Lecture 6 — Vector Autoregressive Models: Notation, Stability, Estimation, Lag Selection, and Granger Causality

Logical continuation after Lectures 1–5: from univariate dynamics to multivariate systems

Jiajing Sun

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

Econometrics and Time Series Methods
Spring 2026

Why VAR is the right next step

I agree that the most logical move after Lectures 1–5 is to start the **multivariate linear time-series** block.

- Lectures 1–3 established the univariate language: stationarity, Wold, ARMA structure, identification, estimation, diagnostics, and forecasting.
- Lectures 4–5 covered nonstationarity, deterministic trend, random walk logic, and empirical interpretation in finance.
- The natural next question is no longer *how one series behaves on its own*, but *how several series move together over time*.
- That is exactly the role of the **vector autoregression**, or VAR.

How these two lectures are organized

Lecture 6 focuses on the core VAR machinery: notation, stability, estimation, lag choice, forecasting, and Granger causality.
Lecture 7 then builds the dynamic-interpretation layer: VMA/VARMA representations, impulse responses, orthogonalization, and inference.

Textbook sequence for Lectures 6 and 7

To keep the sequence coherent and non-repetitive with Lectures 1–5, I follow the chapter logic in this order:

1. **Lecture 6:** Vector autoregressive model
   - VAR notation and lag-polynomial form,
   - companion form and stability,
   - OLS / Gaussian likelihood estimation,
   - lag-order selection and forecasting,
   - Granger causality and interpretation.
2. **Lecture 7:** Dynamic representations and responses
   - VMA and VARMA representations,
   - impulse response functions,
   - orthogonalized shocks,
   - inference and practical interpretation.

Important boundary

I deliberately stop short of cointegration and VECM here. That material fits best in the next lecture, once the stationary VAR framework is fully in place.

Learning goals for this three-hour lecture

By the end of Lecture 6, students should be able to:

1. write down a VAR(p) in vector, componentwise, lag-polynomial, and companion-form notation;
2. explain the stability condition in terms of eigenvalues and characteristic roots;
3. understand why equation-by-equation OLS works for the reduced-form VAR;
4. choose a lag order using information criteria, likelihood-ratio logic, and residual checks;
5. distinguish forecasting usefulness from structural causality when interpreting Granger-causality results.

Practical plan for the three contact hours

Hour 1

Why multivariate models are needed; VAR notation, lag-polynomial form, mean, companion form, and stability.

Hour 2

Estimation of VAR models, Gaussian likelihood, lag-order selection, residual checking, and forecasting.

Hour 3

Granger causality, predictive interpretation, block exogeneity, and the limits of causal language.

Teaching emphasis

The main theme is that a VAR is a **system for organized dependence**: it captures own lags, cross lags, and contemporaneous shock correlation inside one coherent framework.

Why univariate models are not enough anymore

A univariate ARMA model is useful when one series can be studied in relative isolation. But much of economics and finance is inherently systemic.

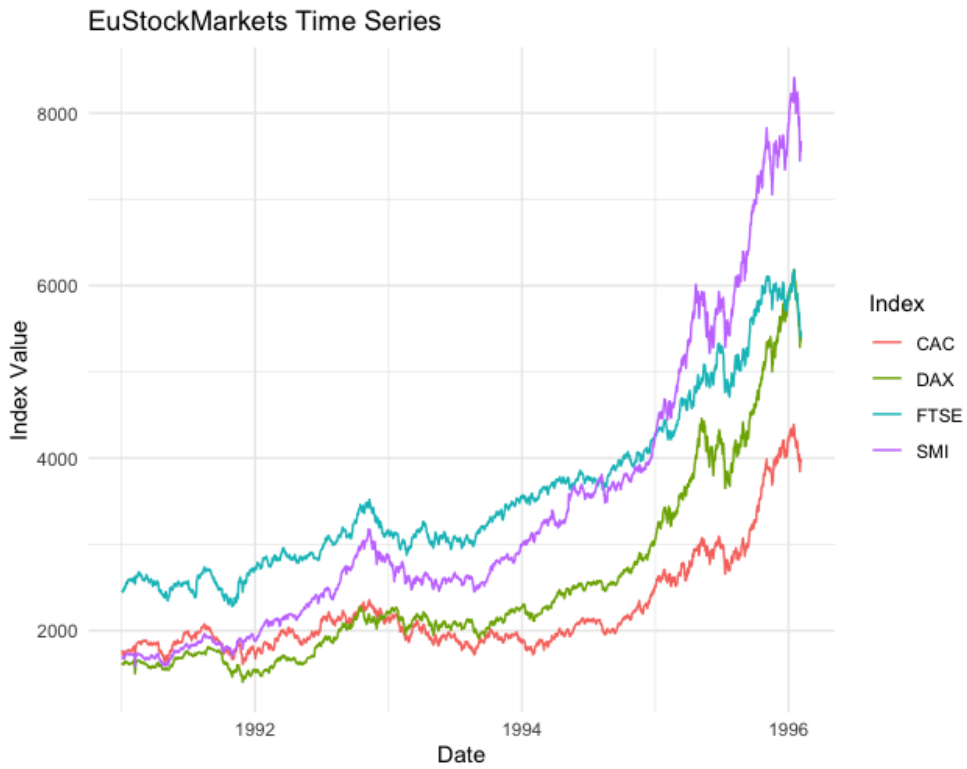
- Output, inflation, and interest rates move together.
- Equity markets in different countries co-move and react to common shocks.
- Stock returns, bond yields, exchange rates, and commodity prices often transmit information across markets.
- Policy analysis usually asks how one variable responds when another is disturbed.

