Lecture 7 — VMA and VARMA Representations, Impulse Responses, Orthogonalization, and Inference

Turning the reduced-form VAR into a dynamic propagation model

Jiajing Sun

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

Econometrics and Time Series Methods
Spring 2026

How Lecture 7 continues the logic of Lecture 6

Lecture 6 estimated the reduced-form VAR and asked whether lagged variables improve forecasting. Lecture 7 now asks a different question:

What happens to the whole system after a shock hits one component?

To answer that, we need to move from the autoregressive form to moving-average representations and then interpret the resulting dynamic coefficients as **impulse responses**.

Big picture

A fitted VAR is only the beginning. The real economic interpretation starts when we translate the coefficient matrices into shock-propagation paths.

Textbook sequence for this lecture

To keep the chapter logic coherent, this lecture follows the next steps of the multivariate-time-series chapter:

1. **VMA and VARMA representations**
   - vector moving-average models,
   - vector ARMA models,
   - infinite-order representations.
2. **Impulse responses**
   - response matrices Ψ_k,
   - orthogonalized shocks,
   - ordering issues and interpretation.
3. **Inference and practice**
   - delta-method logic,
   - bootstrap confidence bands,
   - applied interpretation using ETF-return IRFs.

Learning goals for this three-hour lecture

By the end of Lecture 7, students should be able to:

1. define VMA(q) and VARMA(p, q) processes and explain why they matter conceptually;
2. derive or at least recognize the VMA(∞) representation of a stable VAR;
3. interpret the matrices Ψ_k as impulse-response objects;
4. explain why orthogonalization is needed when reduced-form shocks are contemporaneously correlated;
5. read an IRF figure carefully, including sign, persistence, horizon, and confidence-band information.

Practical plan for the three contact hours

Hour 1

VMA and VARMA models, infinite-order representations, and why direct estimation is harder than VAR estimation.

Hour 2

Impulse-response functions, orthogonalized shocks, ordering issues, and economic interpretation.

Hour 3

Inference for IRFs, bootstrap logic, and an empirical illustration using equity ETF returns.

Why we need a moving-average representation

The autoregressive form is excellent for estimation, but it is not the most transparent way to describe shock propagation.

- In a VAR, shocks enter contemporaneously and then keep reappearing indirectly through future lagged values.
- To see the *full dynamic effect* of one innovation, we want the system written directly as a function of current and past shocks.
- That is exactly what a vector moving-average representation provides.

Core idea

The VMA form tells us how much of today can be written as a weighted sum of past innovations. Those weights become the impulse responses.

Definition of a VMA(q)

A vector moving average of order q is

$$y_t = \mu + \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \cdots - \Theta_q \varepsilon_{t-q},$$

where the Θ_j are $(n \times n)$ matrices.

- Current values depend on the current innovation and a finite number of past innovations.
- Each shock can affect several variables at several future dates through the coefficient matrices.
- Invertibility, not stationarity, is the key additional condition for recovering innovations from observables.

The VMA(1) autocovariance structure

For

$$y_t = \mu + \varepsilon_t - \Theta\varepsilon_{t-1},$$

with innovation covariance matrix Ω, the autocovariances are

$$\Gamma(0) = \Omega + \Theta\Omega\Theta',$$

$$\Gamma(1) = -\Theta\Omega, \qquad \Gamma(-1) = -\Omega\Theta', \qquad \Gamma(k) = 0 \text{ for } |k| > 1.$$

- This extends the familiar univariate MA(1) truncation property.
- Cross-covariances across variables are built into the matrices.

What the VMA form makes easy to see

Suppose one component of ε_t receives a unit shock.

- The contemporaneous effect is given by the corresponding column of the identity matrix.
- The one-period-ahead effect is given by the corresponding column of $-\Theta_1$.
- More generally, the VMA coefficients map shocks directly into future responses.

Interpretive advantage

In the autoregressive form, propagation is implicit. In the moving-average form, propagation is explicit.

