilistic rather than mechanical.

This is why we will later combine ACF/PACF inspection with information criteria and diagnostics instead of trusting one plot blindly.

Identification pattern for pure AR models

- The ACF of a pure AR model typically tails off.
- The PACF of a pure AR(p) cuts off after lag p.
- The tail may be geometric, oscillatory, or a mixture, depending on the roots.

The cut-off property belongs to the PACF, not the ACF. This is one of the most important practical rules in elementary ARMA identification.

Why the PACF cuts off for a pure AR(p)

Once the first p lags are controlled for, the AR recursion says there is no additional direct linear dependence from deeper lags.

- Longer-lag correlation is still present in the raw ACF.
- But it operates indirectly through the first p lags.

That is exactly what the PACF is designed to strip away.

A compact AR checkpoint

Questions students should now be able to answer

1. Why does the AR root condition involve the inverse polynomial $\phi(z)$ rather than the coefficients separately?
2. Why is geometric-decay language more reliable than a generic "half-life" slogan?
3. Why does the ACF of a pure AR process tail off rather than cut off?

Lecture Roadmap

1. Motivation and setup
2. AR(p) models: definitions and basic intuition
3. MA(q) models and invertibility
4. Lag polynomials and ARMA(p, q) models
5. R illustrations
6. Summary and wrap-up

Hour 2 begins here

MA(q) models and invertibility

An MA model is very different from an AR model. The dependence does not come through lagged values of the series. It comes through a finite distributed lag of shocks.

Definition of an MA(q) model

A moving-average model of order q is

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

In centered form,

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

- The model is a finite linear filter of white noise.
- Each innovation affects the process for only finitely many periods.

Dynamic interpretation of an MA model

- In an AR model, persistence travels through the state variable itself.
- In an MA model, persistence travels directly through past innovations.
- Once an innovation is older than q periods, it no longer affects the process at all.

So MA models are natural descriptions of *finite shock memory*.

Why finite-order MA models are automatically stationary

If

$$X_t = \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \qquad \theta_0 = 1,$$

then X_t is just a finite linear combination of white-noise terms.

- The mean is constant.
- The variance is finite whenever $\sigma^2 < \infty$.
- Covariances depend only on lag because the same finite filter is used at every date.

So stationarity is automatic for finite-order MA models.

General covariance formulas for MA(q)

For

$$X_t = \sum_{j=0}^{q} \theta_j \varepsilon_{t-j},$$

we have

$$\gamma_0 = \sigma^2 \sum_{j=0}^{q} \theta_j^2,$$

and for $0 \le k \le q$,

$$\gamma_k = \sigma^2 \sum_{j=0}^{q-k} \theta_j \theta_{j+k}.$$

For $k > q$,

$$\gamma_k = 0.$$

This finite cut-off in the autocovariance sequence is the hallmark of an MA process.

Why the ACF cuts off for MA(q)

If two observations are more than q periods apart, they are built from disjoint sets of white-noise shocks.

- No common shocks means zero covariance.
- Hence $\gamma_k = 0$ and $\rho_k = 0$ for all $k > q$.

This is the exact opposite of the pure AR case, where dependence typically tails off but does not cut off.

MA(1): the benchmark finite-memory model

The simplest moving-average model is

$$X_t = \varepsilon_t + \theta \varepsilon_{t-1}.$$

- The current shock matters immediately.
- The previous shock still matters one period later.
- After that, the shock is forgotten completely.

This is the cleanest example of a process with short but finite shock memory.

Moments of an MA(1)

For

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

we obtain

$$\gamma_0 = (1+\theta^2)\sigma^2, \qquad \gamma_1 = \theta\sigma^2, \qquad \gamma_k = 0 \text{ for } k \geq 2.$$