Textbook point

Multivariate time-series models extend the linear-process logic of the univariate case to a setting where several variables influence each other dynamically.

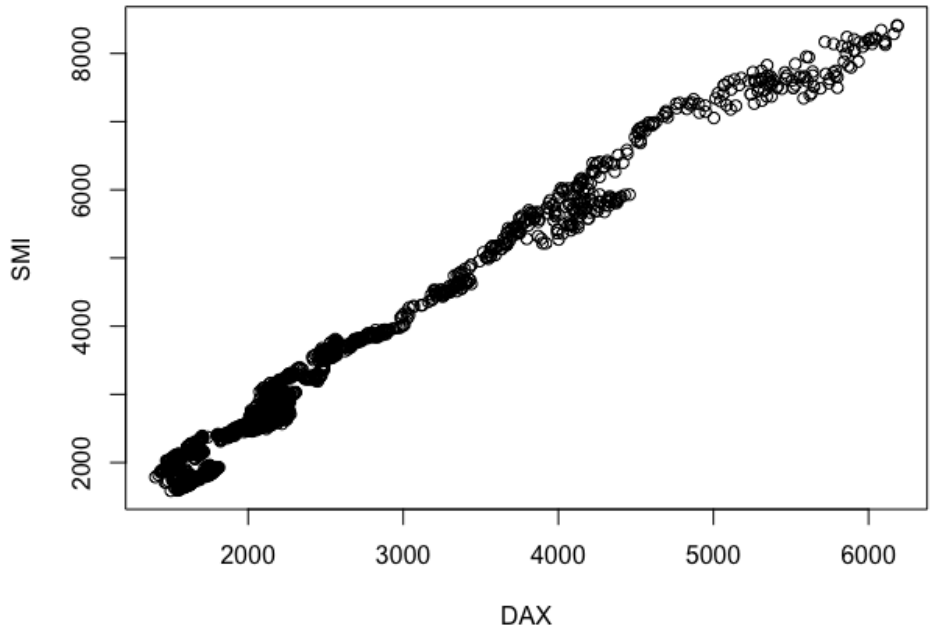
A motivating picture: European equity indices

These market indices do not move independently. They show common long swings, joint crises, and country-specific deviations.



Cross-sectional association is not enough

A scatter plot reveals strong comovement, but it hides timing.



Key limitation

Cross-sectional association does not identify lead-lag structure, feedback, or shock transmission.

Two forms of dependence inside a multivariate system

The textbook emphasizes that a VAR captures **two** conceptually different types of dependence.

Contemporaneous dependence

This comes through the innovation covariance matrix

$$\Omega = \mathrm{E}(\varepsilon_t \varepsilon_t').$$

Different shocks may be correlated at the same date.

Dynamic dependence

This comes through the autoregressive coefficient matrices

$$\Phi_1, \ldots, \Phi_p.$$

Lagged values of one variable can predict another variable.

Reduced-form interpretation

A reduced-form VAR does not automatically tell us which shock is structural; it describes how the observables co-evolve.

Notation for an n-dimensional time series

Let

$$\mathrm{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} \in \mathbb{R}^n.$$

A multivariate time series is a sequence $\{\mathrm{y}_t\}$ indexed by time.

- Each date produces a **vector observation**, not a scalar observation.
- The model must allow each component to depend on its own past *and* the past of the other components.
- The innovation at date t is also a vector:

$$\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})'.$$

Definition of a VAR(p)

The chapter begins with the reduced-form vector autoregression of order p:

$$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \varepsilon_t,$$

where

- c is an $(n \times 1)$ intercept vector,
- Φ_j is an $(n \times n)$ matrix for each lag j,
- ε_t is an $(n \times 1)$ innovation vector with

$$E(\varepsilon_t) = 0, \qquad E(\varepsilon_t \varepsilon_t') = \Omega,$$

and typically $\{\varepsilon_t\}$ is taken to be i.i.d. or at least a martingale difference sequence.

Component-by-component view of a VAR

The first equation of the VAR(p) can be written as

$$y_{1t} = c_1 + \sum_{j=1}^{n} (\Phi_1)_{1j} y_{j,t-1} + \cdots + \sum_{j=1}^{n} (\Phi_p)_{1j} y_{j,t-p} + \varepsilon_{1t}.$$

Likewise for $y_{2t}, \ldots, y_{nt}$.

- Every equation uses the same set of regressors: lags of *all* variables in the system.
- Own lags describe persistence within a variable.
- Cross lags describe predictive interactions across variables.
- This "everything depends on everything lagged" structure is what makes the reduced-form VAR so flexible.

Lag-polynomial representation

Define the matrix lag polynomial

$$\Phi(L) = I_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p.$$

Then the VAR is written compactly as

$$\Phi(L)\mathrm{y}_t = \mathrm{c} + \varepsilon_t.$$

- This is the direct multivariate analogue of the univariate AR polynomial.
- The determinant
$$\det\left(I_n - \Phi_1 z - \cdots - \Phi_p z^p\right)$$
plays the role of the characteristic polynomial.
- Root locations determine whether the model admits a stable one-sided representation.