Example: a common-factor VMA(1)

The textbook gives a useful construction. Suppose

$$x_{it} = \varepsilon_{it} - \theta \varepsilon_{i,t-1}, \qquad i = 1, \ldots, n,$$

with independent univariate MA(1) components, and define

$$\mathrm{y}_t = A\mathrm{x}_t.$$

Then y_t is an n-dimensional VMA(1) with

$$\Gamma(0) = \sigma_\varepsilon^2 (1 + \theta^2) AA', \qquad \Gamma(1) = -\theta \sigma_\varepsilon^2 AA'.$$

Why this example is useful

It shows clearly how common latent factors can generate multivariate moving-average dependence across all variables at once.

Example: a two-dimensional VMA(1)

Consider

$$y_{1t} = \varepsilon_{1t} + \varepsilon_{2,t-1}, \qquad y_{2t} = \varepsilon_{2t} + \theta\varepsilon_{2,t-1}, \qquad |heta| < 1.$$

Here y_{1t} is i.i.d. but y_{2t} is an MA(1). The lag-one cross-autocorrelation can be larger than the own autocorrelation of y_2.

Lesson

In multivariate systems, cross-dependencies can be stronger and more revealing than marginal own-lag dependencies. That is one reason the multivariate framework is richer than running separate univariate models.

Definition of a VARMA(p, q)

A vector ARMA model combines autoregressive and moving-average matrix polynomials:

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

with

$$\Phi(L) = I_n - \Phi_1 L - \cdots - \Phi_p L^p, \qquad B(L) = I_n - \Theta_1 L - \cdots - \Theta_q L^q.$$

Under stability and invertibility,

$$\mathrm{y}_t - \mu = \Psi(L)\varepsilon_t, \qquad \Psi(z) = \Phi(z)^{-1} B(z).$$

- VARMA is the multivariate analogue of univariate ARMA.
- The transfer function $\Psi(z)$ is the object that ultimately generates impulse responses.

Stability and invertibility in the VARMA setting

The two key matrix-polynomial conditions are:

- **Stability / causality:** the roots of

$$\det \Phi(z) = 0$$

  lie outside the unit circle.

- **Invertibility:** the roots of

$$\det B(z) = 0$$

  lie outside the unit circle.

Consequence

Under these conditions, a VARMA process can be written both as an infinite-order VAR and as an infinite-order VMA. That dual representation is central to the theory.

Infinite-order representations

Under suitable conditions, the VARMA model admits both forms:

$$C(L)y_t = c' + \varepsilon_t \qquad \text{(VAR}(\infty)\text{ form)},$$

$$y_t = c'' + D(L)\varepsilon_t \qquad \text{(VMA}(\infty)\text{ form)}.$$

- The VAR(∞) representation emphasizes prediction from lagged observables.
- The VMA(∞) representation emphasizes shock decomposition.
- IRFs live naturally in the VMA(∞) representation.

Why direct estimation of VARMA is harder than estimation of VAR

The textbook emphasizes a practical point: direct VMA and VARMA estimation is substantially harder. Even for a VMA(1),

$$\Gamma(0) = \Omega + \Theta\Omega\Theta', \qquad \Gamma(1) = -\Theta\Omega,$$

which can be combined into the matrix quadratic equation

$$\Gamma(1)P\Gamma(1)'P - \Gamma(0)P + I_n = 0, \qquad P = \Omega^{-1}.$$

- Unlike the scalar MA case, there is no general closed-form solution.
- Invertibility and identification are nonlinear constraints.
- This is why moderately high-order VARs are often used as flexible approximations to VARMA dynamics.

From the VMA(∞) representation to impulse responses

Suppose a stable VAR or VARMA model admits

$$y_t - \mu = \sum_{k=0}^{\infty} \Psi_k \varepsilon_{t-k}.$$

The matrices Ψ_k are the **impulse-response matrices**.

- $\Psi_0 = I_n$ for reduced-form innovations.
- Ψ_1 tells us the one-step-ahead effect of today's shock.
- Ψ_2 tells us the two-step-ahead effect, and so on.

Elementwise interpretation of Ψ_k

The (i, j) element of Ψ_k is

$$(\Psi_k)_{ij} = \frac{\partial y_{i,t+k}}{\partial \varepsilon_{jt}}.$$

So it measures the response of variable i at horizon k to a one-unit innovation in shock component j at date t, holding all other innovations fixed.

