Chapter 1 — Regression Models

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Table of Contents

Regression Models in Econometrics and Finance

- Regression models are central tools in applied econometrics.
- In finance, linear regression underpins:
  - Asset pricing (CAPM, multifactor APT).
  - Portfolio construction and risk control via expected returns, volatilities, correlations.
  - Tests of market efficiency and links from macro conditions to financial outcomes.
- In economics more broadly, regression analysis quantifies how key variables (output, inflation, unemployment, trade, policy instruments, etc.) move together.
- Well-specified regressions support policy evaluation, testing theoretical restrictions, and forecasting.
- Credible analysis requires transparent assumptions, careful design, and disciplined diagnostics.

From Question to Regression Model

- Building a regression model follows a common workflow:
  1. Start from a clear research question and theoretical rationale.
  2. Curate and preprocess data (missing values, outliers, transformations, sources).
  3. Specify the model: dependent and explanatory variables, testable hypotheses.
  4. Estimate parameters (often OLS in linear models).
  5. Diagnose model fit and assumptions (residuals, heteroskedasticity, etc.).
  6. Interpret results, run counterfactuals, and use the model for forecasting if appropriate.
- This chapter implements each step with reproducible `R` examples.

Simple (Univariate) Linear Regression: Model

We study the simple (univariate) linear regression model:

$$y = \beta_0 + \beta_1 x + u,$$

where:

- y: dependent (explained) variable.
- x: independent (explanatory, predictor) variable.
- β_0: intercept, $\mathrm{E}(y \mid x = 0)$.
- β_1: slope, change in $\mathrm{E}(y)$ for a one-unit change in x.
- u: error term, capturing all other determinants of y not in x.

- Goal: choose β_0, β_1 so that predicted values of y are as close as possible to observed values.

Terminology for x and y in Regression

Common terminology:

x	y
Independent variable	Dependent variable
Explanatory variable	Explained variable
Predictor variable	Predicted variable
Regressor	Regressand
Control variable	Response variable

- These names emphasize different roles (causal, predictive, control, etc.).
- The same terminology extends naturally to multiple regression with many regressors.

Applications of Simple Regression

- In financial markets, simple regressions are used to:
  - Estimate the CAPM beta of an asset (return on stock vs. market return).
  - Study risk–return trade-offs and risk-adjusted performance.
- In risk management:
  - Quantify volatility and correlations of assets.
  - Support portfolio construction and Value-at-Risk style measures.
- In macro–finance:
  - Relate stock returns or bond yields to macro variables (inflation, growth, policy rates).
- In later chapters, these ideas extend to time series (autoregressive models, cointegration).

Historical Note: Galton and "Regression"

Galton's Law (1889)

Francis Galton studied the relationship between parents' and children's heights:

- Tall parents tend to have tall children, but on average not as tall as the parents.
- Short parents tend to have short children, but on average not as short as the parents.

He described this as regression towards the mean: offspring heights tend to move towards the population average, regardless of the parents' height.

- The term "regression" comes from this phenomenon.
- Today, regression analysis is a general method for studying linear relationships.

Data for Teaching Regression in R

- `ggplot2::economics`: macroeconomic time series such as unemployment and income.
- `quantmod`: easy access to stock price series for financial regression examples.
- FRED data via `quantmod` or `fredr`:
  - Rich macro and financial series (interest rates, exchange rates, GDP, etc.).
- These built-in or easily accessible datasets make it straightforward to demonstrate:
  - Economic relationships and stylized facts.
  - Practical implementation of regression in R.

Assumptions of the Simple Linear Regression Model (1)

Let $x = \{x_i\}$, $y = \{y_i\}$, $i = 1, \ldots, n$. The simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Assumption 1: Linearity

- The conditional expectation of y_i is linear in x_i.

Assumption 2: Zero Conditional Mean

$$\mathrm{E}(u_i \mid x_i) = 0 \quad \Rightarrow \quad \mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i.$$

- No omitted variables systematically related to x_i.
- Model correctly specifies the conditional mean of y_i.

Assumptions of the Simple Linear Regression Model (2)

Assumption 3: Homoscedasticity (Equal Variance)

$$\mathrm{Var}(u_i \mid x_i) = \sigma^2 \quad \text{for all } i.$$

- Error variance does not depend on the level of x_i.

Assumption 4: Uncorrelated Errors

$$\mathrm{Cov}(u_i, u_j) = 0 \quad \text{for } i \neq j.$$

- Error terms are uncorrelated across observations.
- A stronger version requires $\{u_i\}$ to be mutually independent.

Assumptions of the Simple Linear Regression Model (3)

Assumption 5: Normality (Optional)

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad \Leftrightarrow \quad u_i \sim N(0, \sigma^2).$$

- Not needed for unbiasedness or consistency of OLS.
- Useful for exact small-sample inference (t-tests, F-tests).

Independence vs Zero Covariance

Statistical independence implies zero covariance, but not vice versa. Example: if x is symmetric with $\mathrm{E}(x) = 0$ and $\mathrm{E}(x^3) = 0$, and $y = x^2$, then x and y are clearly dependent, but $\mathrm{Cov}(x, y) = 0$.

Ordinary Least Squares (OLS): Objective

We choose β_0, β_1 to minimize the sum of squared residuals:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- OLS picks the line that best fits the data in the least squares sense.
- "Closeness" is measured by squared prediction errors.
- This gives a regression line that passes through the center of the data cloud.

OLS: Closed-Form Solution

Let

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The OLS estimators are:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus the fitted regression line is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = \bar{y} + \widehat{\beta}_1 (x - \bar{x}).$$

Residuals and Their Key Properties

Residuals are the differences between observed and fitted values:

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i.$$

OLS residuals satisfy several important properties:

- The regression line passes through the sample mean point $(\bar{x}, \bar{y})$.
- The residuals sum to zero:
$$\sum_{i=1}^{n} \widehat{u}_i = 0.$$
- The residuals are "orthogonal" to the regressor:
$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0.$$

More Properties of OLS Residuals

- Since both $\sum \widehat{u}_i = 0$ and $\sum x_i \widehat{u}_i = 0$, we have

$$\text{Cov}(x_i, \widehat{u}_i) = 0.$$

- Similarly, because $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$,

$$\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0 \quad \Rightarrow \quad \text{Cov}(\widehat{y}_i, \widehat{u}_i) = 0.$$

- These orthogonality properties are mechanical consequences of the OLS minimization, not evidence that the model is truly correct (they do not rule out omitted variable bias).

Errors vs Residuals

Error vs Residual

- **Error** u_i:
  - Difference between y_i and the *true* regression line.
  - Depends on unknown parameters β_0, β_1, so it is unobservable.
- **Residual** $\widehat{u}_i$:
  - Difference between y_i and the *fitted* regression line.
  - Observable once OLS estimates are computed.
  - Used for diagnostics (heteroskedasticity, autocorrelation, outliers).

Residual analysis is central to checking whether regression assumptions are reasonable.

The t-Statistic in Simple Regression

To test $H_0 : \beta_i = 0$ (e.g., slope has no effect), the t-statistic is

$$t_i = \frac{\widehat{\beta}_i - 0}{\mathrm{SE}(\widehat{\beta}_i)}, \qquad i = 0, 1,$$

where $\mathrm{SE}(\widehat{\beta}_i)$ is the estimated standard error.

- Under the classical assumptions (including normality), t_i follows a t-distribution with $n - 2$ degrees of freedom.
- Large absolute t_i (and small p-value) indicates a statistically significant effect.

t-Distributions vs the Normal Distribution

- The t-distribution has heavier tails than the standard normal, especially at low degrees of freedom.
- As degrees of freedom increase, the t-distribution approaches the standard normal.
- This reflects the Central Limit Theorem: with larger samples, sampling distributions become approximately normal.

t-Distributions vs the Normal Distribution

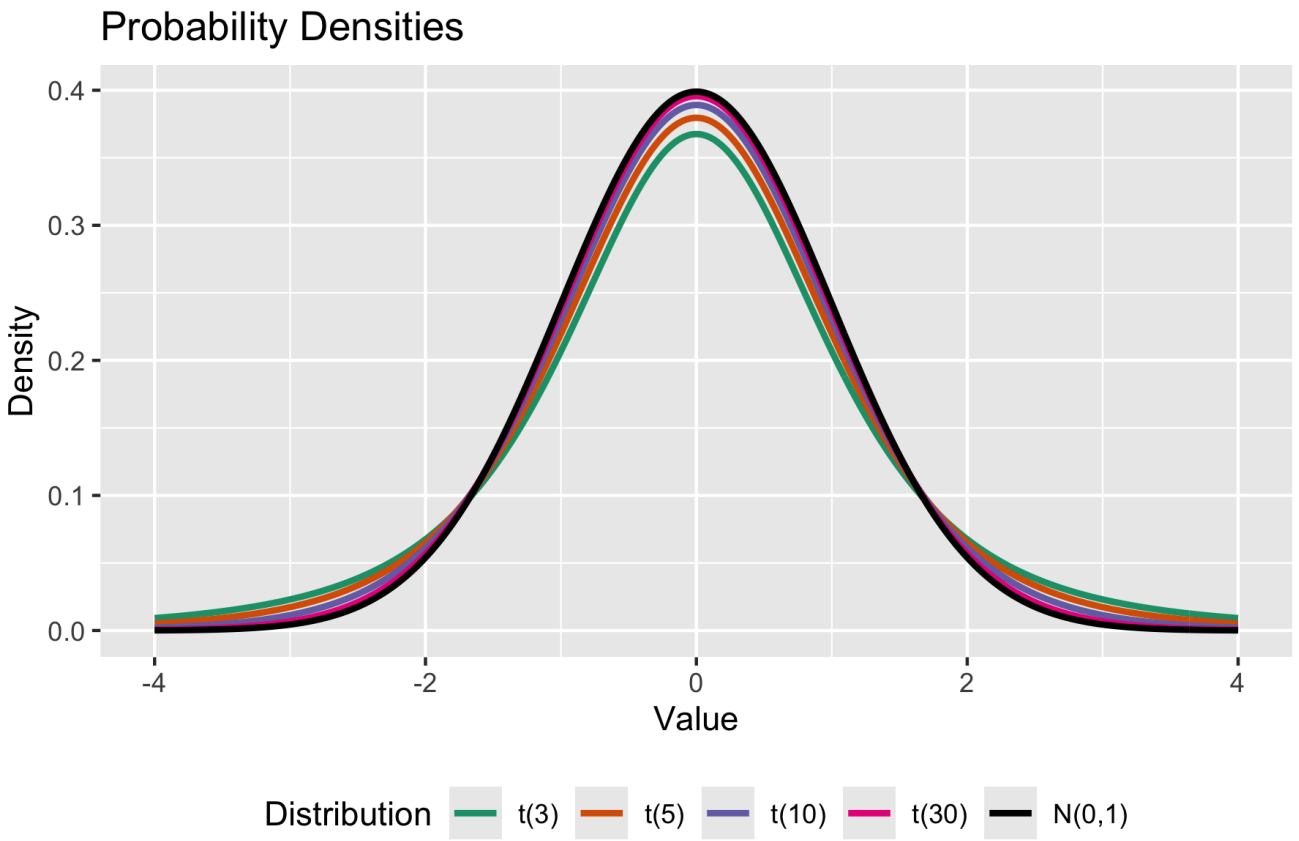


Figure: t-distribution densities and the standard normal density.

Total, Explained, and Residual Variation

Define:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

$$\text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2.$$

- SST: total variability of y.
- SSE: variability explained by the regression line.
- SSR: variability left in the residuals.

They satisfy the decomposition:

$$\text{SST} = \text{SSE} + \text{SSR}.$$

Coefficient of Determination R^2

The coefficient of determination is defined as

$$R^2 \equiv \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

- Measures the fraction of total variation in y explained by the regression.
- $0 \leq R^2 \leq 1$:
  - $R^2 = 1$: perfect linear fit, all points lie on the regression line.
  - $R^2 \approx 0$: regressors explain little of the variation in y.
- In simple regression, R^2 equals the square of the sample correlation between y_i and $\widehat{y}_i$.

Case Study: Car Weight and Fuel Efficiency

- Dataset: `mtcars` (built into `R`).
- Question: Does car weight (`wt`) negatively affect fuel efficiency (miles per gallon, `mpg`)?
- Model:

$$\text{mpg}_i = \beta_0 + \beta_1 \text{wt}_i + u_i.$$

- Estimation: OLS using `lm(mpg ~ wt, data = mtcars)` in `R`.
- We also visualize the relationship and estimate a simple regression line.

Case Study: Regression Output

OLS results (abridged):

- **Intercept** $\widehat{\beta}_0 \approx 37.29$ (highly significant).
- **Slope** $\widehat{\beta}_1 \approx -5.34$, t-value ≈ -9.56, $p \approx 1.3 \times 10^{-10}$.
  - Heavier cars have significantly lower fuel efficiency.
- Residual standard error ≈ 3.05, df $= 30$.
- Multiple $R^2 \approx 0.75$, adjusted $R^2 \approx 0.74$:
  - About 75% of the variation in `mpg` is explained by `wt`.