Mean of a stable VAR

If the VAR is covariance stationary and has unconditional mean $\mu = \mathrm{E}(\mathrm{y}_t)$, then taking expectations gives

$$\mu = \mathrm{c} + \Phi_1 \mu + \cdots + \Phi_p \mu,$$

so that

$$\mu = \left(I_n - \Phi_1 - \cdots - \Phi_p\right)^{-1} \mathrm{c},$$

provided the inverse exists.

Centering the system

If we define $\tilde{\mathrm{y}}_t = \mathrm{y}_t - \mu$, then the centered VAR is

$$\tilde{\mathrm{y}}_t = \Phi_1 \tilde{\mathrm{y}}_{t-1} + \cdots + \Phi_p \tilde{\mathrm{y}}_{t-p} + \varepsilon_t.$$

This is often the cleanest form for theoretical derivations.

Standardization is possible, but it changes the coefficient matrices

If $\Sigma_y = \mathrm{Var}(\mathrm{y}_t)$ is nonsingular, define the standardized process

$$\mathrm{y}_t^* = \Sigma_y^{-1/2}(\mathrm{y}_t - \mu).$$

Then y_t^* is still a VAR(p), with transformed matrices

$$\Phi_j^* = \Sigma_y^{-1/2}\Phi_j\Sigma_y^{1/2}, \qquad j = 1, \ldots, p,$$

and transformed innovation covariance

$$\Omega^* = \Sigma_y^{-1/2}\Omega\Sigma_y^{-1/2}.$$

- Standardization can be useful for interpretation and comparison across variables.
- But the dynamic content remains the same up to a linear re-scaling.

Why stability matters

A stable VAR is the multivariate analogue of a stationary AR model.

- Shocks have effects that eventually die out.
- Unconditional moments exist and do not depend on calendar time.
- Forecasts converge back toward the unconditional mean rather than exploding.
- The model has a convergent VMA(∞) representation, which we will need for Lecture 7.

Economic reading

If a stationary VAR is hit by a temporary innovation, the system may react strongly, but it does not drift forever simply because of that one shock.

Companion form: turning VAR(p) into VAR(1)

For the centered process $\tilde{y}_t = y_t - \mu$, define

$$\xi_t = \begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{t-1} \\ \vdots \\ \tilde{y}_{t-p+1} \end{bmatrix}, \qquad \xi_t = F\xi_{t-1} + v_t, \qquad J = [I_n \quad 0 \quad \cdots \quad 0].$$

With

$$v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \tilde{y}_t = J\xi_t,$$

we obtain the one-sided representation

$$\tilde{y}_t = \sum_{s=0}^{\infty} JF^s J' \varepsilon_{t-s}$$

whenever $F^s \to 0$.

Use later

The same matrices $JF^s J'$ become the impulse-response matrices in Lecture 7.

The companion matrix explicitly

For a VAR(p), the companion matrix is

$$F = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \cdots & 0 & 0 \\ 0 & I_n & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_n & 0 \end{bmatrix}.$$

The corresponding innovation block vector is

$$v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- The top block row carries the economic dynamics.
- The lower block rows simply shift lags down the state vector.

Stability condition in terms of eigenvalues

If all eigenvalues of the companion matrix F lie strictly inside the unit circle, then

$$F^s \to 0 \qquad \text{as } s \to \infty,$$

and the effect of a shock eventually disappears. From the state recursion,

$$\xi_{t+s} = v_{t+s} + Fv_{t+s-1} + \cdots + F^{s-1}v_{t+1} + F^s\xi_t.$$

So the term $F^s\xi_t$ vanishes only when those eigenvalues have modulus less than one.

Operational rule

For a stationary VAR, **all eigenvalues of the companion matrix must lie inside the unit circle**.

Equivalent root condition in lag-polynomial form

The eigenvalue condition is equivalent to the root condition

$$\det\left(I_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right) = 0.$$

The VAR(p) is covariance stationary when **all roots z of this determinant equation lie outside the unit circle**.

- This is the exact multivariate analogue of the root condition for univariate AR models.
- In software, the usual stability check is implemented via the eigenvalues of the companion matrix.
- In lecture notes and proofs, the determinant form is often more compact.

Dynamic intuition behind the stability condition

Suppose one component of ε_t receives a shock at date t.

- The current effect is immediate through the first block of the state vector.
- At horizon $t+1$, the response is shaped by F.
- At horizon $t+2$, it is shaped by F^2.
- More generally, F^h controls how the shock propagates h steps ahead.

Interpretation

If powers of F decay, responses shrink. If powers of F do not decay, responses persist indefinitely or even grow. That is why the stability check is the first non-negotiable step in any VAR analysis.

The VAR(1) case is the workhorse benchmark

For

$$y_t = c + Ay_{t-1} + \varepsilon_t,$$

we obtain after centering:

$$y_t - \mu = A(y_{t-1} - \mu) + \varepsilon_t, \qquad \mu = (I_n - A)^{-1}c.$$

Iterating backward gives

$$y_t = \mu + \varepsilon_t + A\varepsilon_{t-1} + A^2\varepsilon_{t-2} + \cdots,$$

provided all eigenvalues of A lie inside the unit circle.