How to read it

Fix a shock index j and trace the entire column of response variables $i = 1, \ldots, n$ across horizons $k = 0, 1, 2, \ldots$. That is the dynamic footprint of one shock across the whole system.

Recursion for the impulse-response matrices in a VAR(p)

Write the stable VAR in VMA(∞) form as

$$y_t - \mu = \sum_{k=0}^{\infty} \Psi_k \varepsilon_{t-k}, \qquad \Psi_0 = I_n.$$

Substituting into

$$y_t - \mu = \Phi_1 (y_{t-1} - \mu) + \cdots + \Phi_p (y_{t-p} - \mu) + \varepsilon_t$$

and matching coefficients on ε_{t-k} gives

$$\Psi_k = \Phi_1 \Psi_{k-1} + \Phi_2 \Psi_{k-2} + \cdots + \Phi_p \Psi_{k-p}, \qquad k \geq 1,$$

with $\Psi_k = 0$ for $k < 0$. Equivalently, in generating-function notation,

$$\Phi(L)\Psi(L) = I_n, \qquad \Psi(L) = I_n + \Psi_1 L + \Psi_2 L^2 + \cdots .$$

Companion-form route to the same object

Lecture 6 already gave us a shortcut. If the augmented centered system is

$$\tilde{\xi}_t = F\tilde{\xi}_{t-1} + v_t, \qquad \tilde{y}_t = J\tilde{\xi}_t,$$

then

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

So the impulse-response matrices are simply

$$\Psi_k = JF^k J'.$$

Why this is useful

The companion form turns the entire IRF problem into repeated matrix multiplication, which is ideal for proofs, forecasting formulas, and software implementation.

Why raw reduced-form shocks are hard to interpret

Reduced-form innovations are rarely orthogonal.

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

where the off-diagonal elements of Ω are generally not zero.

- A unit increase in ε_{jt} is not necessarily a shock *only* to variable j in an economically clean sense.
- It may come bundled with contemporaneous movements in other innovation components.

Interpretation problem

The raw IRF to ε_{jt} mixes innovation components whenever the reduced-form shocks are correlated.

Standardized shocks versus orthogonalized shocks

Two common alternatives appear in the textbook and in software practice.

- **Standardized shocks:** scale each innovation by its own standard deviation.
- **Orthogonalized shocks:** transform the innovation vector into a new vector with uncorrelated components and unit or diagonal covariance.

If $\varepsilon_t = Pw_t$ with $PP' = \Omega$ and $\mathrm{E}(w_t w_t') = I_n$, then the orthogonalized responses are based on

$$\Psi_k P.$$

Cholesky orthogonalization

A standard approach is to use a Cholesky factorization

$$\Omega = PP',$$

where P is lower triangular. Then define

$$\varepsilon_t = Pw_t, \qquad \mathrm{E}(w_t w_t') = I_n.$$

The response to a one-unit orthogonal shock in component j is the jth column of

$$\Psi_k P.$$

Important caveat

Because the Cholesky factor is triangular, the orthogonalized responses depend on the variable ordering.

LDL decomposition and the textbook formula

The chapter also presents an LDL-style orthogonalization. If

$$\Omega = ADA',$$

where A is lower triangular with ones on the diagonal and D is diagonal, define

$$u_t = A^{-1}\varepsilon_t.$$

Then the components of u_t are uncorrelated, with covariance matrix D. The orthogonalized impulse response to the jth variable is then written as

$$\Psi_k a_j,$$

where a_j is the jth column of A.

Bottom line

Whether you think in Cholesky or LDL terms, the goal is the same: isolate shocks that do not move contemporaneously together.

Ordering dependence is not a technical footnote

In recursive orthogonalization, the ordering encodes contemporaneous priority.

- Variables earlier in the ordering can affect later variables contemporaneously.
- Variables later in the ordering cannot affect earlier ones contemporaneously within the recursive scheme.

Interpretive rule

If the ordering changes, the orthogonalized IRFs can change too. Therefore the ordering must be defended economically, not chosen casually.