Case Study: Visualizing the Relationship

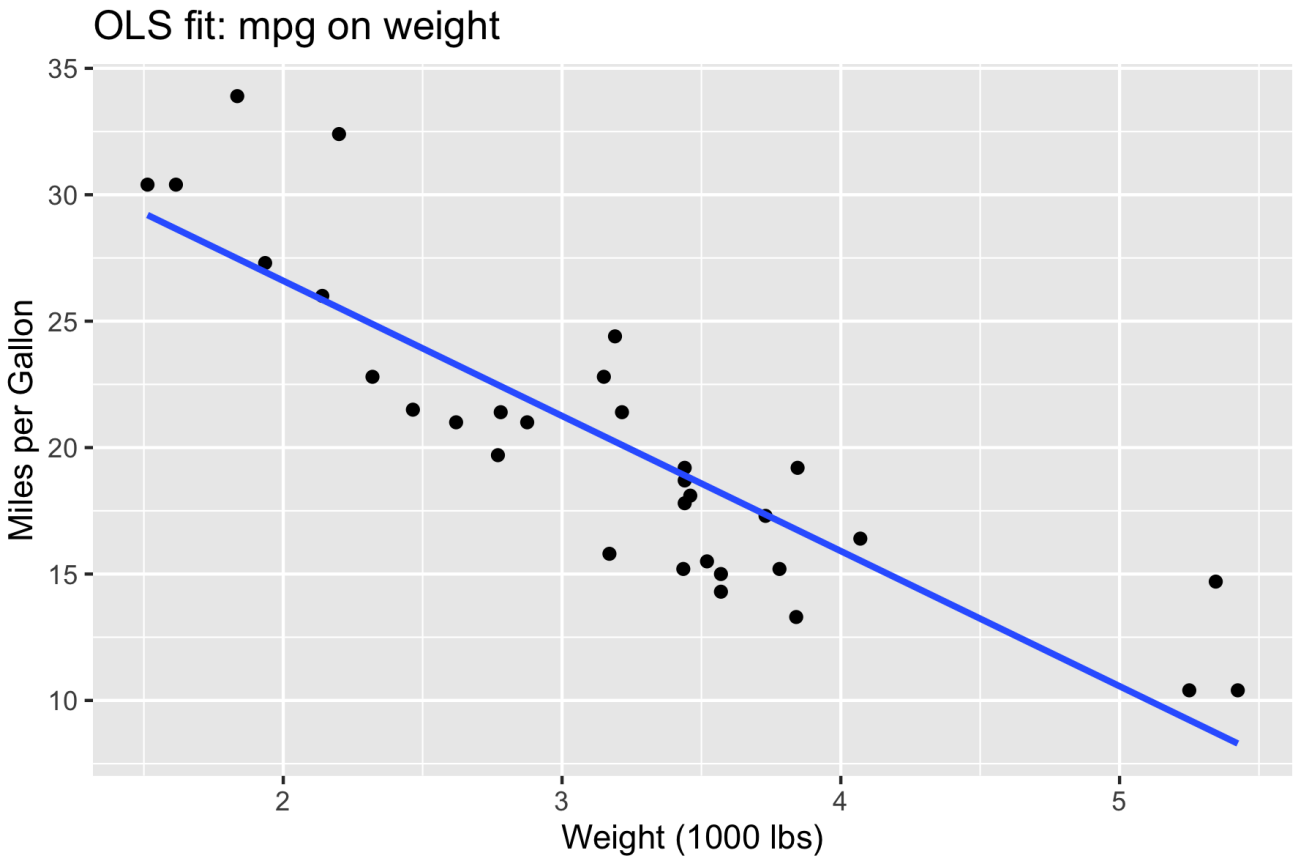


Figure: OLS fit of `mpg` on `wt` in the `mtcars` dataset.

Case Study: Visualizing the Relationship

- Clear negative relationship between weight and fuel efficiency.
- Visualization helps assess linearity and detect potential outliers.

Case Study: ANOVA for mpg ~ wt

```
# ANOVA for the regression model
anova_result <- anova(model)
print(anova_result)
```

Typical output (abridged):

```
Analysis of Variance Table
Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
wt          1 847.73  847.73  91.375 1.294e-10 ***
Residuals  30 278.32    9.28
```

Case Study: Interpreting the ANOVA

- **Source of Variation**:
  - `wt`: variability explained by car weight.
  - `Residuals`: unexplained variability.
- **Degrees of Freedom (Df)**:
  - 1 df for `wt` (one predictor).
  - 30 df for residuals ($n - p - 1$).
- **F-statistic**:
  - $F \approx 91.4$ with $p \approx 1.3 \times 10^{-10}$.
  - Strong evidence that weight is an important predictor of `mpg`.
- **Conclusion**:
  - The simple regression model captures a highly significant negative effect of car weight on fuel efficiency.

Table of Contents

Why Multiple Regression?

- Multiple regression allows us to include several explanatory variables at once and to disentangle different channels.
- The response variable y can be:
  - an asset return, firm output, household consumption,
  - a worker's wage, a country's inflation rate, or its economic growth.
- Covariates $\{x_1, \ldots, x_k\}$ may include:
  - prices, incomes, interest rates, demographic factors,
  - policy indicators, institutional variables, and more.
- Across labour, IO, public, development, macro, and finance, multiple regression is the standard empirical tool for:
  - policy evaluation, structural analysis, and forecasting.

Examples and Interpretation

- **Asset pricing:** portfolio returns depending on several risk factors.
- **Labour economics:** wage equations controlling for education, experience, region, etc.
- **Macroeconomic policy:** effects of interest-rate changes or fiscal shocks holding other determinants of output and inflation fixed.
- Multiple regression provides a unified framework for all such settings.
- **Interpretation:** each slope coefficient measures a *ceteris paribus* (all-else-equal) effect:
  - impact of a one-unit change in a given regressor on the conditional mean of y,
  - holding the other regressors constant.

The Multiple Linear Regression Model

- Model with k explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

- y: dependent (response) variable.
- $x_1, \ldots, x_k$: independent (predictor) variables.
- β_0: intercept, $\mathrm{E}(y)$ when all $x_j = 0$.
- β_j $(j = 1, \ldots, k)$: slope parameters:
  - effect on y of a one-unit change in x_j,
  - holding other regressors fixed.
- u: error term, capturing all other influences on y not in the regressors.

OLS in Multiple Regression

Given a sample $\{(y_i, x_{i1}, \ldots, x_{ik})\}_{i=1}^n$:

- Fitted values:
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}.$$
- Residuals:
$$\widehat{u}_i = y_i - \widehat{y}_i.$$
- OLS chooses $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$ to minimize the sum of squared residuals:
$$\text{SSR} = \sum_{i=1}^n \widehat{u}_i^2 = \sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik})^2.$$
- Setting partial derivatives with respect to each $\widehat{\beta}_j$ to zero gives the *normal equations*.

Normal Equations

The $(k+1)$ normal equations are:

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0,$$

$$\sum_{i=1}^{n} x_{i1}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0,$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0,$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0.$$

- Under a rank condition on the regressors (no perfect multicollinearity), this linear system has a unique solution for $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$.

Matrix Formulation: Model Setup

- Stack the n observations in vectors and matrices:

$$Y = \mathbf{X}\beta + u,$$

where

  - Y: $n \times 1$ vector of y-observations,
  - $\mathbf{X}$: $n \times (k+1)$ regressor matrix,
  - β: $(k+1) \times 1$ parameter vector,
  - u: $n \times 1$ error vector.

- The regressor matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix},$$

where the first column is all ones (intercept).

Matrix Formulation: OLS Solution

- OLS solves:

$$\min_{\beta} u'u = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta).$$

- Differentiating and setting the derivative to zero:

$$-2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\widehat{\beta} = 0 \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'Y.$$

- If $\mathbf{X}$ has full column rank, $\mathbf{X}'\mathbf{X}$ is invertible and

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

- This compact expression is extremely useful both conceptually and computationally.

Matrix Formulation: Interpretation Note

Interpretation of $\mathbf{X}$ and β

If

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & x_1 & x_2 & x_3 \end{bmatrix},$$

then the scalar model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

and Y is the $n \times 1$ vector of observations $\{y_i\}_{i=1}^n$.

- Matrix notation provides a compact way to derive properties of $\widehat{\beta}$, such as unbiasedness and variance.

Gauss–Markov Conditions

We treat $\mathbf{X}$ as fixed (conditional on $\mathbf{X}$).

- Model:

$$Y = \mathbf{X}\beta + u.$$

- Assumptions:

1. **Linearity and full rank:** model linear in parameters, $\operatorname{rank}(\mathbf{X}) = k + 1$.
2. **Zero conditional mean (exogeneity):**

$$\mathrm{E}(u \mid \mathbf{X}) = \mathbf{0}.$$

3. **Homoskedasticity and no autocorrelation:**

$$\mathrm{E}(uu' \mid \mathbf{X}) = \sigma^2 I_n.$$

4. **Normality (optional):**

$$u \mid \mathbf{X} \sim \mathcal{N}(0, \sigma^2 I_n).$$

Gauss–Markov Theorem and BLUE

Gauss–Markov Theorem

Under the Gauss–Markov assumptions, the OLS estimator

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

is the *Best Linear Unbiased Estimator* (BLUE) of β: among all estimators that are linear in Y and unbiased, $\widehat{\beta}$ has the smallest covariance matrix.

- **Linearity:** $\widehat{\beta} = AY$ with $A = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- **Unbiasedness:**

$$\mathrm{E}[\widehat{\beta} \mid \mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[u \mid \mathbf{X}] = \beta.$$

- **Variance:**

$$\mathrm{Var}(\widehat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Efficiency of OLS

- Any other linear unbiased estimator can be written as

$$\beta_0 = CY, \quad C = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + D,$$

  with $D\mathbf{X} = 0$.

- Its variance is

$$\mathrm{Var}(\beta_0 \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2 DD' \succeq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- Because DD' is positive semidefinite, no other linear unbiased estimator has a smaller covariance matrix.
- This is what it means for OLS to be **efficient** within the class of linear unbiased estimators.

Testing Individual Coefficients: t-tests

- For a coefficient β_j, a typical two-sided test is:

$$H_0: \ \beta_j = \beta_j^0 \quad \text{vs.} \quad H_1: \ \beta_j \neq \beta_j^0.$$

- The test statistic is

$$t_j = \frac{\widehat{\beta}_j - \beta_j^0}{\widehat{\text{se}}(\widehat{\beta}_j)} \approx t_{n-k-1},$$

  under the classical assumptions with normality.

- Compare $|t_j|$ to the critical value from a t-distribution or use the p-value.

- One-sided alternatives (e.g., $H_1: \beta_j > c$) use the same statistic but compare to a one-sided critical value.

Joint Linear Restrictions and the F-test

- Many hypotheses involve joint linear restrictions:

$$H_0 : R\beta = r \quad \text{vs.} \quad H_1 : R\beta \neq r,$$

  where R is a $J \times (k+1)$ matrix of rank J and r is $J \times 1$.

- Examples:
  - Overall significance: all slopes zero.
  - Equality of two coefficients: $\beta_2 = \beta_3$.
  - A linear combination: $\beta_2 + 2\beta_3 = 1$.
- Let:
  - SSR_U: residual sum of squares from *unrestricted* model,
  - SSR_R: residual sum of squares from *restricted* model,
  - J: number of restrictions.

General F-statistic

- The general F-statistic for testing $H_0 : R\beta = r$ is:

$$F = \frac{(\mathrm{SSR}_R - \mathrm{SSR}_U)/J}{\mathrm{SSR}_U/(n-k-1)}.$$

- Under the Gauss–Markov assumptions with normality, and if H_0 is true,

$$F \sim F_{J,\, n-k-1}.$$

- At significance level α, reject H_0 if

$$F > F_{1-\alpha}(J,\, n-k-1)$$

or if the p-value is below α.

Overall F-test and Relationship with R^2

- Special case: test if all slopes are zero:

$$H_0: \ \beta_1 = \cdots = \beta_k = 0.$$

- Define:
  - $\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$,
  - $\text{SSR} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$,
  - $\text{SSE} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$,

  with $\text{SST} = \text{SSE} + \text{SSR}$.
- In this case:

$$\text{SSR}_R = \text{SST}, \quad \text{SSR}_U = \text{SSR}, \quad J = k.$$

- The overall F-statistic becomes:

$$F = \frac{(\text{SST} - \text{SSR})/k}{\text{SSR}/(n-k-1)} = \frac{\text{SSE}/k}{\text{SSR}/(n-k-1)}.$$

Expressing the F-statistic via R^2

- Recall:
$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$
- Substituting $\text{SSE} = R^2\text{SST}$ and $\text{SSR} = (1 - R^2)\text{SST}$ into the F-statistic:
$$\begin{aligned} F &= \frac{n-k-1}{k} \frac{\text{SSE}}{\text{SSR}} \\ &= \frac{n-k-1}{k} \frac{R^2}{1-R^2} \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)}. \end{aligned}$$
- For fixed n and k, a larger R^2 corresponds to a larger overall F-statistic.

Expressing the F-statistic via R^2

Interpretation

The overall F-test asks whether the regressors collectively help explain y beyond the intercept; R^2 measures the fraction of variance in y explained by the fitted values. Their algebraic link clarifies why high R^2 typically coincides with a large F-statistic.

Adjusted R^2

- In simple regression, R^2 is natural; in multiple regression it never decreases when adding regressors, even irrelevant ones.
- To penalize overfitting, use adjusted R^2:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} \widehat{u}_i^2 / (n-k-1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}.$$

- Here $\widehat{u}_i$ are OLS residuals, n is sample size, and $k+1$ is number of coefficients (including intercept).
- $\bar{R}^2$ can *decrease* when an unhelpful regressor is added:
  - more informative than R^2 for comparing models with different numbers of regressors.

Residual Standard Error (RSE)

- The residual standard error is an estimate of the standard deviation of the error term:

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-p}} = \sqrt{\frac{\text{SSR}}{n-p}},$$

  where $p = k + 1$ is the number of parameters (including intercept).

- RSE measures the typical size of residuals in the units of y:
  - smaller RSE indicates a tighter fit.
  - as an absolute measure, often interpreted relative to the mean of y (e.g. via coefficient of variation).

Looking Ahead

- So far, we have:
  - defined the multiple regression model and OLS estimator,
  - derived its matrix form and the Gauss–Markov properties (BLUE),
  - discussed t-tests, F-tests, R^2, adjusted R^2, and RSE.
- Real-world data often violate homoskedasticity or feature strongly correlated regressors.
- Next, we turn to two key complications:
  1. **Heteroskedasticity**,
  2. **Multicollinearity**,

  and to diagnostics and remedies for both in multiple regression.

Table of Contents

What is Heteroskedasticity?

In the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, n,$$

a standard assumption is constant conditional variance:

$$\mathrm{Var}(u_i \mid x_{i1}, \ldots, x_{ik}) = \sigma^2 \quad \text{for all } i.$$

- When the variance depends on the regressors,

  $$\mathrm{Var}(u_i \mid x_{i1}, \ldots, x_{ik}) = \sigma_i^2 = h(x_{i1}, \ldots, x_{ik}),$$

  the errors are **heteroskedastic**.
- The function $h(\cdot)$ is some positive, possibly unknown, function.

Consequences and Sources of Heteroskedasticity

Consequences (under exogeneity):

- OLS coefficients remain **unbiased** and **consistent**.
- But OLS is no longer the **most efficient** linear unbiased estimator.
- Usual OLS formulas for:
  - standard errors,
  - t-tests and F-tests,
  - confidence intervals

  are generally invalid.
- Conventional inference can be severely misleading.

Consequences and Sources of Heteroskedasticity

Typical economic sources:

- Cross-sectional heterogeneity (e.g. countries or firms of very different sizes).
- Omitted variables whose effects spill into u_i.
- Inappropriate functional forms (e.g. linear vs nonlinear relationships).
- Structural breaks over time or time-varying volatility (e.g. regime changes).