Why VAR(1) matters

Many intuition-building results are easiest to see in the VAR(1) case; the general VAR(p) then follows by companion-form logic.

The VMA(∞) representation already appears here

Under stability,

$$y_t - \mu = \sum_{j=0}^{\infty} A^j \varepsilon_{t-j}.$$

This is a vector moving-average representation of infinite order.

- The coefficient matrix on ε_{t-j} is A^j.
- So A^j summarizes how a shock at date $t-j$ affects the system today.
- Lecture 7 will generalize this to the sequence of impulse-response matrices Ψ_j for a general VAR(p).

Autocovariance matrices in a VAR(1)

Let

$$\Gamma(k) = \mathrm{E}\big[(\mathrm{y}_t - \mu)(\mathrm{y}_{t+k} - \mu)'\big].$$

Then for a stationary VAR(1),

$$\Gamma(0) = A\Gamma(0)A' + \Omega.$$

This is the multivariate analogue of the univariate variance formula. Iterating yields

$$\Gamma(0) = \sum_{s=0}^{\infty} A^s \Omega (A^s)'.$$

Also,

$$\Gamma(k) = A^k \Gamma(0), \qquad k \geq 0,$$

for the forward-lag convention used here.

The matrix Yule–Walker viewpoint

For a centered VAR(p), the autocovariance matrices satisfy the multivariate Yule–Walker relations

$$\Gamma(k) = \Phi_1\Gamma(k-1) + \cdots + \Phi_p\Gamma(k-p), \qquad k = 1, \ldots, p.$$

- These are matrix equations, not scalar equations.
- They show how second moments encode the autoregressive structure.
- They can be used for estimation, although in empirical work equation-by-equation OLS is usually simpler.

Connection to earlier lectures

This is the direct multivariate extension of the Yule–Walker logic you saw for univariate AR processes.

What happens when stability fails?

Failure of the stability condition can reflect several distinct cases:

- **Unit roots:** some characteristic roots lie on the unit circle.
- **Explosiveness:** some roots lie inside the unit circle in z-space, or equivalently eigenvalues of F exceed one in modulus.
- **Near-instability:** roots are close to the boundary, producing very persistent but still stationary dynamics.

Why we stop here today

Once the system is nonstationary, we need the separate theory of nonstationary VARs, cointegration, and VECMs. That is the next block, not today's one.

How many parameters are we trying to estimate?

A VAR can become large very quickly. For n variables and lag order p:

- there are $n(np+1)$ regression coefficients including intercepts,
- and the innovation covariance matrix contributes

$$\frac{n(n+1)}{2}$$

  free parameters.

So the total parameter count is

$$n(np+1) + \frac{n(n+1)}{2}.$$

Implication

The curse of dimensionality enters much earlier in VAR work than in univariate ARMA work. This is one reason lag selection and parsimony matter so much.

Each VAR equation is a regression

Define the regressor vector

$$x_t = \begin{bmatrix} 1 \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix},$$

and collect coefficients in

$$\Pi' = \begin{bmatrix} c & \Phi_1 & \Phi_2 & \cdots & \Phi_p \end{bmatrix}.$$

Then the VAR becomes

$$y_t = \Pi' x_t + \varepsilon_t.$$

For the ith component,

$$y_{it} = x_t' \beta_i + \varepsilon_{it}.$$

So estimation can be carried out equation by equation using least squares.

Why OLS works in the reduced-form VAR

The regressors are stochastic and endogenous in the broad economic sense, but the reduced-form assumption implies

$$\mathrm{E}(\varepsilon_t \mid \mathrm{y}_{t-1}, \mathrm{y}_{t-2}, \ldots) = 0.$$

That is enough for consistency of OLS in each equation.

- The same lagged variables appear in each equation.
- The error terms may be correlated across equations through Ω.
- But because the regressor matrix is identical across equations, GLS and OLS coincide in the Gaussian case.

Practical consequence

The reduced-form VAR is one of the most computationally convenient multivariate models in econometrics.

System form and SUR equivalence

Stack the observations into matrices:

$$Y = XB + E,$$

where

- Y collects the dependent variables across equations,
- X is the common lagged-regressor matrix,
- B stacks the coefficient vectors,
- E collects the innovations.

This is a seemingly unrelated regressions system with **identical regressors across equations**. In that case,

$$\widehat{B}_{\mathrm{GLS}} = \widehat{B}_{\mathrm{OLS}}.$$

Interpretation

Cross-equation correlation matters for covariance estimation and structural interpretation, but not for the reduced-form point estimates when all equations use the same regressors.

The OLS estimator

For a single equation,

$$\widehat{\beta}_i = (X'X)^{-1}X'Y_i.$$

In stacked matrix form,

$$\widehat{B} = (X'X)^{-1}X'Y.$$

Equivalently, with the notation used in the textbook,

$$\widehat{\Pi}' = \left[\sum_{t=1}^{T} y_t x_t'\right]\left[\sum_{t=1}^{T} x_t x_t'\right]^{-1}.$$

- Once $\widehat{\Pi}$ is obtained, we can recover $\widehat{c}$ and each $\widehat{\Phi}_j$ from its blocks.
- Software such as `vars::VAR()` does exactly this, though with convenient wrappers for diagnostics and inference.