Therefore

$$\rho_1 = \frac{\theta}{1+\theta^2}, \qquad \rho_k = 0 \text{ for } k \geq 2.$$

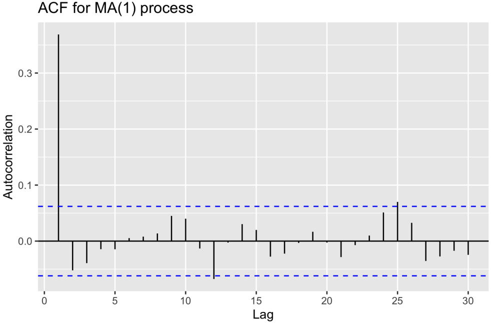
Interpreting the MA(1) autocorrelation

- Only the first autocorrelation can be nonzero.
- The sign of ρ_1 follows the sign of θ.
- The magnitude of ρ_1 is bounded away from one, because

$$\left|\frac{\theta}{1+\theta^2}\right| \leq \frac{1}{2}.$$

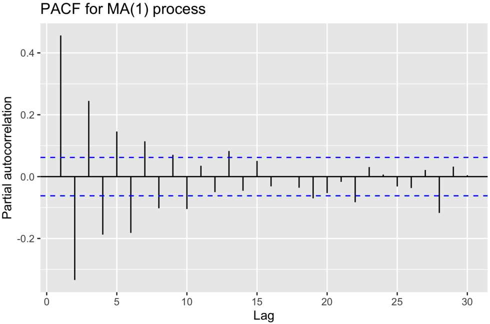
So a pure MA(1) cannot generate arbitrarily strong first-order correlation.

A typical MA(1) ACF pattern



In population, the ACF cuts off after lag 1. In finite samples, later lags fluctuate around zero rather than landing exactly on zero.

A typical MA(1) PACF pattern



The PACF of a pure MA model tails off rather than cutting off. This is the mirror image of the pure AR case.

MA(2): slightly richer but still finite-memory

For

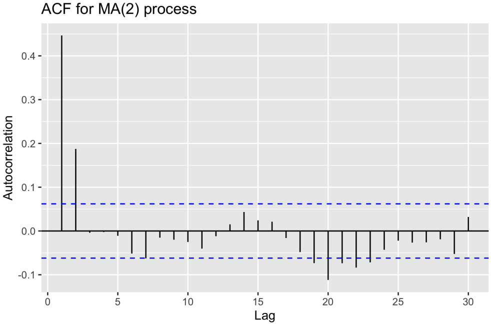
$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2},$$

the first two autocorrelations can be nonzero, but all higher ones are zero.

$$\rho_1 = \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2}, \qquad \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2}, \qquad \rho_k = 0 \text{ for } k \geq 3.$$

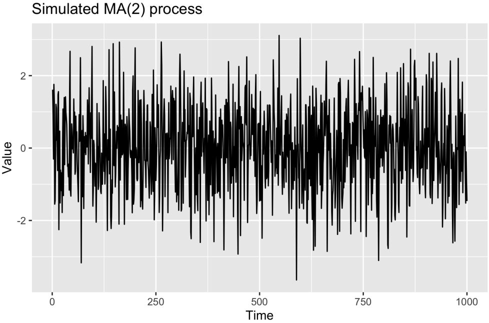
The ACF cut-off moves from lag 1 to lag 2.

A typical MA(2) ACF pattern



The key feature is still the same: beyond the order of the MA filter, the population autocorrelations vanish.

A simulated MA(2) process



The series is stationary, but each shock has only a short-lived effect. Visually, MA data often look “locally irregular” because shocks do not propagate through a persistent state variable.

Why MA models raise an identification problem

For AR models the main issue is stationarity. For finite MA models, stationarity is automatic.

The real problem is different:

different MA parameterizations can generate the same second-order behavior.

This is why we impose invertibility.

Observational equivalence in MA(1)

For MA(1),

$$\rho_1 = \frac{\theta}{1+\theta^2}.$$

This expression is unchanged when θ is replaced by $1/\theta$.

- So θ and $1/\theta$ generate the same autocorrelation pattern.
- After a suitable adjustment of innovation variance, they generate the same second-order observed process.

The data alone do not tell us which representation is the economically meaningful one.

Invertibility: the definition

Definition

An ARMA(p, q) process is **invertible** if there exists an absolutely summable sequence $\{\pi_j\}$ such that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

- Invertibility means shocks can be recovered uniquely as a stable function of current and past observations.
- It is the MA-side analogue of AR causality.

Invertibility for MA(q)

Write the MA model as

$$X_t = \theta(L)\varepsilon_t, \qquad \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q.$$

The model is invertible if

$$\theta(L)^{-1} = \sum_{j=0}^{\infty} \pi_j L^j$$

exists as an absolutely summable one-sided series.

For finite-order MA models, this is equivalent to all roots of $\theta(z) = 0$ lying outside the unit circle.