What an IRF plot is telling you

For any response path, ask five questions.

1. What is the **impact effect** at horizon 0 or 1?
2. Is the response **positive or negative**?
3. How quickly does it **decay**?
4. Is there **overshooting, oscillation, or sign reversal**?
5. Do the confidence bands suggest the response is precisely estimated?

Interpretation mindset

An IRF is not just a line. It is a compact summary of timing, sign, persistence, and uncertainty.

Common dynamic shapes in impulse responses

- **Pure decay:** the response is strongest on impact and then shrinks smoothly.
- **Hump-shaped response:** the effect builds for a few periods before fading.
- **Oscillatory response:** the sign alternates across horizons.
- **Persistent near-unit-root response:** the effect remains large for many periods.

Economic reading

These shapes often correspond to adjustment frictions, feedback loops, overshooting mechanisms, or highly persistent state variables.

Cumulative and long-run responses

Sometimes the object of interest is the cumulative response up to horizon K,

$$\sum_{k=0}^{K} \Psi_k,$$

or the long-run multiplier

$$\sum_{k=0}^{\infty} \Psi_k = \Psi(1) = \Phi(1)^{-1} B(1).$$

For a reduced-form VAR, $B(1) = I_n$, so

$$\sum_{k=0}^{\infty} \Psi_k = \left(I_n - \Phi_1 - \cdots - \Phi_p\right)^{-1}.$$

- Cumulative responses summarize medium-run adjustment.
- Long-run multipliers become especially important in later lectures on structural and nonstationary systems.

Reduced-form OIRFs versus structural interpretation

Orthogonalized IRFs are already more interpretable than raw reduced-form IRFs, but they are still not automatically structural.

- Recursive orthogonalization imposes one particular contemporaneous structure.
- A structural VAR adds explicit identifying restrictions rooted in economic theory.
- We will revisit that in the SVAR lecture rather than mixing it into today's reduced-form discussion.

Careful language again

It is safer to say "orthogonalized shock under a recursive ordering" than "structural shock," unless an identification strategy has been fully justified.

Why IRFs need sampling uncertainty

An estimated IRF is a nonlinear function of estimated VAR coefficients and the estimated innovation covariance matrix.

- Even if the fitted line looks smooth, it is still subject to sampling error.
- Uncertainty grows with horizon in many applications.
- Responses that seem economically meaningful can be statistically imprecise.

Therefore

Any serious IRF analysis should report confidence intervals or confidence bands, not just point estimates.

Delta-method logic for IRFs

Let θ collect the estimated VAR parameters and let

$$\iota = f(\theta) = \begin{bmatrix} \text{vec}(\Psi_1) \\ \vdots \\ \text{vec}(\Psi_K) \end{bmatrix}$$

stack the IRF coefficients up to horizon K. If

$$\sqrt{T}(\widehat{\theta} - \theta) \Longrightarrow N(0, V_\theta),$$

then the delta method gives

$$\sqrt{T}(\widehat{\iota} - \iota) \Longrightarrow N\big(0,\ G(\theta) V_\theta G(\theta)'\big), \qquad G(\theta) = \frac{\partial f(\theta)}{\partial \theta'}.$$

- Conceptually clean.
- Algebraically cumbersome in large systems.
- This is one reason bootstrap bands are so popular in applied work.

Pointwise confidence intervals

A common practice is to report, for each horizon k and each impulse–response pair (i, j),

$$\widehat{\Psi}_{ij,k} \pm z_{\alpha/2}\, \widehat{\mathrm{se}}_{ij,k}.$$

- These intervals are easy to compute and easy to read.
- But they are **pointwise**: they do not provide simultaneous coverage over all horizons and all variables.

Interpretation

A pointwise interval tells you about uncertainty at one specific horizon, not about the whole path jointly.

Simultaneous bands are wider but more demanding

If we want a band that covers the full response path over many horizons simultaneously, the critical value must be larger.

- Simultaneous bands are harder to compute accurately.
- They are usually wider than pointwise intervals.
- But they are closer to the real question many readers ask: "Do I trust the whole path, not just one point on it?"