Heteroskedasticity in Simple Regression

Simple regression model:

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad i = 1, \ldots, n,$$

with

$$\mathrm{E}(u_i \mid x_i) = 0, \qquad \mathrm{Var}(u_i \mid x_i) = \sigma_i^2.$$

- **Homoskedasticity**: $\sigma_i^2 = \sigma^2$ for all i.
- **Heteroskedasticity**: variance changes with x_i,

  $$\mathrm{Var}(u_i \mid x_i) = h(x_i),$$

  for some function $h(\cdot)$.
- Under heteroskedasticity, OLS estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ remain:
  - linear,
  - unbiased,
  - but with different sampling variances.

Effect on the Variance of OLS Estimators

Slope estimator:

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Under heteroskedasticity:

$$\mathrm{Var}(\widehat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}.$$

Under homoskedasticity:

$$\mathrm{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Large variation in σ_i^2 can substantially affect precision.
- Observations with larger variance effectively carry less information.

Variance of the Intercept

Under homoskedasticity:

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Under heteroskedasticity:

$$\mathrm{Var}(\widehat{\beta}_1) = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}\, \omega_i \right)^2 \sigma_i^2,$$

where

$$\omega_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2}.$$

- Exact formulas involve the unknown σ_i^2.
- This motivates robust variance estimators and GLS-type methods.

Can Residuals Look Normal under Heteroskedasticity?

Key Question

If residuals look approximately normal in a histogram, can we still have heteroskedasticity?

- **Yes.** Normality refers to the *shape* of the distribution, not to constant variance.
- It is entirely possible that:
  - Residuals are roughly normal overall.
  - Their variance still changes systematically with x_i or $\widehat{y}_i$.
- A residual-vs-fitted plot can reveal a "fan-shaped" pattern even when the histogram looks normal.

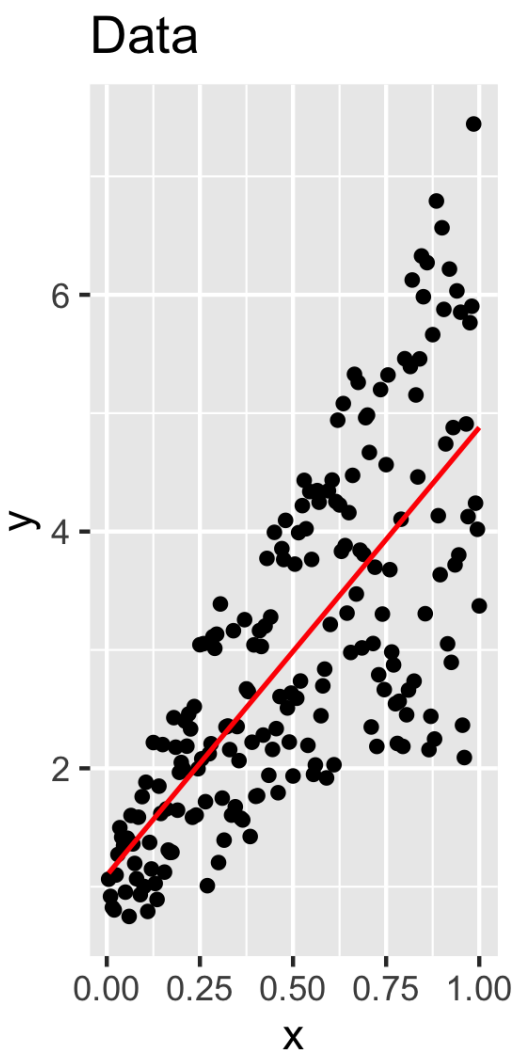
- Simulated examples: draw i.i.d. normal disturbances, then assign larger disturbances to larger fitted values.
- Histogram stays normal; residual-vs-fitted plot shows clear heteroskedasticity.

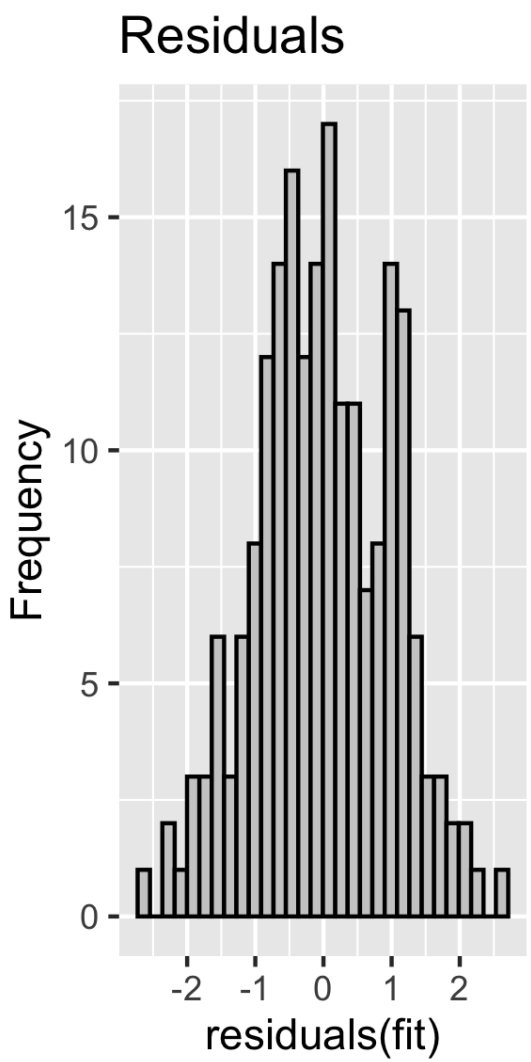
Example: Synthetic Heteroskedastic Data

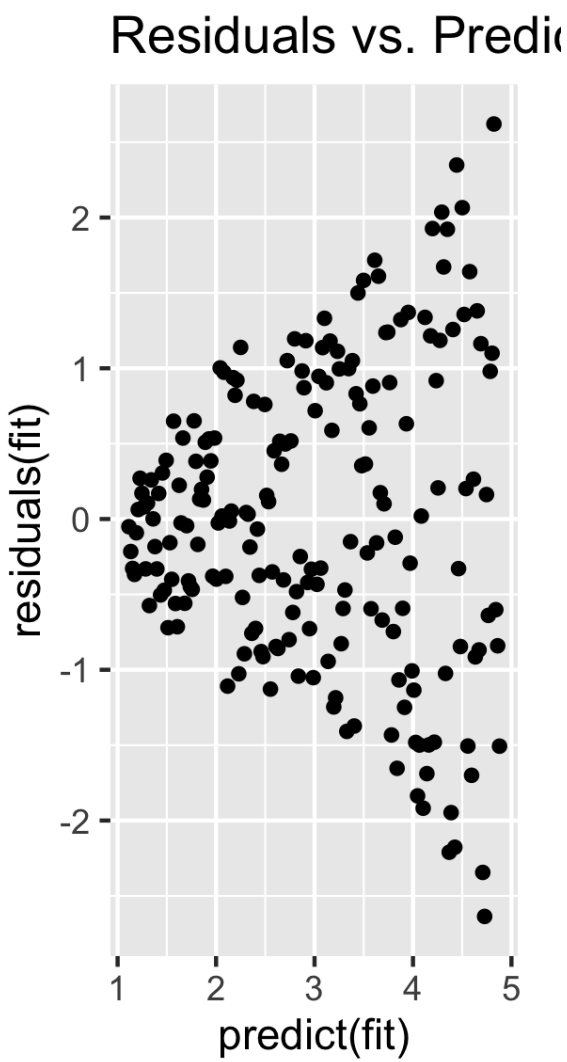
- Generate x on $[0, 1]$ and i.i.d. $N(0, 1)$ disturbances.
- Rearrange disturbances so that large shocks occur at large x, inducing heteroskedasticity.
- Estimate a simple regression $y_i = \beta_1 + \beta_2 x_i + u_i$.

- **Plot 1**: scatter of y vs x with fitted line.
- **Plot 2**: histogram of residuals (appears roughly normal).
- **Plot 3**: residuals vs fitted values, showing increasing spread.

Example: Synthetic Heteroskedastic Data







Example: Engel Curve Heteroskedasticity

Engel data: `engel` dataset from the `quantreg` package.

- Cross-section of working-class households.
- Variables: food expenditure (`foodexp`), household income (`income`).

- Simple regression: $\texttt{foodexp}_i = \beta_1 + \beta_2 \, \texttt{income}_i + u_i$.
- Plot reveals:
  - Increasing mean food expenditure with income.
  - Dispersion around the regression line grows with income.

Example: Engel Curve Heteroskedasticity

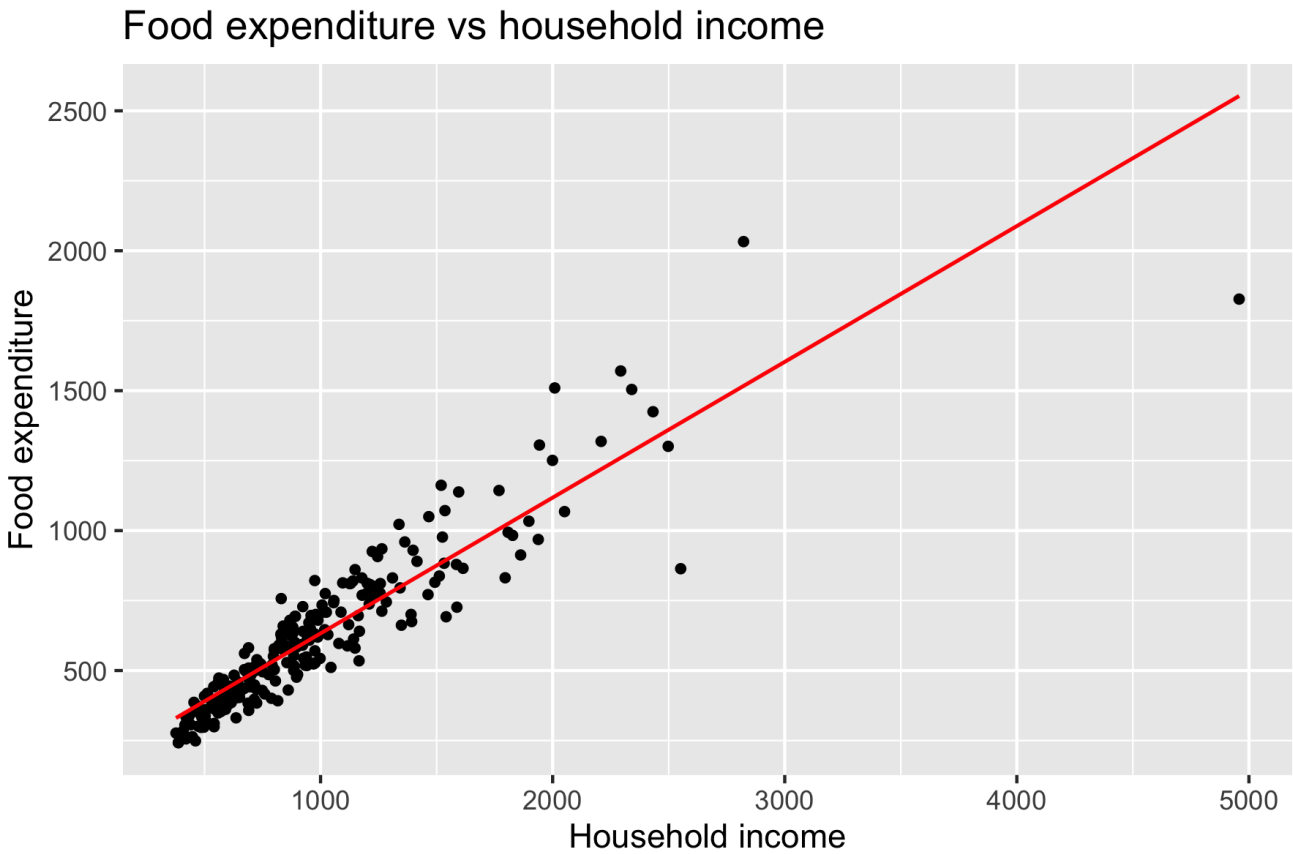


Figure: Food expenditure vs household income (`engel` data).

Formal Tests for Heteroskedasticity: Overview

We consider three classical tests:

- Breusch–Pagan / Lagrange Multiplier (LM) test.
- White test.
- Goldfeld–Quandt test.

General strategy:

- Estimate the original regression by OLS.
- Use OLS residuals to construct an *auxiliary* regression for squared residuals.
- Form a test statistic (often nR^2 or a variance ratio) with an asymptotic reference distribution.

Breusch–Pagan / LM Test

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i.$$

Assume:

$$\mathrm{Var}(u_i \mid z_{i2}, \ldots, z_{iS}) = \sigma_i^2 = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}).$$

Hypotheses:

$$H_0 : \alpha_2 = \cdots = \alpha_S = 0 \quad \Rightarrow \quad \sigma_i^2 = h(\alpha_1) \text{ (constant)}.$$

$$H_1 : \text{ at least one of } \alpha_2, \ldots, \alpha_S \neq 0.$$

Test procedure:

1. OLS on original model; obtain residuals $\hat{u}_i$.
2. Auxiliary regression:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i.$$

3. Let R^2 be from this auxiliary regression. The LM statistic is

$$\chi^2_{\mathrm{LM}} = nR^2.$$

Why is it Called an LM Test?

Lagrange Multiplier Perspective

- Start from a likelihood with variance depending on z-variables.
- Under homoskedasticity, restrictions $\alpha_2 = \cdots = \alpha_S = 0$ are imposed.
- LM principle: test whether relaxing these restrictions yields a significant gain in likelihood.
- In large samples, the LM statistic can be shown to be proportional to nR^2 from the auxiliary regression.

- This makes the Breusch–Pagan test easy to implement using standard regression output.
- We return to likelihood-based reasoning in the time-series chapters.

White Test

- The White test (White 1980) is more general:
  - Does not require specifying the exact functional form of the variance.

Example: mean function with two regressors

$$\mathrm{E}(y_i \mid x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

- Include as variance drivers:

$$z_{i2} = x_{i2}, \quad z_{i3} = x_{i3}, \quad z_{i4} = x_{i2}^2, \quad z_{i5} = x_{i3}^2,$$

  and possibly the interaction $x_{i2}x_{i3}$.
- Auxiliary regression: $\hat{u}_i^2$ on constant and $z_{i2}, \ldots, z_{i5}$.
- Test statistic:

$$\chi^2_{\text{White}} = nR^2,$$

  with asymptotic χ^2_{S-1} under homoskedasticity (where $S-1$ is the number of non-constant regressors in the auxiliary regression).