MA(1) invertibility condition

For

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

the polynomial is $\theta(z) = 1 + \theta z$.
Its root is

$$z = -\frac{1}{\theta}.$$

So the root-outside-the-unit-circle condition gives

$$|\theta| < 1.$$

Under our sign convention, the invertible MA(1) representation is therefore the one with $|\theta| < 1$.

The invertible MA(1) as an AR(∞)

If $|\theta| < 1$, then

$$\varepsilon_t = (1+\theta L)^{-1} X_t = (1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \cdots) X_t.$$

Hence

$$\varepsilon_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \cdots .$$

So an invertible MA process can be rewritten as an infinite autoregression in the observed data.

Why invertibility is operationally essential

- Forecasting formulas are written in terms of innovations.
- Likelihood-based estimation treats innovations as the primitive shocks.
- Structural interpretation also requires a unique mapping from observed data back to shocks.

Without invertibility, the same observed process can correspond to multiple shock representations, and that ambiguity is typically unacceptable in practice.

AR versus MA: the central contrast

Model class	Main issue	Signature pattern
AR(p)	stationarity / causality	ACF tails off
MA(q)	invertibility	ACF cuts off

This table is simple, but it encodes a very large amount of theory.

Finite memory versus propagated memory

- In an MA model, a shock lives for a finite number of periods.
- In a stationary AR model, a shock usually lives forever in theory, but with declining weight.
- So both models can be short-memory, but they organize that short memory in very different ways.

This difference is why AR and MA processes often look similar in raw time plots but very different in their ACF/PACF structure.

A compact MA checkpoint

Questions students should now be able to answer

1. Why is every finite-order MA model covariance stationary?
2. Why do we still need invertibility even though stationarity is automatic?
3. Why can θ and $1/\theta$ generate the same MA(1) autocorrelation?

Lecture Roadmap

1. Motivation and setup
2. AR(p) models: definitions and basic intuition
3. MA(q) models and invertibility
4. Lag polynomials and ARMA(p, q) models
5. R illustrations
6. Summary and wrap-up

Hour 3 begins here

ARMA(p, q), lag polynomials, causality, and invertibility

ARMA models are the natural mixed class: AR for persistent propagation, MA for short-run shock transmission.

Lag polynomials as dynamic algebra

- $\phi(L)$ encodes autoregressive feedback through past values.
- $\theta(L)$ encodes moving-average transmission through current and past shocks.
- Multiplication of polynomials corresponds to composition of filters.
- Factorization exposes the root structure immediately.

This is why lag-polynomial notation is more than compact writing. It is the algebraic language of dynamic systems.

Two basic polynomials

We write

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \qquad \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q.$$

- The AR polynomial is attached to the observed series.
- The MA polynomial is attached to the innovation sequence.
- Causality is a root condition on $\phi(z)$.
- Invertibility is a root condition on $\theta(z)$.

Definition of an ARMA(p, q) model

An ARMA(p, q) model is

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

In centered form,

$$\phi(L) X_t = \theta(L) \varepsilon_t.$$

This single compact equation is the standard starting point for ARMA theory.

Why ARMA is often more parsimonious

- A high-order AR may be approximated well by a lower-order ARMA.
- A high-order MA may also be approximated by an ARMA.
- The mixed class lets us capture both persistent propagation and short-run innovation effects with relatively few parameters.

This is why ARMA models are often preferred empirically to a purely high-order AR or purely high-order MA specification.

Special cases are nested inside ARMA

- AR(p) is ARMA($p, 0$).
- MA(q) is ARMA($0, q$).
- White noise is ARMA($0, 0$).

So the mixed class does not replace the pure classes. It contains them.

The mean of an ARMA process

If the process is stationary and $1 - \phi_1 - \cdots - \phi_p \neq 0$, then

$$\mu = \mathrm{E}(Y_t) = \frac{c}{1 - \phi_1 - \cdots - \phi_p}.$$

The MA coefficients do not affect the unconditional mean directly because the innovation process has zero mean.

As before, most analysis is cleaner in terms of the centered process $X_t = Y_t - \mu$.

Causality for ARMA

Definition

An ARMA(p, q) process is **causal** if there exists an absolutely summable sequence $\{\psi_j\}$ such that

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

- Causality means the process is generated by current and past shocks only.
- It is the operational notion of stationarity used throughout classical ARMA theory.