Practical compromise

Many applied papers show pointwise bootstrap bands and then interpret the path cautiously, horizon by horizon.

Residual bootstrap algorithm for VAR IRFs

A basic residual-bootstrap procedure works as follows:

1. estimate the VAR and compute residuals $\widehat{\varepsilon}_t$;
2. resample the centered residuals with replacement;
3. generate a bootstrap sample recursively under the estimated VAR;
4. re-estimate the VAR and recompute the IRFs;
5. repeat many times and use the empirical distribution of $\widehat{\Psi}_k^*$ to form bands.

In symbols, bootstrap data are generated from

$$y_t^* = \widehat{c} + \sum_{\ell=1}^{p} \widehat{\Phi}_\ell y_{t-\ell}^* + \varepsilon_t^*.$$

Why bootstrap is popular

It handles the nonlinear mapping from VAR coefficients to IRFs without forcing us to derive every asymptotic covariance term explicitly.

Parametric versus residual bootstrap

Two common variants are:

- **Residual bootstrap:** resample estimated residuals from the fitted model.
- **Parametric bootstrap:** simulate Gaussian shocks from

$$N(0, \widehat{\Omega}).$$

- Residual bootstrap preserves the empirical distribution of the estimated shocks.
- Parametric bootstrap is cleaner if Gaussianity is believed and sample size is not too small.

Local projections as an alternative

The chapter briefly notes that impulse responses can also be estimated using **local projections**. For each horizon h, estimate a separate regression for y_{t+h} on current shocks or current variables and lags.

- Local projections are often more robust to model misspecification.
- Under correct VAR specification, they are typically less efficient than VAR-based IRFs.

How to think about them

VAR-based IRFs are the natural choice inside the linear-system framework. Local projections are a useful robustness tool.

Practical inference checklist for IRFs

Before interpreting an impulse response seriously, ask:

1. Is the underlying VAR stable and well specified?
2. Are the variables and ordering economically defensible?
3. Are the bands pointwise or simultaneous?
4. How many bootstrap replications were used?
5. Do the results survive reasonable robustness checks?

Empirical discipline

The economic story should come after the specification and inference story, not before it.

Textbook case study: equity ETF returns

The chapter includes a useful applied example with four liquid ETFs:

- **SPY:** large U.S. equities,
- **QQQ:** growth-oriented Nasdaq equities,
- **IWM:** U.S. small-cap equities,
- **EFA:** developed ex-U.S. equities.

The workflow is:

1. download daily prices,
2. convert to logarithmic returns,
3. choose the lag order by AIC,
4. estimate a reduced-form VAR,
5. compute orthogonalized IRFs with bootstrap confidence intervals.

Why ETF returns are a good teaching example

This example is convenient because it brings out several core themes at once.

- Returns are closer to stationarity than prices, so the reduced-form stationary VAR framework is more defensible.
- The assets are related but not identical, so spillovers are plausible.
- Contemporaneous shock correlation is almost unavoidable, so orthogonalization matters.
- The resulting IRFs are short-horizon and economically interpretable.

R workflow in compact form

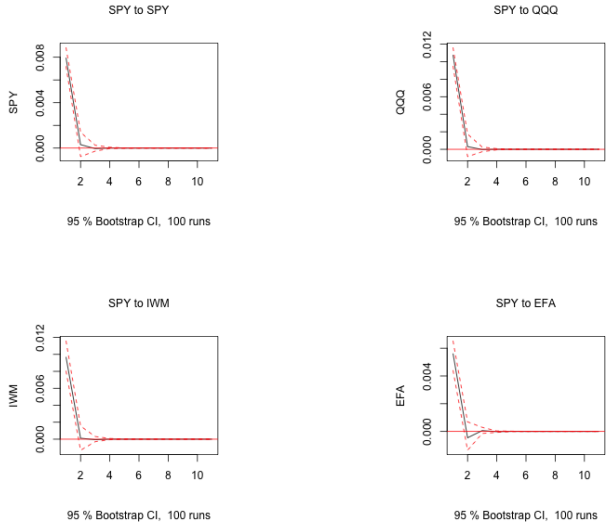
The chapter's script uses `quantmod` and `vars`. The essential steps are:

```
library(quantmod)
library(vars)

getSymbols(c("SPY","QQQ","IWM","EFA"), src = "yahoo",
           from = "2024-01-01", to = "2024-12-31")

prices <- na.omit(merge(Ad(SPY), Ad(QQQ), Ad(IWM), Ad(EFA)))
returns <- as.data.frame(na.omit(diff(log(prices))))

lag_selection <- VARselect(returns, lag.max = 10, type = "const")
var_model <- VAR(returns, p = lag_selection$selection["AIC(n)"], type = "const")

irf(var_model, impulse = "SPY", response = "QQQ", ortho = TRUE, boot = TRUE)
```