Goldfeld–Quandt Test: Groupwise Version

Model:

$$y_i = x_i'\beta + u_i, \quad i = 1, \ldots, n,$$

with x_i a $k \times 1$ regressor vector.

Null vs alternative:

$$H_0 : \mathrm{Var}(u_i) = \sigma^2 \text{ for all } i,$$

$$H_1 : \mathrm{Var}(u_i) = \sigma_g^2 \text{ differs across groups } g.$$

Goldfeld–Quandt Test: Groupwise Version

1. Split sample into two groups $g = 1, 2$ (e.g. by some observable).
2. Estimate the regression separately in each group and compute

$$\hat{\sigma}_g^2 = \frac{\text{SSR}_g}{n_g - k}.$$

3. Form

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2},$$

placing the larger variance in the numerator so $F \geq 1$.

4. Under H_0 and normal errors, $F \sim F_{n_2-k,\, n_1-k}$.

- Because large or small ratios are suspicious, a two-sided decision rule is appropriate.

Goldfeld–Quandt Test: Ordering by a Scale Variable

Often we suspect that the error variance is related to a *scale* variable z_i (e.g. income, size).

Procedure:

1. Order observations by z_i (smallest to largest).
2. Drop a middle block of observations to create two groups with low and high z_i.
3. Estimate the regression separately in each group; compute $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.
4. Form the variance ratio F and compare to an appropriate F-critical value.

- Dropping the middle enhances the contrast between groups, increasing power.
- But it reduces sample size and degrees of freedom in each group.

Dealing with Heteroskedasticity: Two Issues

When heteroskedasticity is present:

1. **Efficiency:**
   - OLS remains unbiased and consistent under exogeneity.
   - But OLS is no longer efficient; better linear estimators may exist.
2. **Inference:**
   - Usual OLS standard errors are incorrect.
   - Confidence intervals and tests based on them are not valid.

Two main approaches:

- Keep OLS coefficients but use **robust (heteroskedasticity-consistent) standard errors**.
- Use **GLS** or **Feasible GLS (FGLS)** when we model the error variance explicitly.

White's Heteroskedasticity-Consistent Standard Errors

Simple regression:

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad \mathrm{Var}(u_i \mid x_i) = \sigma_i^2.$$

Exact variances under heteroskedasticity:

$$\mathrm{Var}(\widehat{\beta}_1) = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}\, \omega_i \right)^2 \sigma_i^2, \quad \mathrm{Var}(\widehat{\beta}_2) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2},$$

with

$$\omega_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2}.$$

- White's idea: replace σ_i^2 by squared residuals e_i^2, and include a degrees-of-freedom correction.
- This yields **heteroskedasticity-consistent** variance estimators, often denoted HC0, HC1, HC2, HC3, etc.
- In multiple regression, compact matrix formulas exist and are implemented in standard software.

Properties of Robust Standard Errors

- Robust standard errors are asymptotically valid *whether or not* the errors are homoskedastic.
- They fix the *second* problem (invalid standard errors) while leaving OLS point estimates unchanged.
- Robust t- and F-tests are based on:

$$t_j = \frac{\widehat{\beta}_j}{\widehat{\text{se}}_{\text{robust}}(\widehat{\beta}_j)}.$$

- In practice:
  - OLS coefficients often change little,
  - but naive (homoskedastic) standard errors can be seriously understated.

GLS with Known Variance Function

Suppose:

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad \mathrm{Var}(u_i \mid x_i) = \sigma^2 x_i.$$

Define transformed variables:

$$\check{y}_i = \frac{y_i}{\sqrt{x_i}}, \quad \check{x}_{i1} = \frac{1}{\sqrt{x_i}}, \quad \check{x}_{i2} = \sqrt{x_i}, \quad \check{u}_i = \frac{u_i}{\sqrt{x_i}}.$$

Then:

$$\check{y}_i = \beta_1 \check{x}_{i1} + \beta_2 \check{x}_{i2} + \check{u}_i,$$

with

$$\mathrm{Var}(\check{u}_i \mid x_i) = \frac{\mathrm{Var}(u_i \mid x_i)}{x_i} = \frac{\sigma^2 x_i}{x_i} = \sigma^2.$$

- Transformed errors are homoskedastic.
- **GLS estimator**: apply OLS to the transformed regression.

GLS: General Weighting Idea

- If $\mathrm{Var}(u_i \mid w_i) = \sigma^2 w_i$ for known positive weights w_i:
  - Divide both sides and all regressors by $\sqrt{w_i}$.
  - Run OLS on the transformed data.
- Equivalent to weighted least squares with weights $1/w_i$.
- When the variance model is correct, GLS is more efficient than OLS.

- In matrix form, GLS solves:

$$\widehat{\beta}_{\mathrm{GLS}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}Y,$$

where $\Omega = \mathrm{Var}(u \mid \mathbf{X})$.

Feasible GLS (FGLS) with Unknown Variance Function

Assume:

$$\mathrm{Var}(u_i \mid x_i) = \sigma_i^2 = \sigma^2 x_i^{\gamma},$$

with unknown γ.

Take logs:

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \gamma \ln(x_i).$$

Let:

$$z_i = \ln(x_i), \quad \alpha_1 = \ln(\sigma^2), \quad \alpha_2 = \gamma,$$

then

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i.$$

- Approximate σ_i^2 with squared residuals u_i^2:

$$\ln(u_i^2) = \alpha_1 + \alpha_2 z_i + v_i.$$

- Replace u_i^2 with e_i^2 (OLS residuals) and estimate $\ln(e_i^2)$ on z_i by OLS.
- Obtain fitted variances:

$$\hat{\sigma}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 z_i).$$

Feasible GLS: Weighted Regression Step

With $\hat{\sigma}_i^2$ from the auxiliary regression:

Define:

$$\check{y}_i = \frac{y_i}{\hat{\sigma}_i}, \quad \check{x}_{i1} = \frac{1}{\hat{\sigma}_i}, \quad \check{x}_{i2} = \frac{x_i}{\hat{\sigma}_i}.$$

Run OLS on:

$$\check{y}_i = \beta_1 \check{x}_{i1} + \beta_2 \check{x}_{i2} + \check{u}_i.$$

- This two-step procedure yields **feasible GLS** estimates.
- If the variance model is reasonable, FGLS improves efficiency relative to OLS.

Case Study: Food Expenditure and Income (Engel)

OLS regression:

$$\texttt{foodexp}_i = \beta_1 + \beta_2\, \texttt{income}_i + u_i.$$

OLS results:

- $\widehat{\beta}_1 \approx 147.5$, $\widehat{\beta}_2 \approx 0.485$.
- Both coefficients highly significant.
- $R^2 \approx 0.83$: income explains a large share of variation in food expenditure.

Tests for heteroskedasticity:

- Breusch–Pagan (income as variance driver): $nR^2 \approx 109.3$, df $= 1$, $p \approx 0$.
- White test (income and income^2): $nR^2 \approx 181.1$, df $= 2$, $p \approx 0$.
- Strong evidence that error variance increases with income.

Case Study: Robust SEs and GLS for Engel Data

White's HC1 robust standard errors:

- Intercept SE: increases from ≈ 16.0 to ≈ 46.6.
- Slope SE: increases from ≈ 0.014 to ≈ 0.052.
- Point estimates unchanged, but uncertainty is much larger than homoskedastic OLS suggests.

GLS with known variance form $\mathrm{Var}(u_i|\texttt{income}_i) \propto \texttt{income}_i$**:**

- After transformation, slope $\widehat{\beta}_2 \approx 0.54$.
- Standard error is smaller; RL fit is tighter (RSE ≈ 2.9 on transformed scale).

Feasible GLS with estimated variance function:

- Auxiliary regression $\log(e_i^2)$ on $\log(\texttt{income}_i)$ yields $\gamma > 0$ and highly significant.
- FGLS slope $\widehat{\beta}_2 \approx 0.57$ with SE ≈ 0.015.
- Weighted regression attains even higher R^2 (on the weighted scale) and tighter standard errors.

Case Study: Wages and Groupwise Heteroskedasticity

CPS1985 wage data:

- OLS regression:

$$\texttt{wage}_i = \beta_0 + \beta_1 \texttt{education}_i + \beta_2 \texttt{experience}_i + u_i.$$

- Both education and experience have positive, significant effects on wages.

Goldfeld–Quandt-type test (male vs female):

- Separate regressions for males and females; compute residual variances.
- Variance ratio $F \approx 1.22$, with a two-sided p-value ≈ 0.10.
- Not strong evidence at 5% level, but suggests some difference in dispersion.

Groupwise FGLS:

- Assign group-specific $\hat{\sigma}_i$ based on gender.
- Divide wages and regressors by $\hat{\sigma}_i$ and run OLS.
- Coefficients close to OLS, but with slightly different standard errors reflecting groupwise heteroskedasticity.

Takeaways on Heteroskedasticity

- Heteroskedasticity is ubiquitous in economic and financial data, especially cross-sections.
- It does not bias OLS under exogeneity, but:
  - invalidates usual standard errors and tests,
  - reduces efficiency of OLS.
- Formal tests (Breusch–Pagan, White, Goldfeld–Quandt) are easy to implement and should accompany graphical diagnostics.
- Robust (White) standard errors fix inference while keeping OLS coefficients.
- GLS and FGLS can improve efficiency when we are willing to model the variance.
- In practice, a combination of:
  - careful diagnostics,
  - robust inference,
  - and, where appropriate, GLS-type methods

  yields more reliable regression analysis.

Table of Contents

What is Multicollinearity?

- Multiple regression model in matrix form:

$$y = \mathbf{X}\beta + u, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}.$$

- **Multicollinearity** (Frisch, 1934): one or more exact or near-exact linear relationships among explanatory variables.
- In matrix terms: linear dependence or near-dependence among the columns of $\mathbf{X}$.
- Consequences:
  - $\mathbf{X}'\mathbf{X}$ nearly singular.
  - In the extreme (perfect collinearity), $\mathbf{X}'\mathbf{X}$ is singular and OLS is not uniquely defined.

Historical Note: Ragnar Frisch

Ragnar Frisch (1895–1973)

- Norwegian economist and one of the founders of modern econometrics.
- Coined the term "econometrics"; helped shape micro vs macro distinction.
- Key contributions:
  - Consumer and production theory (Frisch elasticities).
  - Regression and time-series analysis; Frisch–Waugh–Lovell theorem.
  - Dynamic macro models and impulse–propagation analysis.
- Co-founded the Econometric Society and was the first editor of *Econometrica*.
- Shared the inaugural Nobel Prize in economics (1969) with Jan Tinbergen.

Exact vs Approximate Multicollinearity

Exact multicollinearity

- There exists nonzero $a \in \mathbb{R}^{k+1}$ such that

$$\mathbf{X}a = 0.$$

- At least one column is an exact linear combination of others.
- Examples:
  - Dummy and its complement with an intercept (e.g. MALE_i and $1 - \text{MALE}_i$).
  - Including both parts and their total: $\text{TOTAL}_i = \text{PART1}_i + \text{PART2}_i$.
- $\mathbf{X}'\mathbf{X}$ is singular; normal equations have no unique solution.
- Software typically drops perfectly collinear regressors.

Approximate (Near) Multicollinearity

Approximate multicollinearity

- Linear relationships are only approximate:

$$x_{ij} \approx c_0 + c_1 x_{i\ell_1} + \cdots + c_m x_{i\ell_m},$$

  with small but nonzero deviations.

- $\mathbf{X}'\mathbf{X}$ remains invertible but has very small eigenvalues.
- OLS exists:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y,$$

  but:

  - Sampling variance can be very large in directions linked to small eigenvalues.
  - Estimates become unstable and sensitive to minor data changes.

Consequences for OLS

Under Gauss–Markov assumptions:

$$\mathrm{Var}(\widehat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

1. **Inflated variances and SEs**
   - Highly correlated regressors make it hard to distinguish separate effects.
   - Relevant diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ become large.
2. **Unstable estimates**
   - Small data changes can cause large shifts or sign flips in coefficients.
3. **Weak individual significance**
   - High R^2 and significant overall F-test,
   - but individual t-tests insignificant due to large SEs.
4. **Predictive fragility**
   - In-sample fit good, but predictions for new covariate combinations can be unreliable.

A Simple Two-Regressor Illustration

Consider standardized regressors x_2, x_3 with correlation ρ_{23}.

Under homoskedasticity:

$$\mathrm{Var}(\widehat{\beta}_2 \mid \mathbf{X}) \propto \frac{1}{1-\rho_{23}^2}.$$

- As $|\rho_{23}| \to 1$, the denominator $\to 0$.
- The variance of $\widehat{\beta}_2$ explodes.
- Same idea generalizes via small eigenvalues of $\mathbf{X}'\mathbf{X}$ in higher dimensions.

Diagnosing Multicollinearity: Correlation Matrix

Correlation matrix of regressors

- First step: examine pairwise correlations (often after centering/scaling).
- Large $|\text{corr}|$ between two columns suggests potential collinearity.

Limitations

- Multicollinearity can involve *combinations* of three or more regressors.
- Pairwise correlations may look moderate even when a subset is nearly collinear.
- Correlation matrices are useful but not sufficient on their own.

Variance Inflation Factors (VIFs)

For each regressor x_j (excluding intercept):

$$x_{ij} = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{j-1} x_{i,j-1} + \gamma_{j+1} x_{i,j+1} + \cdots + \gamma_k x_{ik} + v_{ij},$$

and let R_j^2 be the R^2 from this regression.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

- If x_j is orthogonal to other regressors: $R_j^2 = 0$, $\text{VIF}_j = 1$.
- If x_j is nearly a linear combination: $R_j^2 \approx 1$, VIF_j large.
- Rules of thumb:
  - VIF > 5 or > 10 often viewed as problematic.
  - But interpretation must reflect economic context and model purpose.

Condition Number and Eigenvalue Diagnostics

Let $\mathbf{X}^*$ be centered/scaled regressors and

$$\mathbf{R} = \frac{1}{n}(\mathbf{X}^*)'\mathbf{X}^*.$$

Suppose eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0.$$

- Condition index for λ_j:

$$\kappa_j = \sqrt{\frac{\lambda_1}{\lambda_j}}.$$

- Overall condition number:

$$\kappa = \max_j \kappa_j = \sqrt{\frac{\lambda_1}{\lambda_p}}.$$

Condition Number and Eigenvalue Diagnostics

Interpretation

- $\kappa > 10$: moderate collinearity.
- $\kappa \in [30, 100]$: serious collinearity.
- $\kappa > 100$: near-singular design matrix.