Estimating the innovation covariance matrix

Residuals are

$$\widehat{\varepsilon}_t = y_t - \widehat{c} - \widehat{\Phi}_1 y_{t-1} - \cdots - \widehat{\Phi}_p y_{t-p}.$$

The innovation covariance matrix is estimated by

$$\widehat{\Omega} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_t'.$$

- Diagonal elements measure innovation variances of each equation.
- Off-diagonal elements measure contemporaneous covariance between reduced-form shocks.
- These off-diagonal terms are precisely why orthogonalization becomes necessary for IRF interpretation in Lecture 7.

Gaussian conditional likelihood

If the innovations are Gaussian,

$$\varepsilon_t \sim \text{i.i.d. } N(0, \Omega),$$

then conditioning on the first p observations gives

$$\mathcal{L}(\theta) = -\frac{(T-p)n}{2}\log(2\pi) - \frac{T-p}{2}\log|\Omega| - \frac{1}{2}\sum_{t=p+1}^{T}(y_t - \Pi' x_t)'\Omega^{-1}(y_t - \Pi' x_t).$$

Maximizing over Π for fixed Ω is a quadratic least-squares problem, so

$$\widehat{\Pi}_{\text{MLE}} = \widehat{\Pi}_{\text{OLS}}.$$

- The Gaussian likelihood is useful for AIC, BIC, LR tests, and asymptotic theory.
- But the reduced-form point estimates are still the familiar OLS estimates.

Why Gaussian MLE and OLS coincide

The equivalence between OLS and Gaussian conditional MLE is not accidental.

- The Gaussian likelihood is built from quadratic forms in the residuals.
- OLS minimizes the sum of squared residuals equation by equation.
- Because all equations use the same regressor matrix, the system maximum-likelihood problem collapses neatly into the same least-squares estimator for the coefficient matrices.

Good news for practice

You get a likelihood-based framework for information criteria and hypothesis testing without needing a numerically difficult nonlinear optimization for the reduced-form VAR.

Large-sample distribution of the coefficient estimates

Under stability and suitable moment conditions,

$$\sqrt{T}\big(\operatorname{vec}(\widehat{\Phi}) - \operatorname{vec}(\Phi)\big) \Longrightarrow N\big(0,\, \Gamma(0)^{-1} \otimes \Omega\big).$$

Equivalently,

$$\sqrt{T}(\widehat{\Phi} - \Phi) \Longrightarrow MN_{n\times np}\big(0, \Omega, \Gamma(0)^{-1}\big),$$

where MN denotes a matrix-normal limit law. For the intercept vector,

$$\sqrt{T}(\widehat{c} - c) \Longrightarrow N\big(0, \tau\Omega\big), \qquad \tau = 1 + \mu'\Gamma(0)^{-1}\mu.$$

- The Kronecker-product covariance structure is the natural multivariate extension of regression asymptotics.
- These formulas deliver Wald tests and asymptotic standard errors for blocks of coefficients.

A robust-inference remark

If the innovations are only a martingale difference sequence, OLS can still be consistent, but the covariance matrix changes. Define

$$\widehat{\Upsilon} = (\widehat{\Gamma}(0) \otimes I_n)^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} \left((y_{t-1} - \bar{y})(y_{t-1} - \bar{y})' \right) \otimes \widehat{\varepsilon}_t \widehat{\varepsilon}_t' \right] (\widehat{\Gamma}(0) \otimes I_n)^{-1}.$$

Then robust Wald-type inference can be based on $\widehat{\Upsilon}$ instead of the i.i.d.-Gaussian covariance formula.

Practical reading

Reduced-form VAR estimation is easy. Reliable inference is easy only when the shock process is well behaved. With conditional heteroskedasticity, use sandwich-style robust covariance matrices.

Multivariate Yule–Walker as an alternative estimator

For a centered VAR(p), the matrix Yule–Walker equations are

$$\Gamma(k) = \Phi_1\Gamma(k-1) + \cdots + \Phi_p\Gamma(k-p), \qquad k = 1, \ldots, p.$$

After vectorization, these become a linear system in the unknown coefficients.

- This generalizes the univariate Yule–Walker method from Lecture 2.
- It is theoretically useful because it ties moments directly to dynamic parameters.
- In applied work, OLS is generally preferred because it is simpler and integrates naturally with model selection, diagnostics, and forecasting.

Why lag-order selection is a central issue

Choosing p is not a minor technical detail.

- If p is too small, the model omits relevant dynamic structure and leaves serial correlation in the residuals.
- If p is too large, the model consumes degrees of freedom and can overfit badly.
- In a multivariate system, overfitting is especially costly because each extra lag adds an entire *matrix* of coefficients.

Textbook message

The chosen lag length should balance fit, parsimony, and residual adequacy.

Information criteria for choosing p

Let $\widehat{\mathcal{L}}$ be the maximized Gaussian log-likelihood for a candidate lag order p. Then common criteria take the form

$$\mathrm{AIC}(p) = -2\widehat{\mathcal{L}} + 2d(p, n),$$

$$\mathrm{BIC}(p) = -2\widehat{\mathcal{L}} + d(p, n) \log T,$$

with similar logic for HQIC.

- $d(p, n)$ is the total number of free parameters.
- Smaller values are preferred.
- In practice, `vars::VARselect()` computes these automatically over a set of candidate lag lengths.

How AIC, BIC, and HQIC differ

All three criteria trade off fit against complexity, but the penalty strength differs.