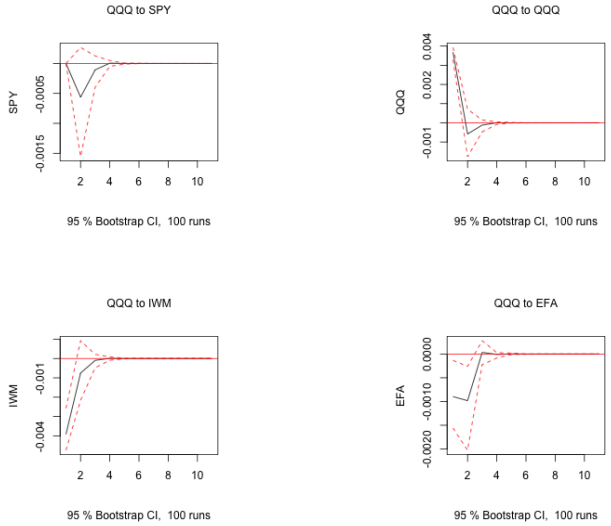
What the output gives us

For each ordered impulse–response pair, we obtain an orthogonalized impulse response with bootstrap confidence bands across horizons.

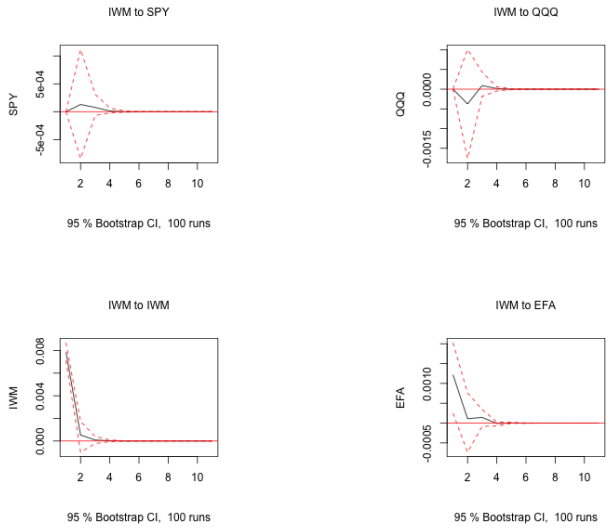
Shock to SPY: responses of all four ETFs



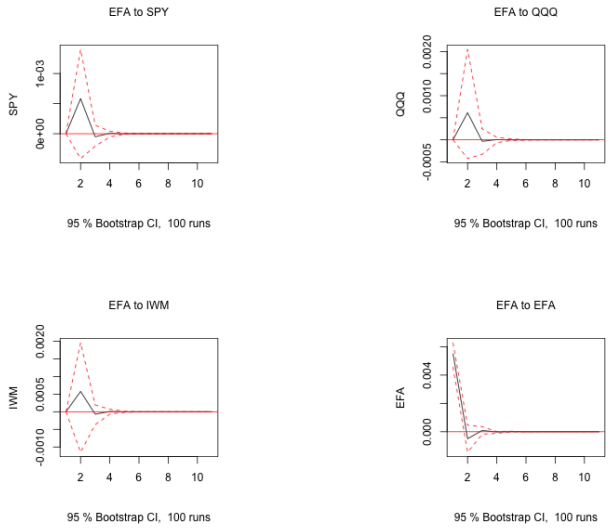
Shock to QQQ: responses of all four ETFs



Shock to IWM: responses of all four ETFs



Shock to EFA: responses of all four ETFs



How to read the ETF IRF panels

Even without imposing a structural asset-pricing model, the panels typically show a few standard reduced-form features.