Eigenvalue Diagnostics (Cont.)

- Condition indices are often paired with *variance-decomposition proportions*:
  - Decompose variance of each coefficient across eigen-directions.
- If several coefficients draw a large share of their variance from small eigenvalues:
  - Those coefficients are jointly affected by multicollinearity.
- This helps identify which variables are entangled in collinear relationships.

Auxiliary F-tests and Sensitivity Checks

Auxiliary F-tests

- Test subsets of coefficients jointly:

$$H_0 : \beta_j = 0 \text{ for } j \in J,$$

  using F-tests.

- Pattern: joint F significant, individual t-tests insignificant $\Rightarrow$ likely multicollinearity among that subset.

Sensitivity checks

- Examine how coefficients change when:
  - adding/removing regressors,
  - slightly modifying the data.
- Large swings in estimates or sign reversals suggest unstable estimation due to multicollinearity.

Remedies: Centering and Scaling

Centering and scaling

- Center: subtract sample mean $\bar{x}_j$.
- Scale: divide by standard deviation s_j.
- Useful when:
  - polynomials (e.g. x, x^2) or interactions are present,
  - regressors have very different scales.
- Example: replace x, x^2 with $(x - \bar{x})$, $(x - \bar{x})^2$ to reduce correlation.
- Centering/scaling improve numerical stability and interpretation, but do *not* fundamentally remove multicollinearity.

Remedies: Redefining or Combining Variables

Redefinition and aggregation

- Drop one of a pair of near-duplicates, if theory does not distinguish them.
- Aggregate highly correlated components into an index or factor:
  - e.g. combine similar risk measures into a single factor.
- Replace levels by more meaningful transformations:
  - ratios, spreads, principal components, etc.

- Aim: capture underlying variation with fewer, better behaved predictors.
- Often reduces collinearity with minor loss of explanatory power.

Remedies: Prior Information and Restrictions

- Economic theory may impose linear restrictions on coefficients:
  - equality, proportionality, or sum-to-one constraints.
- Imposing restrictions reduces effective dimensionality and can stabilize estimates.
- Methods:
  - restricted least squares,
  - Bayesian regression with informative priors (regularization through priors).
- When theory is credible, such restrictions are more than just statistical devices.

Remedies: Regularization and Dimension Reduction

Regularization

- **Ridge regression**:
  - Adds $\lambda \|\beta\|_2^2$ penalty.
  - Shrinks coefficients; inflates diagonal of $\mathbf{X}'\mathbf{X}$.
  - Reduces variance at cost of bias; stabilizes inversion.
- **Lasso / Elastic Net**:
  - Combine shrinkage and variable selection.
  - In presence of strong collinearity, may select one representative from a cluster.

Dimension reduction

- **Principal components regression (PCR)**:
  - Use principal components of regressors as predictors.
  - Discard components associated with very small eigenvalues.
- These methods are particularly useful when number of predictors is large.

Remedies: More or Better Data

- Multicollinearity is a property of the joint distribution of regressors in the sample.
- If regressors move almost in lockstep in observed data:
  - no purely statistical trick can fully separate their effects.
- Potential solutions:
  - Collect more data, especially where regressors vary more independently.
  - Redesign data collection (experiments, quasi-experiments) to generate independent variation.

Multicollinearity, Inference, and Model Purpose

- Multicollinearity mainly harms *inference on individual coefficients*.
- Some linear combinations of coefficients may still be well identified:
  - e.g. effect of a policy package, factor-mimicking portfolio.
  - Frisch–Waugh–Lovell helps understand which combinations are pinned down.
- For **prediction**:
  - multicollinearity is less problematic if:
    - predictions are made within the sample-like range,
    - regularization is used appropriately.
- For **causal interpretation** of individual coefficients:
  - severe multicollinearity makes ceteris paribus interpretations fragile,
  - often calls for model redesign or additional data.

Table of Contents

Fuel-Efficiency Regression: Setup

- Data: `mtcars` (built into `R`).
- Outcome: miles per gallon (`mpg`).
- Predictors:
  - numeric: `wt`, `hp`, `disp`, `drat`, `qsec`;
  - factors: `cyl`, `am` (Auto/Manual).
- Model:

$$\texttt{mpg}_i = \beta_0 + \beta_{\text{wt}}\texttt{wt}_i + \beta_{\text{hp}}\texttt{hp}_i + \cdots + \beta_{\text{cyl}}\texttt{cyl}_i + \beta_{\text{am}}\texttt{am}_i + u_i.$$

- Diagnostics:
  - Multicollinearity (VIF/GVIF, condition number),
  - Heteroskedasticity (BP, NCV, "White-like" test),
  - Robust SEs (HC3),
  - Functional form (RESET),
  - Influence (Cook's distance),
  - Correlation structure.

OLS Fit and Multicollinearity

OLS summary

- $R^2 \approx 0.875$, $\bar{R}^2 \approx 0.831$: high explanatory power.
- Overall F-test highly significant.
- But few individual coefficients are strongly significant.

VIF / GVIF diagnostics

- Adjusted $\text{GVIF}^{1/(2\text{Df})}$ for key regressors: between ≈ 1.8 and ≈ 3.7.
- Indicates moderate multicollinearity, but not extreme.

Condition number

- Condition number (scaled X) ≈ 10.2.
- Typical interpretation: moderate collinearity; standard errors can be inflated, but estimates are not wildly unstable.

Heteroskedasticity and Robust Inference

Heteroskedasticity tests

- Baseline Breusch–Pagan (vs fitted values): $p \approx 0.21$.
- Non-constant variance score test (NCV): $p \approx 0.11$.
- "White-like" BP using $\widehat{y}_i$ and $\widehat{y}_i^2$: $p \approx 0.035$.
- Mixed evidence; richer variance function suggests some heteroskedasticity.

HC3 robust standard errors

- Computed via `vcovHC(fit, type = "HC3")` and `coeftest()`.
- OLS coefficients largely unchanged.
- Only vehicle weight remains clearly significant at 5% level under HC3: $\widehat{\beta}_{\text{wt}} \approx -3.43$, $p \approx 0.043$.
- Robust SEs guard against misspecified variance.

Functional Form, Influence, and Correlations

Functional form

- Ramsey RESET (powers of fitted values): $p \approx 0.19$.
- No strong evidence of omitted nonlinearities or interactions in this specification.

Influence diagnostics

- Standard residual plots: residuals vs fitted, QQ, scale–location, leverage.
- Cook's distances highlight a few influential cars (e.g. *Chrysler Imperial*, *Lotus Europa*).
- These points merit inspection but do not dominate the overall fit.

Correlation structure

- Strong correlations among numeric predictors: $\mathrm{Corr}(\texttt{disp}, \texttt{wt}) \approx 0.89$, etc.
- Consistent with VIFs and condition number: size and power variables move together, explaining wide SEs on individual slopes.

Lessons from the Fuel-Efficiency Case Study

- High R^2 and significant overall F-test do not guarantee precise inference on individual coefficients.
- Moderate multicollinearity among size and power variables inflates SEs and makes some coefficients fragile.
- Heteroskedasticity tests motivate robust standard errors:
  - HC3 SEs change which coefficients are deemed significant.
- Functional-form tests and influence diagnostics help ensure that results are not driven by:
  - misspecified relationships, or
  - a handful of outlying or high-leverage observations.
- Overall, this example shows how multiple diagnostics work together to move from a naïve OLS fit to a *scrutinized* model with transparent assumptions.

Table of Contents

Chapter Summary: Estimation and Inference

- **Simple linear regression**
  - Derived OLS estimators by minimizing squared residuals.
  - Established key properties: regression line passes through $(\bar{x}, \bar{y})$, residuals sum to zero, orthogonality of regressors and residuals.
  - Introduced sampling uncertainty via t-statistics and the t-distribution.
  - Defined measures of fit: R^2, adjusted R^2, residual standard error (RSE).
- **Multiple regression**
  - Matrix formulation: $Y = \mathbf{X}\beta + u$, $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
  - Gauss–Markov conditions under which OLS is BLUE.
  - Joint significance via general F-tests and their link to SST, SSE, SSR.

Chapter Summary: Diagnostics and Remedies

- **Heteroskedasticity**
  - Defined as non-constant conditional variance.
  - Effects: OLS remains unbiased under exogeneity, but is inefficient; usual SEs and tests are invalid.
  - Diagnostics: graphical checks, Breusch–Pagan, White-type tests, Goldfeld–Quandt.
  - Remedies:
    - Robust (HC/White) covariance estimators.
    - GLS and Feasible GLS when variance can be modelled.
- **Multicollinearity**
  - Defined as (near) linear dependence among regressors.
  - Consequences: inflated variances, unstable coefficients, fragile inference on individual slopes.
  - Diagnostics: correlation matrices, VIFs/GVIFs, condition numbers, sensitivity checks.
  - Remedies: centering/scaling, variable redefinition, prior restrictions, regularization, more or better data.

Chapter Summary: Case Studies and Practice

- **Case studies** illustrated how tools interact in practice:
  - `mtcars`: simple regression of fuel efficiency on weight; ANOVA and overall F-test.
  - Engel curves: strong income-dependent heteroskedasticity; robust SEs and GLS/FGLS.
  - CPS wages: groupwise heteroskedasticity; Goldfeld–Quandt ideas and groupwise FGLS.
  - Multivariate fuel-efficiency model: combined diagnostics for multicollinearity, heteroskedasticity, functional form, and influence.
- **Overarching message**
  - A single OLS printout is not the end of the analysis.
  - Robust empirical work requires:
    - transparent assumptions,
    - systematic diagnostics,
    - appropriate corrections and model refinement.

Table of Contents

References I

 White, Halbert (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *Econometrica* 48.4, pp. 817–838.

Chapter 2 — Univariate Time Series

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Univariate Time Series: Overview

- We study a single stochastic process $\{Y_t\}$ observed over time.
- Key probabilistic foundations:
  - **Stationarity**: mean, variance, and autocovariances do not change over time.
  - **Ergodicity**: time averages converge to ensemble averages.
  - **Mixing**: dependence between distant observations decays.
- Wold decomposition: any stationary process can be written as
  - deterministic component + linear filter of white noise.
  - motivates ARMA-type models.

Univariate Time Series: Overview

- ARMA models:
  - lag-operator representation, ACF/PACF patterns, causality and invertibility.
  - estimation (Yule–Walker, maximum likelihood), order selection (AIC, BIC), forecasting.
- Applications: macro forecasting, asset pricing, risk management, volatility modeling, and many other fields where processes evolve over time.

Table of Contents

What Do We Mean by Stationarity?

- Goal of time-series analysis:
  - learn the dynamics behind past data,
  - use them to predict future values.
- Informally: future observations are "probabilistically similar" to past ones.
- This idea is captured by **stationarity**.

What Do We Mean by Stationarity?

Strict Stationarity

$\{Y_t\}$ is strictly stationary if, for any integer k and any time shift h, the joint distribution of $(Y_t, Y_{t-1}, \ldots, Y_{t-k+1})$ is the same as that of $(Y_{t+h}, Y_{t+h-1}, \ldots, Y_{t+h-k+1})$.

- The whole joint distribution is invariant to time shifts.
- Very strong and usually unverifiable directly.

Weak (Covariance) Stationarity

In practice we use a weaker notion:

Weak (Covariance) Stationarity

$\{Y_t\}$ is weakly stationary if:

$$\mathrm{E}(Y_t) = \mu < \infty,$$

$$\mathrm{Var}(Y_t) = \gamma_0 < \infty \quad \text{for all } t,$$

$$\mathrm{Cov}(Y_t, Y_{t-k}) = \gamma_k \quad \text{depends only on lag } k, \text{ not on } t.$$

- γ_k is the **autocovariance** at lag k.
- Often we work with the autocorrelation function (ACF):

$$\rho_k = \gamma_k / \gamma_0.$$

- Weak stationarity is enough for many LLN/CLT results and for ARMA modeling.

Strict vs Weak Stationarity: Two Quizzes

Quiz 1: Is every strictly stationary series also weakly stationary?

- **No.** Strict stationarity does not guarantee finite mean/variance.
- Example: Cauchy process:
  - each Y_t has a Cauchy distribution,
  - the distribution is time-invariant (strictly stationary),
  - but mean and variance are undefined, so weak stationarity fails.

Quiz 2: Can a process be weakly stationary but *not* strictly stationary?

- **Yes.** Example:
  - Let $X_{2t-1} = \varepsilon_t \sim N(0,1)$, $X_{2t} = \eta_t \sim U(-\sqrt{3}, \sqrt{3})$.
  - ε_t and η_t independent, both have mean 0 and variance 1.
  - Mean and autocovariances of X_t are time-invariant (weakly stationary).
  - But the marginal distributions alternate between normal and uniform, so strict stationarity fails.

Heavy-Tailed Distributions and Undefined Moments

- Not all pre-specified distributions have finite mean and variance.
- **Heavy tails** can lead to undefined or infinite moments.
- Examples:
  - **Cauchy** (Lorentz) distribution:
    - pdf: $f(x; x_0, \gamma) = \dfrac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}$.
    - No finite mean or variance.
  - **Lévy** and other stable laws: often lack finite mean/variance.
  - **Student** t:
    - df ≤ 1: mean undefined;
    - df ≤ 2: variance undefined.
  - **Pareto**:
    - shape $\alpha \leq 1$: mean infinite;
    - $\alpha \leq 2$: variance infinite.
- These examples show why finite-moment conditions must be checked explicitly.

Innovations: i.i.d., MDS, and White Noise

We often specify time series via a sequence of **innovations** $\{\varepsilon_t\}$.

- **i.i.d. innovations**:
  - $\varepsilon_t \sim \text{i.i.d.}(0, \sigma^2)$.
  - Strong assumption: independence and identical distribution.
- **Martingale difference sequence (MDS)**:

$$\mathrm{E}(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0,$$

  where $\mathcal{F}_{t-1}$ is past information.
  - No predictable component in the conditional mean,
  - but variance may be time-varying (e.g. ARCH/GARCH).

Innovations: i.i.d., MDS, and White Noise (cont.)

- **White noise** $\varepsilon_t \sim WN(0, \sigma^2)$:

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

  - Zero autocovariances, but not necessarily independent unless i.i.d.

- i.i.d. $\Rightarrow$ MDS (with finite moments) $\Rightarrow$ white-noise autocovariances.
- The converse directions generally do not hold.