- **AIC** penalizes additional parameters relatively lightly, so it often chooses a richer model.
- **BIC** penalizes more heavily, so it is typically more parsimonious.
- **HQIC** sits between them.

No universal winner

In time-series practice it is common to compare the recommendations of several criteria, check residual adequacy, and then make a reasoned empirical choice rather than blindly following one number.

Likelihood-ratio testing for lag order

Suppose we test a restricted VAR with p_0 lags against an unrestricted VAR with $p_1 > p_0$ lags. Let $\widehat{\Omega}_0$ and $\widehat{\Omega}_1$ be the residual covariance matrices under the two models. The likelihood-ratio statistic is

$$\mathrm{LR} = T \left\{ \log |\widehat{\Omega}_0| - \log |\widehat{\Omega}_1| \right\}.$$

Under the null, this is asymptotically χ^2 with degrees of freedom equal to the number of imposed zero restrictions.

- Information criteria compare many models at once.
- LR tests compare a smaller model against a larger nested model.

A practical lag-selection workflow

A sensible empirical workflow is:

1. choose a reasonable maximum lag $p_{\max}$;
2. compare AIC, BIC, and HQIC across $p = 0, 1, \ldots, p_{\max}$;
3. test down from a larger lag order if appropriate using LR tests;
4. estimate the candidate VAR and inspect residual serial correlation;
5. prefer the most parsimonious model that leaves residuals approximately white noise.

What not to do

Do not pick the lag length only because it produces a "nice" Granger-causality result or an attractive impulse response later on.

Residual diagnostics still matter after estimation

Even after choosing p, you should check whether the fitted model is dynamically adequate.

- inspect residual autocorrelation and cross-correlation,
- look for obvious conditional heteroskedasticity or outliers,
- verify stability of the estimated system,
- confirm that coefficient estimates are economically sensible and not purely driven by overfitting.

Reduced-form adequacy first

Impulse responses, variance decompositions, and causality tests are only as credible as the underlying fitted VAR.

Forecasting with a VAR

One-step-ahead forecasts are

$$\widehat{y}_{t+1|t} = \widehat{c} + \widehat{\Phi}_1 y_t + \cdots + \widehat{\Phi}_p y_{t-p+1}.$$

For the centered companion state, multi-step forecasts satisfy

$$\widehat{\xi}_{t+h|t} = F^h \xi_t, \qquad \widehat{y}_{t+h|t} = \mu + J F^h \xi_t.$$

The h-step forecast-error covariance matrix is

$$\mathrm{MSFE}(h) = \sum_{j=0}^{h-1} J F^j Q (F^j)' J'.$$

For a VAR(1), this reduces to

$$\mathrm{MSFE}(h) = \sum_{j=0}^{h-1} A^j \Omega (A^j)'.$$

Bridge to Granger causality

If lagged values of one variable materially reduce forecast error variance for another variable, that is exactly the predictive content summarized by Granger causality.

Why Granger causality belongs in a VAR lecture

The Granger-causality idea is inseparable from the VAR framework because it is fundamentally about **whether lagged variables improve prediction**.

- The test is not about philosophical causation.
- It is not about contemporaneous structural channels either.
- It is about whether the history of one variable helps forecast another once the latter's own history is already included.

In one line

"y Granger-causes x" means that lagged values of y contain predictive information about future x beyond what lagged values of x already contain.

Textbook definition in mean-squared-error language

If y does *not* Granger-cause x, then for every horizon $s > 0$,

$$\mathrm{MSE}\Big[\mathrm{E}(x_{t+s} \mid x_t, x_{t-1}, \ldots)\Big] = \mathrm{MSE}\Big[\mathrm{E}(x_{t+s} \mid x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)\Big].$$

If including the history of y lowers the forecast mean-squared error, then y is said to Granger-cause x.

Interpretation

This is a forecasting definition. It is entirely operational: compare predictive ability with and without the lagged information set of another variable.

Bivariate VAR and the lower-triangular restriction

Consider a bivariate VAR(p) in $(x_t, y_t)'$. If all lag matrices are lower triangular,

$$\Phi_j = \begin{bmatrix} \phi_{11}^{(j)} & 0 \\ \phi_{21}^{(j)} & \phi_{22}^{(j)} \end{bmatrix}, \qquad j = 1, \ldots, p,$$

then lagged values of y do not appear in the x equation.

- In that case, y does not help predict x.
- Therefore y is not a Granger cause of x.

Restriction form

Testing non-causality becomes a test of whether a block of cross-lag coefficients is jointly equal to zero.

Restricted versus unrestricted models

To test whether y fails to Granger-cause x, compare:

- a **restricted** model in which all coefficients on lagged y in the x equation are set to zero;
- an **unrestricted** model in which those coefficients are freely estimated.

The null hypothesis is

$$H_0 : \phi_{12}^{(1)} = \phi_{12}^{(2)} = \cdots = \phi_{12}^{(p)} = 0.$$

Logic

If relaxing those restrictions materially improves fit or forecast performance, the data support the claim that y Granger-causes x.