Root condition for ARMA causality

For ARMA(p, q),

$$\phi(L)X_t = \theta(L)\varepsilon_t.$$

The process is causal if and only if all roots of

$$\phi(z) = 0$$

lie outside the unit circle.

Under that condition,

$$X_t = \frac{\theta(L)}{\phi(L)}\varepsilon_t = \psi(L)\varepsilon_t,$$

with an absolutely summable one-sided filter $\psi(L)$.

Invertibility for ARMA

Definition

An ARMA(p, q) process is **invertible** if there exists an absolutely summable sequence $\{\pi_j\}$ such that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

- Invertibility recovers shocks from the observed data.
- It is the natural identification restriction for the MA side of the model.

Root condition for ARMA invertibility

The ARMA process is invertible if and only if all roots of

$$\theta(z) = 0$$

lie outside the unit circle.
Under that condition,

$$\varepsilon_t = \frac{\phi(L)}{\theta(L)} X_t = \pi(L) X_t,$$

where $\pi(L)$ is an absolutely summable one-sided filter.

Why standard ARMA modeling wants both conditions

- **Causality** gives a stable forward representation of the data in terms of shocks.
- **Invertibility** gives a stable backward representation of shocks in terms of the data.

Together they ensure that the model is both dynamically well behaved and uniquely interpretable.

A warning about common factors

Suppose $\phi(L)$ and $\theta(L)$ share a common factor:

$$\phi(L) = \tilde{\phi}(L)(1 - aL), \qquad \theta(L) = \tilde{\theta}(L)(1 - aL).$$

Then the factor cancels and the model is observationally equivalent to a lower-order ARMA process.

So a properly specified ARMA model should not contain redundant common factors.

Why cancellation matters economically

If an AR and an MA factor nearly cancel, the process can look deceptively simple.

- The observed series may look almost like white noise.
- Yet there is still hidden dynamic structure in the parameterization.
- Estimation can become numerically delicate when near-cancellation occurs.

This is one reason ARMA identification is harder than pure AR or pure MA identification.

ARMA(1,1): the benchmark mixed model

The simplest mixed model is

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}.$$

- The AR part propagates the state through time.
- The MA part modifies the immediate and short-run impact of shocks.

ARMA(1,1) is the workhorse example for understanding the interaction between persistence and innovation smoothing.

Repeated substitution for ARMA(1,1)

If $|\phi| < 1$, then

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$

implies

$$X_t = \varepsilon_t + (\phi+\theta)\varepsilon_{t-1} + \phi(\phi+\theta)\varepsilon_{t-2} + \phi^2(\phi+\theta)\varepsilon_{t-3} + \cdots .$$

So the ARMA(1,1) model has a causal MA(∞) representation whenever the AR root condition holds.

Impulse-response coefficients of ARMA(1,1)

For causal ARMA(1,1),

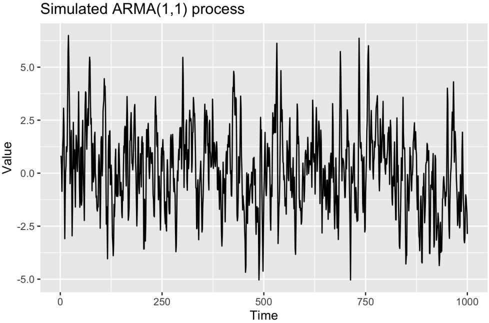
$$\psi_0 = 1, \qquad \psi_j = (\phi + \theta)\phi^{j-1}, \qquad j \geq 1.$$

- The impact effect is always one.
- The post-impact propagation depends jointly on ϕ and θ.
- The AR parameter controls the geometric decay rate after the first step.

How to interpret $\phi + \theta$ in ARMA(1,1)

- If $\phi + \theta$ is large in magnitude, the first few impulse weights are substantial.
- If $\theta \approx -\phi$, near-cancellation occurs and the series can look almost uncorrelated.
- So the MA parameter does not simply add "more persistence." It changes the short-run transmission of shocks.

A simulated ARMA(1,1) process



The process combines lagged-value feedback and lagged-shock transmission. This is why ARMA data can look smoother than MA data but less rigidly persistent than a pure AR with the same first-order correlation.

What the ARMA ACF usually looks like

- For a genuine ARMA model, the ACF usually tails off.
- The PACF also usually tails off.
- Neither side has the clean finite cut-off property that identifies pure AR or pure MA models.