Why Ergodicity and Mixing Matter

- In time series, observations are dependent; we cannot apply iid LLN/CLT directly.
- **Ergodicity**:
  - ensures that time averages converge to ensemble (population) averages,
  - justifies using a single long time series to learn about the underlying process.
- **Mixing conditions**:
  - quantify how fast dependence between past and future decays,
  - allow extending LLN and CLT to dependent data.
- In econometrics and finance:
  - serial correlation is the rule rather than the exception,
  - ergodicity and mixing underpin asymptotic inference and forecasting.

A Stationary but Non-Ergodic Example

Example

Let $Z \sim N(0,1)$ and define $Y_t = Z$ for all t.

- $\{Y_t\}$ is **strictly stationary**:
  - every finite-dimensional distribution is the distribution of repeated copies of Z,
  - mean and variance constant over time.
- But the sample mean

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t = Z$$

  does not converge to $\mathrm{E}(Z) = 0$; it is just a constant random draw.
- LLN fails: time average $\bar{Y}_T$ converges to Z, not to the population mean.
- Conclusion: stationarity alone is not enough; we also need some form of "forgetfulness" of the past—ergodicity.

Time Average vs Ensemble Average

- **Ensemble (cross-sectional) average**:

$$\overline{\overline{Y}} = \frac{1}{m} \sum_{i=1}^{m} Y_i,$$

  across independent copies of the process.

- **Time average**:

$$\bar{Y}_T = \frac{1}{T} \sum_{t=1}^{T} Y_t,$$

  along a single realization.

Time Average vs Ensemble Average

- In theory:
  - ensemble average is natural, but we rarely have many independent copies.
  - in practice, we usually have one long time series, so we rely on time averages.
- **Ergodicity in mean**:
  - ensures $\bar{Y}_T \xrightarrow{p} \mu = \mathrm{E}(Y_t)$,
  - so the sample mean is a consistent estimator of the true mean.

Variance of the Sample Mean and Ergodicity

For a covariance-stationary process $\{Y_t\}$ with mean μ, variance γ_0, and autocovariances γ_k:

$$\mathrm{Var}(\bar{Y}_T) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}(Y_t, Y_s) = \frac{\gamma_0}{T^2}\sum_{k=-(T-1)}^{T-1}(T-|k|)\rho_k,$$

where $\rho_k = \gamma_k/\gamma_0$.

- Rearranging:

$$\mathrm{Var}(\bar{Y}_T) = \frac{\gamma_0}{T}\sum_{k=-(T-1)}^{T-1}\left(1-\frac{|k|}{T}\right)\rho_k.$$

- If the autocorrelations decay sufficiently fast, the sum is finite and $\mathrm{Var}(\bar{Y}_T) \to 0$ as $T \to \infty$.
- Then $\bar{Y}_T$ is consistent for μ: $\bar{Y}_T \xrightarrow{p} \mu$.

Ergodicity in Mean: Sufficient Condition

Ergodicity for the Mean

A covariance-stationary process $\{Y_t\}$ is **ergodic for the mean** if

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu = \mathrm{E}(Y_t)$$

as $T \to \infty$.

Ergodicity in Mean: Sufficient Condition

Sufficient Condition (Absolute Summability)

A sufficient condition for ergodicity for the mean is:

$$\sum_{k=0}^{\infty} |\gamma_k| < \infty \quad \text{or equivalently} \quad \sum_{k=0}^{\infty} |\rho_k| < \infty.$$

- Intuition: long-lag autocorrelations must die out fast enough.
- Then LLN holds for the sample mean of a single, long realization.

Random Walk: Nonstationarity and Non-Ergodicity

Consider a random walk:

$$Y_t = Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } (0, \sigma^2).$$

- $\mathrm{E}(Y_t)$ may be constant if Y_0 fixed, but

$$\mathrm{Var}(Y_t) = t\sigma^2 \to \infty.$$

- The process is *not* covariance-stationary:
  - variance and covariances depend on t.
- Sample mean of a single random walk path does not stabilize around a fixed value.
- Different random-walk paths at a fixed time t have large cross-sectional variance.
- Random walks illustrate why nonstationary series violate LLN-type reasoning and why we need stationarity/ergodicity for meaningful averaging and inference.

From Independence to Ergodicity

For random variables X and Y:

$$X \perp Y \quad \Longleftrightarrow \quad \mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)]\mathrm{E}[g(Y)]$$

for all measurable functions f, g.

- This factorization extends naturally to time series.
- Idea of **ergodicity**:
  - as time separation grows, the joint behavior of "blocks" of the series factors into products of marginal expectations.
  - captures asymptotic independence of distant past and future.

From Independence to Ergodicity

Ergodicity (Informal)

A strictly stationary process $\{Y_t\}$ is ergodic if, for bounded functions f and g, applied to finite blocks of past and future, the joint expectation $\mathrm{E}[f(\text{past})g(\text{future})]$ factorizes into $\mathrm{E}[f(\text{past})]\mathrm{E}[g(\text{future})]$ in the limit as the time gap between the blocks $\to \infty$.

LLN and CLT Under Ergodicity

WLLN for Ergodic Stationary Processes

If $\{Y_t\}$ is stationary and ergodic with $\mathrm{E}(Y_t) = \mu$ and $\mathrm{E}|Y_t| < \infty$, then

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu \quad \text{as } T \to \infty.$$

LLN and CLT Under Ergodicity

CLT for Ergodic Stationary MDS

If $\{Y_t\}$ is stationary, ergodic, and a martingale difference sequence with $\mathrm{Var}(Y_t) = \sigma^2 < \infty$, then

$$\sqrt{T}\,\bar{Y}_T \Longrightarrow N(0, \sigma^2),$$

so

$$\sigma^{-1}\sqrt{T}\,\bar{Y}_T \Longrightarrow N(0, 1).$$

- These results justify standard asymptotic inference (confidence intervals, tests) for dependent but ergodic time series.

CLT for Linear Stationary Processes

Consider a linear process:

$$Y_t = \mu + \sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j},$$

where

- $\{\varepsilon_t\}$ i.i.d. with $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{E}(\varepsilon_t^2) < \infty$,
- $\sum_{j=0}^{\infty} |\xi_j| < \infty$.

CLT for Linear Stationary Processes

CLT for a Stationary Linear Process

Under these conditions,

$$\sqrt{T}\left(\bar{Y}_T - \mu\right) \Longrightarrow N\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right),$$

where γ_j are the autocovariances of $\{Y_t\}$.

- This covers many ARMA-type models.
- The long-run variance $\sum_j \gamma_j$ plays the role of asymptotic variance.

Mixing Conditions: Intuition

- Mixing conditions formalize the idea that the process *forgets* its past.
- Define $\mathcal{F}_a^b = \sigma(X_a, \ldots, X_b)$, the σ-algebra generated by observations from time a to b.
- For a process $\{X_t\}$, mixing coefficients measure how much:
  - probabilities of future events (based on $\mathcal{F}_{k+n}^{\infty}$) depend on past events (based on $\mathcal{F}_1^k$),
  - and how fast this dependence decays as $n \to \infty$.
- These coefficients are key in proving LLNs and CLTs for dependent sequences.

Alpha (Strong) Mixing

Alpha Mixing (Strong Mixing)

$\{X_t\}$ is **alpha mixing** if

$$\alpha(n) = \sup_{k \geq 1} \sup_{A \in \mathcal{F}_1^k,\, B \in \mathcal{F}_{k+n}^{\infty}} |P(A \cap B) - P(A)P(B)| \longrightarrow 0 \quad \text{as } n \to \infty.$$

Alpha (Strong) Mixing

- $\alpha(n)$ measures the maximal deviation from independence between past events A and future events B separated by n time units.
- As n grows, $\alpha(n) \to 0$ means the distant future becomes almost independent of the distant past.
- Widely used in establishing CLTs and other limit theorems for time series.

Beta (Absolute Regularity) and Phi Mixing

Beta Mixing (Absolute Regularity)

$\{X_t\}$ is **beta mixing** if

$$\beta(n) = \sup_{k \geq 1} \mathrm{E}\left[\sup_{B \in \mathcal{F}_{k+n}^{\infty}} |P(B \mid \mathcal{F}_1^k) - P(B)| \right] \longrightarrow 0 \quad \text{as } n \to \infty.$$

Beta (Absolute Regularity) and Phi Mixing

- Measures, on average, how far conditional probabilities of future events given the past are from their unconditional probabilities.

Phi Mixing (Uniform Mixing)

$\{X_t\}$ is **phi mixing** if

$$\phi(n) = \sup_{k \geq 1} \sup_{\substack{A \in \mathcal{F}_1^k,\, B \in \mathcal{F}_{k+n}^{\infty} \\ P(A) > 0}} \left| \frac{P(A \cap B)}{P(A)} - P(B) \right| \longrightarrow 0 \quad \text{as } n \to \infty.$$

- Stronger than alpha mixing: requires uniform convergence of $P(B \mid A)$ to $P(B)$.

Rho Mixing and Relationships Among Conditions

Rho Mixing

$\{X_t\}$ is **rho mixing** if

$$\rho(n) = \sup_{k \geq 1} \sup_{\substack{f \in L^2(\mathcal{F}_1^k),\, g \in L^2(\mathcal{F}_{k+n}^{\infty}) \\ \mathrm{Var}(f)>0,\, \mathrm{Var}(g)>0}} |\rho(f, g)| \longrightarrow 0 \quad \text{as } n \to \infty,$$

where $\rho(f, g)$ is the correlation between f and g.

- Focuses on *linear* dependence:
  - correlations between functions of past and future go to zero.

Rho Mixing and Relationships Among Conditions

Relationships (roughly):

- Phi mixing $\Rightarrow$ beta mixing $\Rightarrow$ alpha mixing.
- Rho mixing also implies alpha mixing.
- Converse implications generally fail.

- In practice, we choose conditions:
  - strong enough to get LLN/CLT,
  - weak enough to be satisfied by models (e.g. many ARMA, GARCH processes).

Table of Contents

Why the Wold Decomposition Matters

- Any covariance–stationary process can be split into:
  - a **deterministic component** (perfectly predictable from the past), and
  - a **stochastic component** that is a purely nondeterministic infinite moving average of white noise.
- This is the content of the **Wold decomposition**.
- Interpretation:
  - separates predictable structure from genuine shocks,
  - provides a theoretical foundation for ARMA and more general linear models,
  - guides forecasting: forecast the deterministic part; model and forecast the stochastic part.

Wold Decomposition: Statement

Theorem (Wold Decomposition)

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a covariance–stationary process with finite variance. Then there exist:

- a white-noise sequence $\{\varepsilon_t\}$ with $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2 \in (0, \infty)$, $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0$ for $t \neq s$,
- a linearly deterministic process $\{v_t\}$ (perfectly predictable from the infinite past),

such that

$$Y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + v_t,$$

with $\sum_{j=0}^{\infty} \theta_j^2 < \infty$.

- The stochastic part is a linear filter of white noise.
- The deterministic part v_t can be constant, periodic, or more general predictable structure.

Interpretation and Normalization

- Example: $v_t = v$, a (possibly random) constant with $\mathrm{E}(v) = 0, \mathrm{Var}(v) < \infty$.
  - Then $\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} v$ (WLLN).
- The decomposition is not unique unless we normalize:
  - we can absorb a scale factor into θ_j vs ε_t,
  - typically we normalize $\mathrm{Var}(\varepsilon_t) = 1$ and let the scale sit in $\{\theta_j\}$.
- **Innovations interpretation**:

$$\varepsilon_t = Y_t - \mathrm{E}\big(Y_t \mid Y_{t-1}, Y_{t-2}, \dots\big),$$

  so that $\mathrm{E}(\varepsilon_t \mid Y_{t-1}, Y_{t-2}, \dots) = 0$.
- The innovations are the *one-step-ahead forecast errors* under optimal linear prediction.

Linear Processes and ARMA Models

- If the deterministic component v_t is zero, we have a **purely nondeterministic** process:

$$Y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j},$$

  called a **linear process**.

- If the shocks $\{\varepsilon_t\}$ are i.i.d., this class includes:
  - all ARMA(p, q) models (with finite p, q),
  - many other models with exponentially decaying θ_j.
- For ARMA(p, q), the Wold representation has:
  - $v_t = 0$,
  - θ_j that decay geometrically fast.

Example: Long Memory via Wold Representation

Example (Linear Process with Power-Law Coefficients)

$$Y_t = \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j}, \quad \theta_j = c_j j^{-\gamma}, \quad |c_j| \le c,$$

with white-noise $\{\varepsilon_t\}$ and $\gamma > 0$.

Example: Long Memory via Wold Representation

- If $\gamma > \frac{1}{2}$, then $\sum_j \theta_j^2 < \infty$, so Y_t has finite variance.
- If $\frac{1}{2} < \gamma \leq 1$:
  - autocovariances γ_k decay so slowly that $\sum_{k=-\infty}^{\infty} |\gamma_k| = \infty$,
  - the process has **long memory**.
- If $\gamma > 1$:
  - autocovariances are absolutely summable,
  - the process is **short memory**.
- Wold's representation lets us link the decay of θ_j to long- vs short-run dependence.

Wold Decomposition in Practice: AR(1) as MA(∞)

- Take a stationary AR(1) process:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \quad |\phi| < 1.$$

- Its Wold representation has $v_t = 0$ and

$$Y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j},$$

  i.e. an MA(∞) process.
- In practice, we can approximate this by a high-order MA model (e.g. MA(10)).

Wold Decomposition in Practice: AR(1) as MA(∞)

Implementation idea in R:

- simulate AR(1) data,
- fit an MA(10) model to the simulated series,
- compare the original AR(1) series with the MA(10) fitted values.

- The MA(10) approximation tracks the AR(1) series closely.
- This illustrates Wold's idea: a causal AR model can be viewed as a (possibly infinite) moving average of shocks.