Wald, LM, LR, and F-style testing logic

Write the restrictions as

$$H_0: \quad R\,\mathrm{vec}(\Phi) = r.$$

A Wald statistic is

$$W = T(R\widehat{a} - r)'\left[R\left(\widehat{\Gamma}(0)^{-1} \otimes \widehat{\Omega}\right)R'\right]^{-1}(R\widehat{a} - r) \overset{a}{\sim} \chi^2_q.$$

A likelihood-ratio statistic compares restricted and unrestricted covariance estimates:

$$LR = T\left(\log\det(\widetilde{\Omega}) - \log\det(\widehat{\Omega})\right) \overset{a}{\sim} \chi^2_q.$$

For a single equation with q zero restrictions, the familiar finite-sample form is

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(T - k_u)}.$$

- Wald: estimate the unrestricted model once.
- LR: compare restricted and unrestricted fit.
- LM / score: evaluate the restricted model only.

Block exogeneity in larger systems

In a system with many variables, Granger non-causality is often tested in block form.

- Example: do all lags of foreign variables fail to help predict domestic output?
- Example: do all lags of financial variables fail to help predict inflation?

This becomes a joint restriction on several rows and columns across the coefficient matrices.

Useful term

When a block of variables does not help forecast another block once its own history is controlled for, we often say the first block is **Granger non-causal** or **block exogenous** for the second block.

Typical economic applications of Granger causality

The predictive-causality question appears everywhere in applied economics and finance.

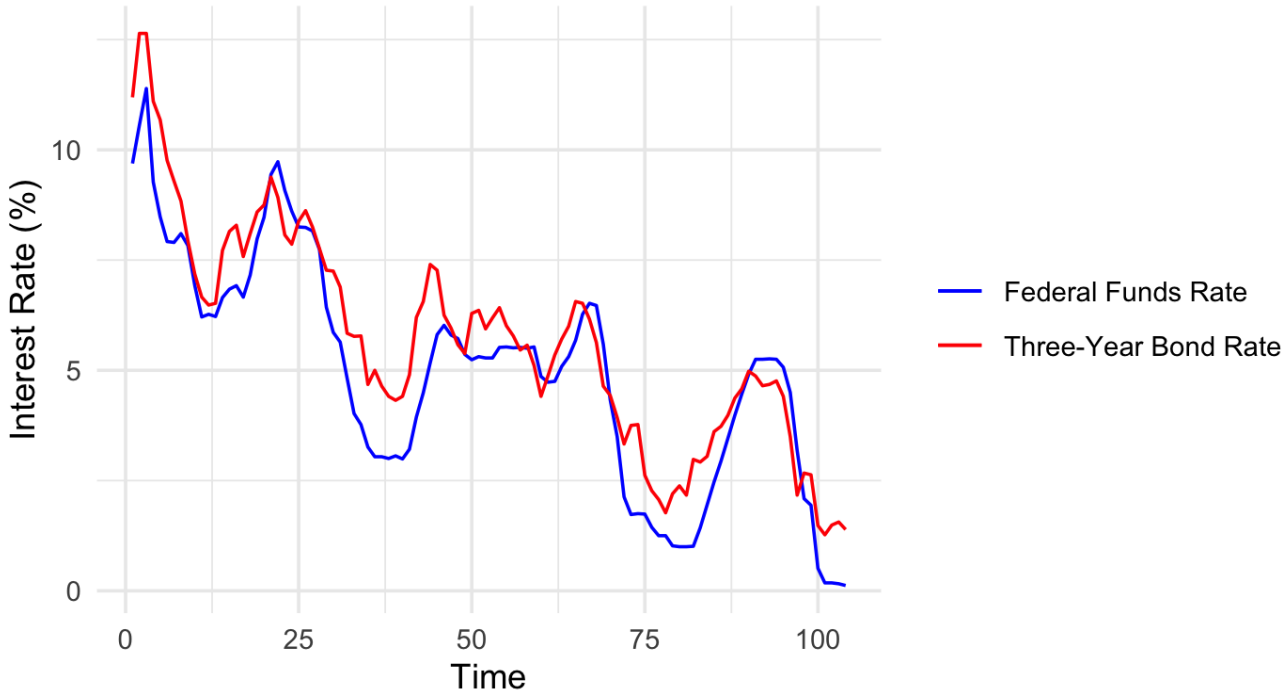
- Do short-term rates help predict longer-term rates?
- Do stock returns help forecast dividend growth or macro indicators?
- Do exchange-rate movements help forecast trade flows or inflation?
- Do foreign market returns help forecast domestic returns?

Important discipline

The test answers a narrowly defined forecasting question. Economic theory is still needed to interpret why the predictive content appears.