This is why mixed models are theoretically elegant but empirically harder to identify.

Why ARMA identification is harder

- Different combinations of AR and MA coefficients can generate similar short-run correlation patterns.
- Near-cancellation can mask dynamic structure.
- Sample ACF and PACF plots are noisy finite-sample objects rather than perfect population signatures.

So ARMA identification usually combines theory, diagnostics, and information criteria rather than relying on one visual rule.

Causal representation and invertible representation

If both conditions hold, then the same ARMA process admits two useful infinite representations:

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad \varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

- The first is useful for dynamic interpretation and forecasting.
- The second is useful for identification and likelihood construction.

ARMA as a parsimonious approximation to Wold

Wold says a covariance-stationary process can be written as a one-sided innovation filter.

ARMA modeling says:

approximate that filter by a ratio of two finite polynomials.

This is the representation-theoretic reason ARMA models are so central.

Forecast intuition for ARMA models

- Forecasts are conditional expectations formed from the observed past.
- Under causality and invertibility, shocks can be updated recursively from past data.
- Multi-step forecasts gradually converge to the unconditional mean in stationary models.

The exact formulas come later, but the qualitative logic already follows from the ARMA structure.

Long-horizon forecasting in a stable ARMA

For a stationary ARMA process,

$$\mathrm{E}(Y_{t+h} \mid \mathcal{F}_t) \to \mu \qquad \text{as } h \to \infty.$$

- Stability means current information eventually fades.
- Forecasts lose the details of today's shocks and drift back toward the unconditional mean.

Again, the speed of that convergence is controlled by the roots.

What causality is not

Causality in ARMA theory does *not* mean causal inference in the policy-evaluation sense.

Here it simply means:

> X_t can be written as a stable function of current and past shocks only.

This terminology is classical in time-series analysis, but it is easy to misunderstand if one is also thinking about modern causal inference.

What invertibility is not

Invertibility does *not* mean that every observed realization can be inverted point by point in a deterministic algebraic sense.

Instead it means that the innovation sequence can be recovered through a stable linear filter:

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

This is a representation statement, not a literal finite-step algebraic inversion of the sample path.

A compact root-condition summary

Object	Polynomial	Desired root condition
AR(p) stability / causality	$\phi(z)$	all roots outside unit circle
MA(q) invertibility	$\theta(z)$	all roots outside unit circle
ARMA(p, q) causality	$\phi(z)$	all roots outside unit circle
ARMA(p, q) invertibility	$\theta(z)$	all roots outside unit circle

A compact AR–MA–ARMA comparison

Class	Dependence source	ACF pattern	Key condition
AR(p)	lagged values	tails off	AR roots
MA(q)	lagged shocks	cuts off	MA roots
ARMA(p, q)	both	usually tails off	both

Lecture Roadmap

1. Motivation and setup
2. AR(p) models: definitions and basic intuition
3. MA(q) models and invertibility
4. Lag polynomials and ARMA(p, q) models
5. R illustrations
6. Summary and wrap-up

Why simulation matters pedagogically

Simulation is not only for pictures.

- It lets us see how coefficients map into dynamic behavior.
- It reminds us what population formulas look like in finite samples.
- It helps students connect algebraic theory to sample paths, ACF plots, and estimation output.

In a 3-hour theory lecture, a few simulations are often worth many paragraphs of prose.

R illustration: simulating a stationary AR(1)

```
set.seed(1)
n <- 500
phi <- 0.8
eps <- rnorm(n)
x <- numeric(n)
for (t in 2:n) x[t] <- phi * x[t - 1] + eps[t]
plot(x, type = "l", main = "Stationary AR(1)")
acf(x, main = "Sample ACF")
pacf(x, main = "Sample PACF")
```

When $|\phi| < 1$, the series fluctuates around a stable center and the sample ACF decays gradually.

R illustration: solving Yule–Walker numerically

```
acf_vals <- acf(x, plot = FALSE, lag.max = 2)$acf
gamma0 <- acf_vals[1]
gamma1 <- acf_vals[2]
phi_hat <- gamma1 / gamma0
phi_hat
```

For AR(1), the Yule–Walker estimator is especially simple because $\rho_1 = \phi$ in population.