Wold Decomposition: AR(1) vs MA(10)

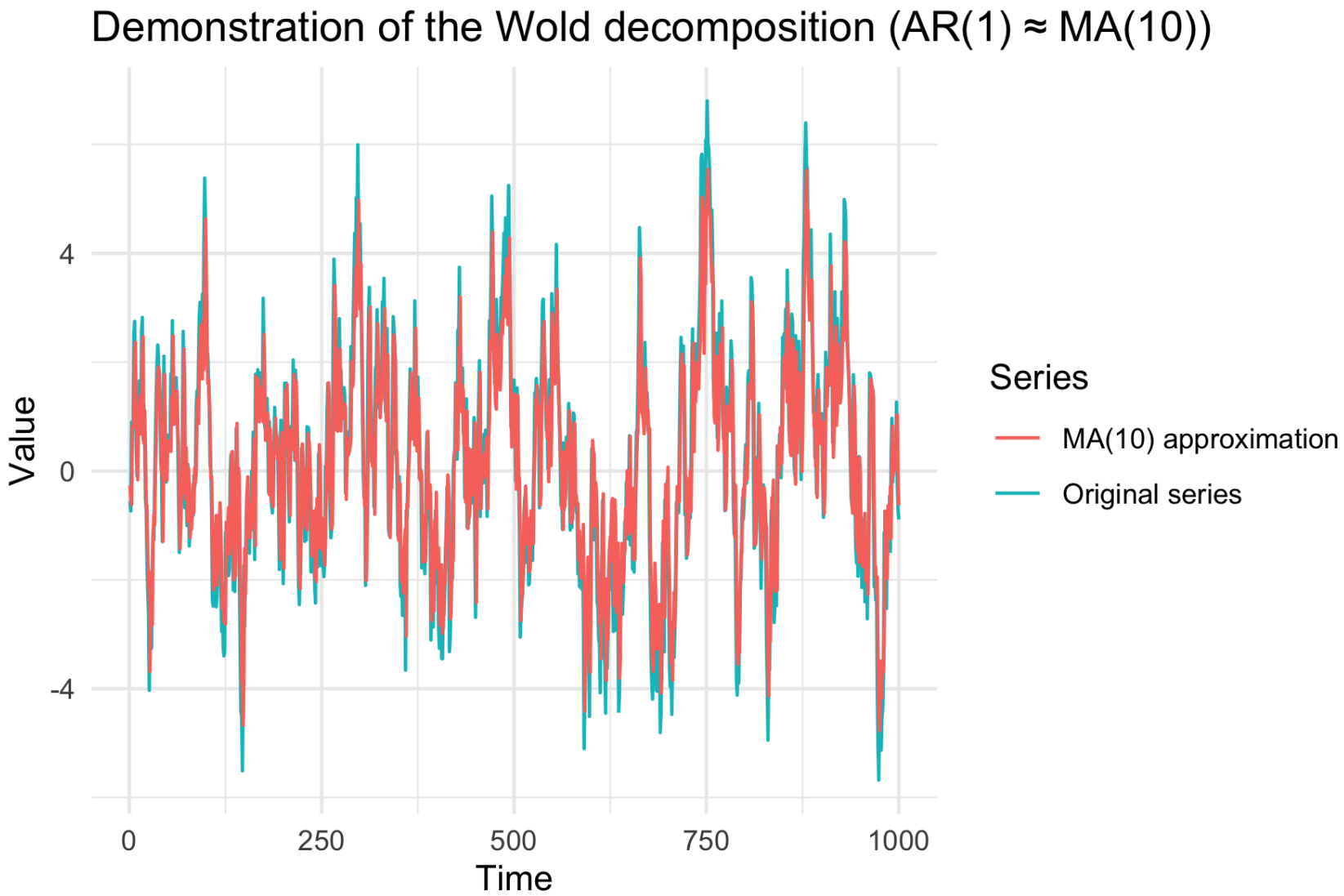


Figure: AR(1) process and its MA(10) approximation (finite order Wold

Autocovariances and Long-Run Variance from Wold

Assume $v_t = 0$ and

$$Y_t = \sum_{j=0}^{\infty} \delta_j \varepsilon_{t-j}, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2.$$

Autocovariance at lag k:

$$\mathrm{E}(Y_t Y_{t-k}) = \sigma^2 \sum_{j=0}^{\infty} \delta_j \delta_{j+k}.$$

Long-run variance (LRV):

$$\mathrm{lrvar}(Y_t) = \sum_{k=-\infty}^{\infty} \mathrm{E}(Y_t Y_{t-k}) = \sigma^2 \left(\sum_{j=0}^{\infty} \delta_j \right)^2.$$

- Only uncorrelatedness of shocks is needed for these second-order properties (white-noise assumption).
- i.i.d. shocks are stronger and allow higher-moment results.

Skewness and Kurtosis in Linear Processes

With i.i.d. shocks ε_t and linear process $Y_t = \sum \delta_j \varepsilon_{t-j}$:

$$\kappa_3(Y) = \frac{\mathrm{E}(Y_t^3)}{\mathrm{E}(Y_t^2)^{3/2}} = \kappa_3(\varepsilon) \frac{\sum_{j=0}^{\infty} \delta_j^3}{\left(\sum_{j=0}^{\infty} \delta_j^2\right)^{3/2}},$$

$$\kappa_4(Y) = \frac{\mathrm{E}(Y_t^4)}{\mathrm{E}(Y_t^2)^2} - 3 = \kappa_4(\varepsilon) \frac{\sum_{j=0}^{\infty} \delta_j^4}{\left(\sum_{j=0}^{\infty} \delta_j^2\right)^2},$$

where κ_3 is skewness and κ_4 is excess kurtosis.

Skewness and Kurtosis in Linear Processes

- Inequalities:

$$\sum_{j=0}^{\infty} \delta_j^4 \leq \Big(\sum_{j=0}^{\infty} \delta_j^2\Big)^2, \quad \Big(\sum_{j=0}^{\infty} \delta_j^3\Big)^2 \leq \Big(\sum_{j=0}^{\infty} \delta_j^2\Big)^3.$$

- Hence $|\kappa_j(Y)| \leq |\kappa_j(\varepsilon)|$ for $j = 3, 4$:
  - linear filtering *attenuates* skewness and kurtosis.
  - sign of skewness can change depending on $\sum \delta_j^3$.
- If ε_t is Gaussian, so is Y_t, and both skewness and excess kurtosis are zero.

Linear vs Nonlinear: A Subtle Boundary

- Linear processes are convenient and tractable.
- However, Bickel and Bühlmann (1997) show:
  - the set of linear processes is *not closed*: sequences of linear processes can converge to nonlinear processes.
  - the "boundary" between linear and nonlinear processes lies outside the linear class.
- Implication:
  - even with infinitely large samples, no test can perfectly distinguish all linear from all nonlinear processes.
  - in practice, we focus on whether a linear model is an *adequate approximation* for the task at hand.

Table of Contents

ARMA Models: Motivation

- **ARMA** = Autoregressive Moving Average.
- General ARMA(p, q) model combines:
  - autoregressive (AR) part: dependence on past Y_t,
  - moving average (MA) part: dependence on past shocks.
- Introduced and formalized in work by Whittle (1951) and popularized by Box–Jenkins.
- Wold decomposition motivates ARMA as:
  - parametric approximations to the infinite MA representation of stationary processes.

Why Stationarity, Ergodicity, and Mixing Matter for ARMA

- **Stationarity**:
  - mean, variance, autocorrelation structure are time-invariant.
  - ARMA assumes the data-generating mechanism does not change over time.
- **Ergodicity**:
  - ensures time averages estimate population moments,
  - critical for consistent estimation from a single time series.
- **Mixing conditions**:
  - guarantee dependence decays sufficiently fast,
  - underpin consistency and asymptotic normality of estimators (e.g. MLE, Yule–Walker).
- **Wold decomposition**:
  - ensures any stationary process can be represented as deterministic + linear filter of noise,
  - ARMA models approximate this filter with finite-order polynomials.

Autoregressive (AR) Models: AR(1)

AR(1) model:

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t,$$

where

- Y_t: time series,
- c: constant,
- ϕ_1: autoregressive coefficient,
- ε_t: white-noise error term.

- Captures dependence on the immediate past.
- Stationarity requires $|\phi_1| < 1$ (for a standard AR(1) with constant).
- Interpretation:
  - positive ϕ_1: persistence,
  - negative ϕ_1: mean-reversion with oscillation.

White Noise Revisited

White Noise $\{\varepsilon_t\}$

- $\mathrm{E}(\varepsilon_t) = 0$.
- $\mathrm{Var}(\varepsilon_t) = \sigma^2$ (constant over time).
- Autocovariance:

$$\gamma(\tau) = \mathrm{E}[\varepsilon_t \varepsilon_{t-\tau}] = \begin{cases} \sigma^2, & \tau = 0, \\ 0, & \tau \neq 0. \end{cases}$$

- No linear predictability from past values.
- Frequency-domain view: equal power at all frequencies (hence "white" noise).
- Serves as the building block for ARMA and Wold-type representations.

Higher-Order AR(p) Models

AR(p) model:

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

- Captures dependence on multiple lags.
- Stationarity is governed by the *roots* of the characteristic polynomial:

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

- Standard condition:
  - all roots of $\phi(z) = 0$ lie outside the unit circle.
- Under stationarity, AR(p) processes have:
  - exponentially decaying autocorrelations,
  - a Wold representation as an MA(∞).

Moving Average (MA) Models: MA(1) and MA(q)

MA(1) model:

$$Y_t = \mu + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

with white-noise ε_t.

MA(q) model:

$$Y_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t.$$

- Each Y_t is a finite linear combination of current and past shocks.
- Autocorrelation function (ACF) of MA(q):
  - nonzero only up to lag q; exactly zero at larger lags.
- Unlike AR, MA models are automatically stationary for finite q, but have invertibility conditions on θ-polynomials.

Example: Simulated MA(2) Process

- Consider an MA(2):

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

  with $\varepsilon_t \sim N(0, 1)$.

- Example parameter values:

$$\theta_1 = 0.5, \quad \theta_2 = -0.3.$$

- In R:
  - simulate white noise,
  - recursively construct Y_t using the MA(2) formula,
  - plot the resulting time series.

Example: Simulated MA(2) Process

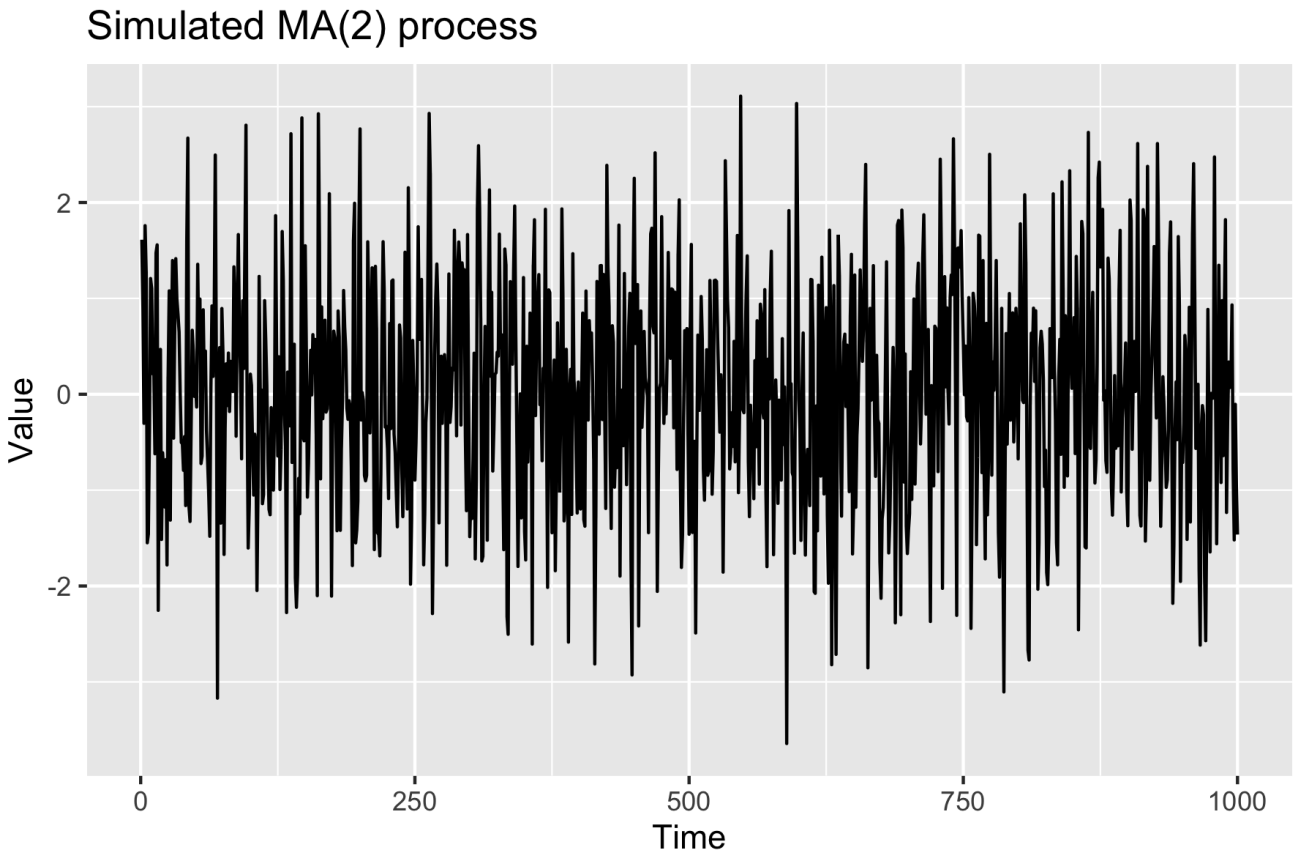


Figure: Simulated MA(2) process ($\theta_1 = 0.5,\ \theta_2 = -0.3$).

Lag Operator: Basics

- For a time series $\{Y_t\}$, define the lag operator L by

$$LY_t = Y_{t-1}.$$

- More generally:

$$L^p Y_t = Y_{t-p}, \quad L^0 Y_t = Y_t, \quad L^{-1} Y_t = Y_{t+1}.$$

- Constants are unaffected: $L\mu = \mu$.
- The lag operator provides a compact notation for ARMA models and their extensions.

AR Models in Lag-Operator Form

- AR(1):

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t \quad \Leftrightarrow \quad (1 - \phi_1 L) Y_t = c + \varepsilon_t.$$

- AR(p):

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

can be written as

$$\phi(L) Y_t = c + \varepsilon_t, \quad \phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p.$$

- Similarly, MA(q) can be written using a lag polynomial $\theta(L)$ acting on ε_t.

Algebra of Lag Polynomials

- A general lag polynomial:

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

- Applying to Y_t:

$$\phi(L) Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}.$$

- Lag polynomials can be:
  - added, multiplied, and (under conditions) inverted.
- Invertibility of lag polynomials is crucial for:
  - representing AR as MA(∞),
  - representing MA as AR(∞),
  - ensuring uniqueness of ARMA representations.

Invertibility: Geometric Series Representation

Consider $1 - \phi L$ with $|\phi| < 1$.