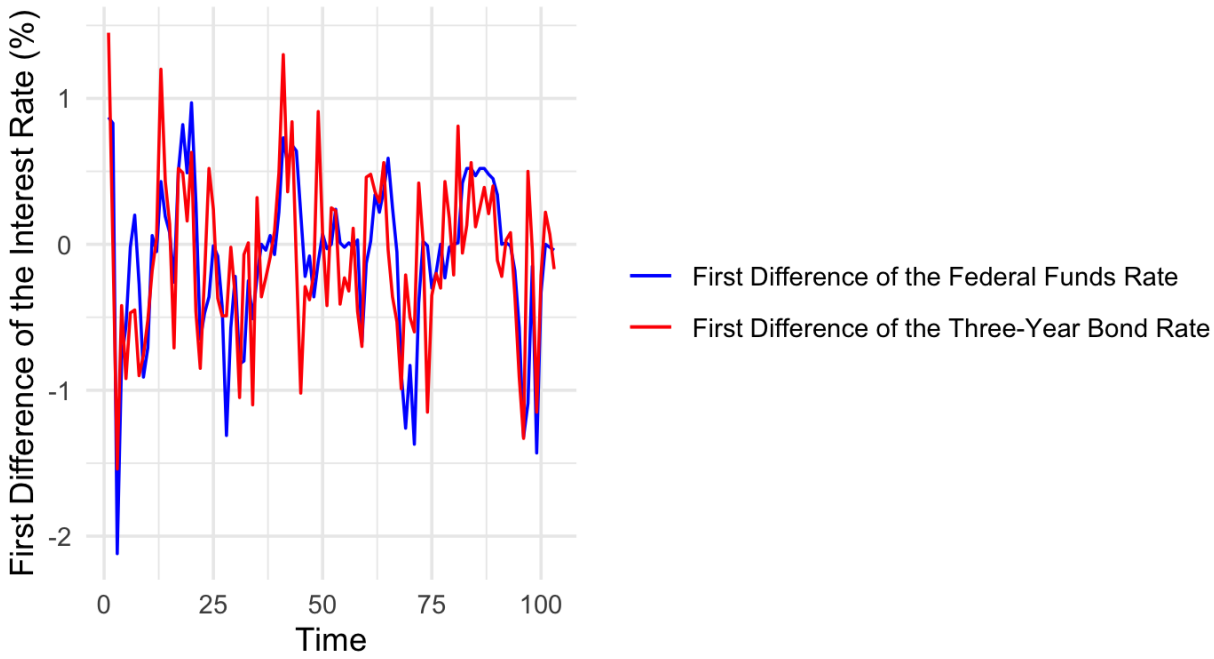
Interest rates as a predictive system

A natural bivariate example is the joint evolution of a policy-related short rate and a medium-term bond yield.



Why predictive tests often use differences

Persistent level variation can dominate predictive regressions. Differencing often isolates shorter-run movements.



Practical message

Predictive and Granger-causality results may differ sharply between levels and first differences when variables are highly persistent.

Granger causality is not the same as true causality

This warning is central and should not be treated as a minor footnote.

- If x predicts y, that does not prove that shocks to x are the structural cause of y.
- A hidden common driver may cause both.
- Forward-looking behavior may make one variable contain information about another variable's future without being its underlying cause.

Safe language

Say "contains predictive information for" or "Granger-causes" rather than "causes" unless a structural identification argument is also available.

Hamilton's classic warning: stock prices and dividends

A well-known example is that stock prices may Granger-cause dividends in the data, even though expected future dividends are part of the economic value of stocks.

- Markets are forward-looking.
- Prices can move today because traders anticipate future dividends.
- So prices may help forecast dividends without being the deep structural cause of dividend payments.

Lesson

Predictive priority and economic causality are not the same thing.

L"utkepohl's transformation example

Even absence of Granger causality need not mean absence of any causal channel. Start with a lower-triangular VAR(1):

$$\begin{bmatrix} z_t \\ x_t \end{bmatrix} = \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} z_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}.$$

Now premultiply the system by a nonsingular matrix

$$B = \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix}.$$

The transformed representation can display contemporaneous influence terms even though the original reduced-form Granger non-causality statement has not changed.

Moral

Reduced-form predictive restrictions are representation-dependent and should not be over-interpreted as deep structural statements.

Checklist for an empirical VAR project

A disciplined reduced-form VAR workflow usually looks like this:

1. choose the variables and transformation carefully;
2. decide whether the stationary VAR framework is appropriate;
3. set a reasonable lag search range;
4. estimate candidate VARs and check stability;
5. choose lag order using information criteria and residual diagnostics;
6. inspect residual covariance and coefficient plausibility;
7. only then move on to Granger causality, impulse responses, or variance decompositions.

Common pitfalls students make at this stage

- treating correlated reduced-form shocks as if they were structural shocks;
- reading one significant coefficient as proof of an economically meaningful effect;
- ignoring lag-order uncertainty;
- fitting a stationary VAR to nonstationary data without checking the integration properties;
- forgetting how fast the parameter count grows with system size.

Good habit

Running a VAR is easy. Credible analysis lives in specification, interpretation, and diagnostic discipline.

What Lecture 7 adds to today's framework

Today's lecture estimated the reduced-form system and discussed predictive content.
The next lecture asks a different question:

How does a shock propagate through the system over time?

To answer that, we need

- the VMA(∞) representation,
- impulse-response matrices,
- orthogonalization of contemporaneously correlated shocks,
- confidence intervals or bootstrap bands for the estimated responses.

Lecture 6 summary

- A VAR(p) is the multivariate extension of the autoregressive model, with coefficient *matrices* rather than coefficients.
- Stability is checked through the companion matrix or, equivalently, through the roots of the matrix lag polynomial.
- In the reduced-form stationary VAR, equation-by-equation OLS is the core estimator, and under Gaussianity it coincides with conditional MLE.
- Lag-order choice must balance fit and parsimony, and it should always be backed up by residual diagnostics.
- Granger causality is a predictive concept, not a structural-causal conclusion.

Take-away

Lecture 6 gives us the **estimated system**. Lecture 7 will turn that system into dynamic economic stories via impulse responses.