- Own responses are usually strongest on impact and decay quickly.
- Cross-market responses are often positive but smaller than own responses.
- Confidence bands tend to widen with horizon, so far-out responses become hard to distinguish from zero.
- The relative size of cross responses provides a descriptive map of short-run market connectedness.

Interpret cautiously

These figures describe dynamic reaction patterns under a particular recursive orthogonalization. They are informative, but they are not yet a structural decomposition of world equity shocks.

What the ETF illustration teaches us

This example neatly summarizes why VAR methods remain so popular.

- The VAR is easy to estimate.
- The IRFs turn the coefficient matrices into economically interpretable trajectories.
- The orthogonalization step forces us to confront contemporaneous correlation and ordering assumptions.
- Bootstrap bands remind us that these trajectories are estimated objects, not deterministic truths.

Connectedness as a natural extension

Once impulse responses are available, it is natural to summarize how much one variable contributes to the movement of another over a finite horizon.

- Impulse responses show the path of a response to a specific shock.
- Forecast-error variance decompositions summarize how much of a variable's forecast uncertainty comes from different shocks.
- Connectedness or spillover indices aggregate these effects across the system.

Why this matters

The path from VAR coefficients to connectedness measures goes through the same VMA/IRF machinery we studied today.

A spillover-type formula from the VMA representation

The textbook mentions measures based on the VMA coefficients such as

$$d_{ij}^{K} = \frac{\sigma_{jj}^{-1} \sum_{k=0}^{K-1} \left(e_i' \Psi_k \Omega e_j\right)^2}{\sum_{k=0}^{K-1} e_i' \Psi_k \Omega \Psi_k' e_i}.$$

- The numerator tracks the contribution of shocks from variable j to variable i over the horizon range.
- The denominator normalizes by the total forecast-error variance of variable i.

Interpretation

This takes us from "how does variable i respond to shock j?" to "how important are shocks from j for explaining the uncertainty of i?"

Why high-order VAR approximation is common in practice

Direct VMA or VARMA estimation often brings more pain than gain.

- A flexible VAR with carefully selected lag length is easy to estimate.
- It often provides a good approximation to the short-run multivariate dynamics of more complicated systems.
- Once the VAR is estimated, IRFs, forecasts, causality tests, and variance decompositions come almost for free.

Practical takeaway

In many empirical settings, "estimate a sensible VAR first" is not a compromise. It is the standard professional workflow.

A recommended workflow after Lectures 6 and 7

1. transform the variables so the stationary reduced-form framework is plausible;
2. estimate a stable VAR with a defensible lag length;
3. inspect residuals and innovation covariance;
4. compute Granger-causality tests if predictive questions matter;
5. compute orthogonalized IRFs with bootstrap bands if dynamic interpretation matters;
6. only then consider stronger structural claims or nonstationary extensions.

Preview of the next lecture block

So far we have assumed the stationary VAR framework is appropriate. The next lecture asks what changes when the variables are nonstationary but move together in the long run. That takes us to:

- nonstationary VAR systems,
- cointegration,
- vector error-correction models (VECMs),
- and the interpretation of permanent and transitory components.

Why the order matters

Cointegration is much easier to understand once the stationary VAR, the VMA representation, and the impulse-response machinery are already familiar.

Lecture 7 summary

- VMA and VARMA models express multivariate dynamics directly in terms of current and lagged innovations.
- For a stable VAR, the VMA(∞) representation provides the impulse-response matrices Ψ_k.
- Orthogonalization is needed because reduced-form shocks are generally contemporaneously correlated.
- Cholesky or LDL decompositions give practical orthogonalized IRFs, but the results depend on variable ordering.
- IRFs are nonlinear objects, so confidence intervals or bootstrap bands are essential.

Take-away

Lecture 6 estimated the VAR. Lecture 7 turned it into a map of dynamic shock transmission.