R illustration: simulating an MA(2)

```
simulate_ma2 <- function(n, theta1, theta2, sigma = 1) {
  eps <- rnorm(n + 2, sd = sigma)
  x <- numeric(n)
  for (t in 1:n) {
    x[t] <- eps[t + 2] + theta1 * eps[t + 1] + theta2 * eps[t]
  }
  x
}
```

The point is to see finite shock memory directly in the time path and in the ACF.

R illustration: sample behavior of MA(2)

```
set.seed(123)
x <- simulate_ma2(1000, theta1 = 0.5, theta2 = -0.3)
plot(x, type = "l", main = "Simulated MA(2)")
acf(x, main = "Sample ACF")
```

In population, the ACF cuts off after lag 2. In sample, later bars should be near zero rather than exactly zero.

R illustration: simulating an ARMA(1,1)

```
set.seed(42)
x <- arima.sim(model = list(ar = 0.7, ma = -0.4), n = 1000)
plot(x, type = "l", main = "Simulated ARMA(1,1)")
acf(x, main = "Sample ACF")
pacf(x, main = "Sample PACF")
```

For a genuine ARMA process, both the ACF and the PACF usually tail off.

R illustration: fitting a simple ARMA model

```
fit <- arima(x, order = c(1, 0, 1), include.mean = FALSE)
fit
```

Formal estimation, model comparison, residual diagnostics, and forecasting will be the main topic of Lecture 3. Today the code is mainly for dynamic intuition.

Lecture Roadmap

1. Motivation and setup
2. AR(p) models: definitions and basic intuition
3. MA(q) models and invertibility
4. Lag polynomials and ARMA(p, q) models
5. R illustrations
6. Summary and wrap-up

The lecture roadmap revisited

1. From Wold to finite-dimensional time-series models.
2. AR(p): roots, stationarity, causality, Yule–Walker.
3. MA(q): finite memory, ACF cut-off, invertibility.
4. ARMA(p, q): lag polynomials, mixed dynamics, two root conditions.

This is the conceptual backbone of classical univariate short-memory modeling.

Main takeaways from the AR block

- AR models propagate shocks through lagged values of the series.
- Stability is governed by the roots of the AR polynomial.
- A stable AR model has a causal MA(∞) representation.
- The Yule–Walker equations translate recursion parameters into autocovariance restrictions.

Main takeaways from the MA block

- MA models propagate shocks through a finite distributed lag of innovations.
- Finite-order MA models are automatically covariance stationary.
- Their ACF cuts off after the lag order.
- Invertibility is the key condition that makes the shock representation unique and operational.

Main takeaways from the ARMA block

- ARMA models combine lagged-value feedback and lagged-shock transmission.
- Causality is an AR-root condition.
- Invertibility is an MA-root condition.
- The mixed class is parsimonious precisely because it approximates the general Wold filter by a ratio of finite polynomials.

Three contrasts to remember

Concept	AR side	MA side
Dependence source	lagged values	lagged shocks
Key condition	causality / stationarity	invertibility
Signature in ACF	tails off	cuts off

Three mistakes students often make

1. Confusing white noise with i.i.d. noise.
2. Thinking that a large coefficient alone proves nonstationarity.
3. Treating the sample ACF as if it were the exact population ACF.

These mistakes disappear once the root conditions and the representation logic are understood clearly.

A board plan for the instructor

Write the following items carefully on the board:

- the centered AR(p) recursion;
- the AR root condition and AR(1) geometric-series inverse;
- the Yule–Walker derivation for AR(1);
- the MA(1) autocorrelation formula;
- the definitions of causality and invertibility;
- the ARMA(p, q) lag-polynomial identity.

Questions for review before Lecture 3

1. Why does a stationary AR(1) have an MA(∞) representation?
2. Why is the ACF cut-off property associated with MA models rather than AR models?
3. Why does invertibility matter even though a finite-order MA model is already stationary?
4. Why is ARMA identification harder than pure AR or pure MA identification?

Reading for Lecture 3

Review the following until they can be stated without notes:

- AR and MA definitions in centered form;
- the two root conditions;
- the AR(1) variance and autocorrelation formulas;
- the Yule–Walker idea;
- the difference between causality and invertibility.

Next time we move to empirical identification, estimation, and forecasting.

Final slide: the lecture in one sentence

AR, MA, and ARMA models are the classical finite-dimensional language for describing how stationary shocks propagate, decay, and can be recovered from the data.