Goal: Find $(1 - \phi L)^{-1}$ such that

$$(1 - \phi L)(1 - \phi L)^{-1} = 1.$$

Formally, we can write

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

so that

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

- In $\mathcal{L}^2$ (square-integrable series), one must show:
  - the partial sums $\sum_{j=0}^{J} \phi^j L^j Y_t$ form a Cauchy sequence,
  - the limit is a well-defined weakly stationary process.
- If these conditions hold, then the infinite series defines $(1 - \phi L)^{-1}$ as an operator.

Higher-Order Lag Polynomials and Factorization

Example: Quadratic lag polynomial

$$1 - \alpha_1 L - \alpha_2 L^2 = (1 - \lambda_1 L)(1 - \lambda_2 L),$$

where $|\lambda_j| < 1$ for $j = 1, 2$.

- Then
$$(1 - \alpha_1 L - \alpha_2 L^2)^{-1} = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}.$$
- Using the geometric series:
$$(1 - \lambda_j L)^{-1} = \sum_{j=0}^{\infty} \lambda_j^j L^j.$$
- Multiplying the series yields:
$$(1 - \alpha_1 L - \alpha_2 L^2)^{-1} = \sum_{j=0}^{\infty} L^j \Big(\sum_{k=0}^{j} \lambda_1^k \lambda_2^{j-k} \Big).$$
- Stationarity requires $|\lambda_1| < 1$ and $|\lambda_2| < 1$.

Characteristic Roots and Unit Roots

- Lag polynomial:
$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2.$$
- Characteristic equation: $\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2 = 0$.
- Roots z_1, z_2 satisfy:
$$\alpha(z) = (1 - \lambda_1 z)(1 - \lambda_2 z), \quad \lambda_j = 1/z_j.$$
- Stationarity:
  - $|\lambda_j| < 1 \Leftrightarrow |z_j| > 1$.
- If any $|z_j| \leq 1$, we have a **unit root** or worse:
  - polynomial is non-invertible,
  - associated process is nonstationary (unit-root process).

AR vs MA: Two Sides of the Same Coin

- Thanks to invertible lag polynomials, there is no fundamental difference between AR and MA processes:

$$Y_t = \phi(L)^{-1}\varepsilon_t \quad \text{or} \quad \varepsilon_t = \phi(L)Y_t.$$

- Example: AR(1)

$$Y_t = \phi Y_{t-1} + \varepsilon_t \quad \Rightarrow \quad (1-\phi L)Y_t = \varepsilon_t.$$

Under $|\phi| < 1$,

$$Y_t = (1-\phi L)^{-1}\varepsilon_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j},$$

i.e. an MA(∞).

- Different ARMA specifications can deliver essentially the same fit:
  - in practice, we prefer the **parsimonious** specification (fewest parameters) that fits well (see model selection discussion).

Table of Contents

ARMA(p, q) Model: Definition and Motivation

- **ARMA(p, q)** combines autoregressive and moving-average components:

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

  with $\{\varepsilon_t\}$ white noise.

- Parameters:
  - c: constant;
  - $\phi_1, \ldots, \phi_p$: AR coefficients;
  - $\theta_1, \ldots, \theta_q$: MA coefficients.
- Advantages vs pure AR or MA:
  - richer dynamics with relatively few parameters,
  - can capture both persistent dependence and short-lived shocks,
  - flexible but still interpretable.
- In R, we can simulate ARMA models using `arima.sim()` (e.g. ARMA(1,1)).

Example: Simulated ARMA(1,1)

- Example specification:

$$Y_t = 0.6\,Y_{t-1} + \varepsilon_t + 0.7\varepsilon_{t-1}, \quad \varepsilon_t \sim N(0,1).$$

- In R, `arima.sim(list(order = c(1,0,1), ar = 0.6, ma = 0.7), n = 1000)`.
- This series exhibits both:
  - autoregressive persistence, and
  - moving-average smoothing of shocks.

Example: Simulated ARMA(1,1)

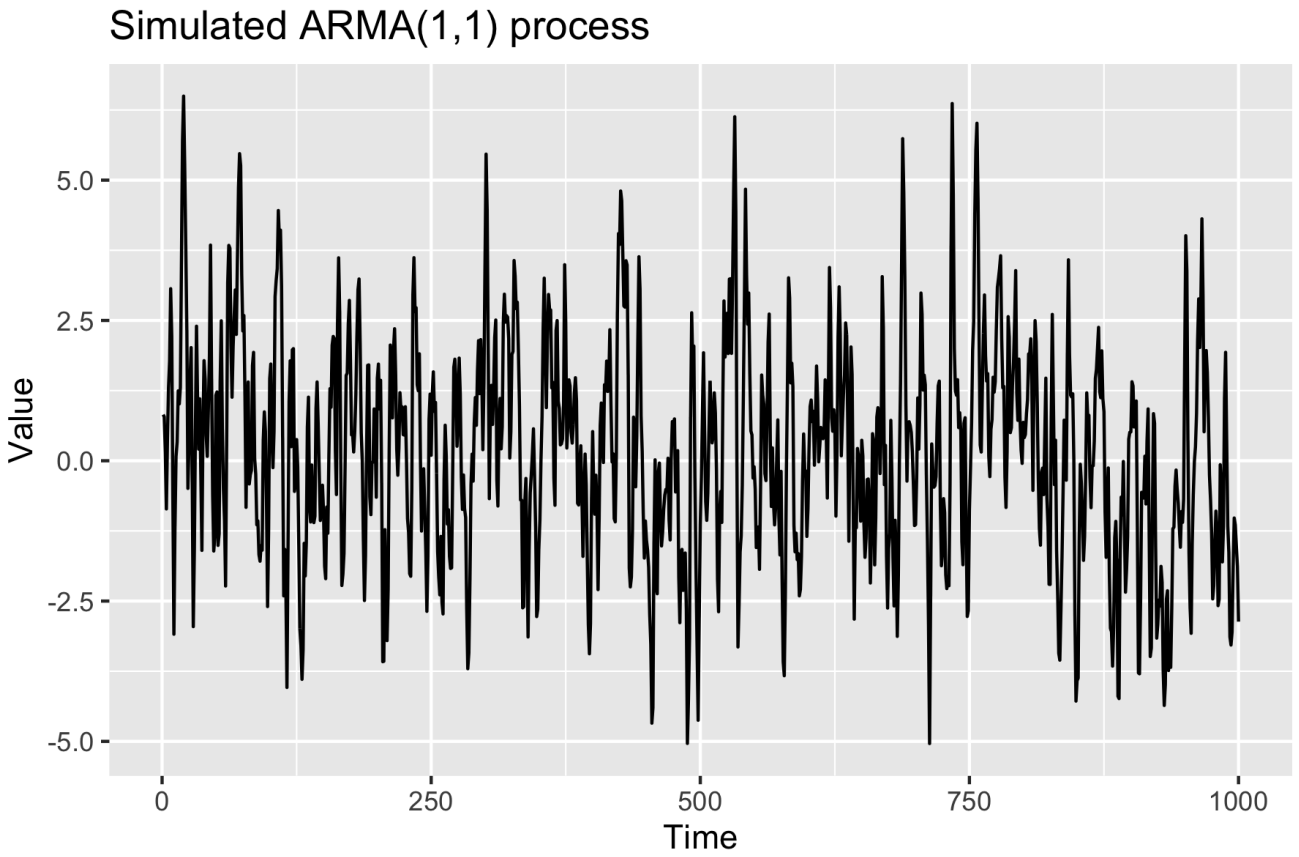


Figure: Simulated ARMA(1,1) process ($\phi_1 = 0.6,\ \theta_1 = 0.7$).

Autocovariance and Autocorrelation

Autocovariance

For a stationary time series $\{Y_t\}$ with mean μ,

$$\gamma_k = \mathrm{Cov}(Y_t, Y_{t-k}) = \mathrm{E}\big[(Y_t - \mu)(Y_{t-k} - \mu)\big].$$

$\gamma_0 = \mathrm{Var}(Y_t)$.

Autocorrelation

$$\rho_k = \frac{\gamma_k}{\gamma_0}, \qquad -1 \le \rho_k \le 1.$$

- ρ_k measures linear dependence between Y_t and Y_{t-k}.
- ACF (autocorrelation function) is ρ_k plotted vs lag k.
- For stationary series: $\gamma_{-k} = \gamma_k$, so $\rho_{-k} = \rho_k$.

ACF of AR(1)

AR(1) model:

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t, \quad |\phi_1| < 1, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

Variance:

$$\mathrm{Var}(Y_t) = \phi_1^2 \mathrm{Var}(Y_{t-1}) + \sigma^2.$$

Stationarity implies $\mathrm{Var}(Y_t) = \mathrm{Var}(Y_{t-1}) = \gamma_0$, so

$$\gamma_0 = \frac{\sigma^2}{1 - \phi_1^2}.$$

ACF:

- For lag 1: $\rho_1 = \phi_1$.
- By recursion, for general lag k:

$$\rho_k = \phi_1^k, \quad k = 0, 1, 2, \ldots$$

- ACF decays geometrically; sign and speed governed by ϕ_1.

Yule–Walker Equations for AR(2)

AR(2) model:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

Take covariances with Y_t, Y_{t-1}, Y_{t-2}:

$$\begin{aligned}
\gamma_0 &= \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2, \\
\gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_1, \\
\gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0.
\end{aligned}$$

- These are the **Yule–Walker equations** for AR(2).
- Higher-lag covariances satisfy recursion:

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}, \quad k \geq 3.$$

- AR(p) generalization: similar linear recursions for γ_k.

Yule–Walker Equations for AR(p)

AR(p) model:

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

Yule–Walker system:

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \cdots + \phi_p \gamma_p + \sigma^2;$$

for $k = 1, 2, \ldots,$

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

- In matrix form: $\Gamma \phi = \gamma$, where

$$\Gamma = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}.$$

- These equations connect AR parameters and autocovariances and are used in estimation.

ACF of MA(1) and MA(2)

MA(1):

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

$$\rho_0 = 1, \quad \rho_1 = \frac{\theta}{1+\theta^2}, \quad \rho_k = 0 \text{ for } k \geq 2.$$

MA(2):

$$Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}.$$

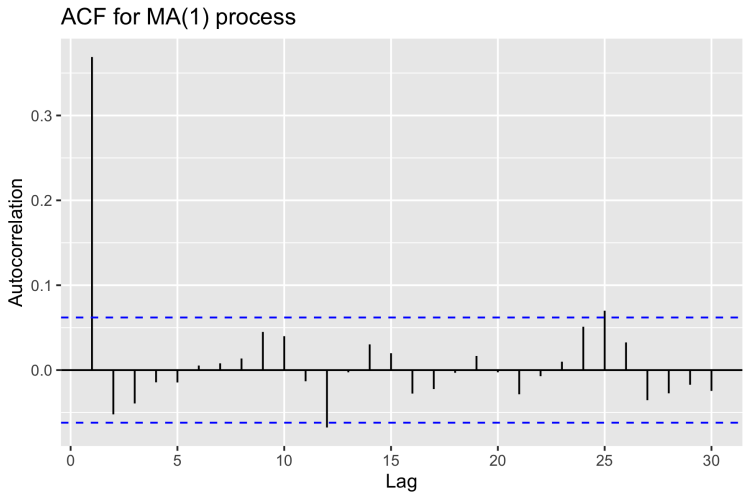
$$\rho_1 = \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2}, \quad \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2}, \quad \rho_k = 0 \text{ for } k \geq 3.$$

- ACF of MA(q): nonzero only up to lag q, then *cuts off* exactly.

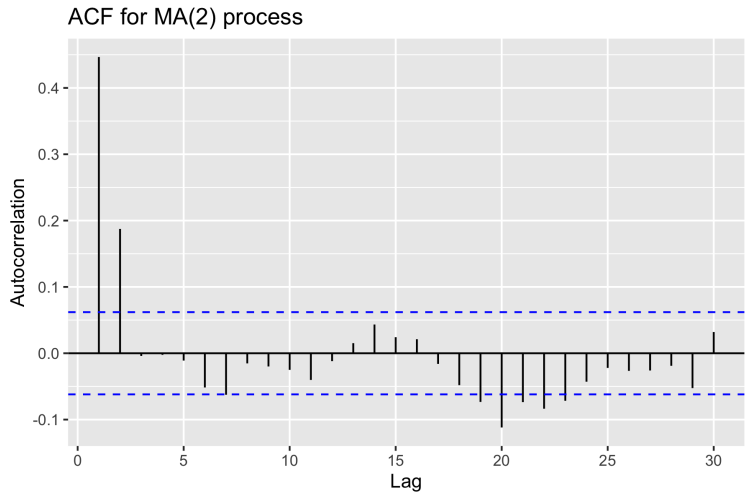
ACF Plots for MA(1) and MA(2)

Simulate MA(1) and MA(2) processes and compute empirical ACF.

- ACF of MA(1) has one nonzero spike at lag 1 (up to sampling noise).
- ACF of MA(2) shows nonzero spikes at lags 1 and 2, then zero (not statistically different from zero).



(a) MA(1) ACF



(b) MA(2) ACF

Figure: Empirical ACFs for simulated MA(1) and MA(2) processes.

Partial Autocorrelation and AR Models

- PACF at lag k = correlation between Y_t and Y_{t-k} after removing linear effects of intermediate lags.
- For AR(k) model:

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_k Y_{t-k} + \varepsilon_t,$$

the k-**th partial autocorrelation** equals ϕ_k.

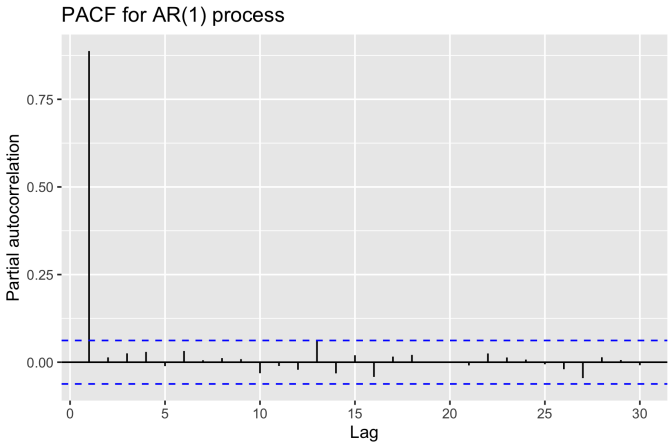
Examples:

- AR(1): $Y_t = \delta + \phi_1 Y_{t-1} + \varepsilon_t \Rightarrow$ PACF: $\phi_{11} = \phi_1$ at lag 1; zero at higher lags (in theory).
- AR(2): $Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \Rightarrow$ PACF: $\phi_{22} = \phi_2$ at lag 2; zero thereafter.

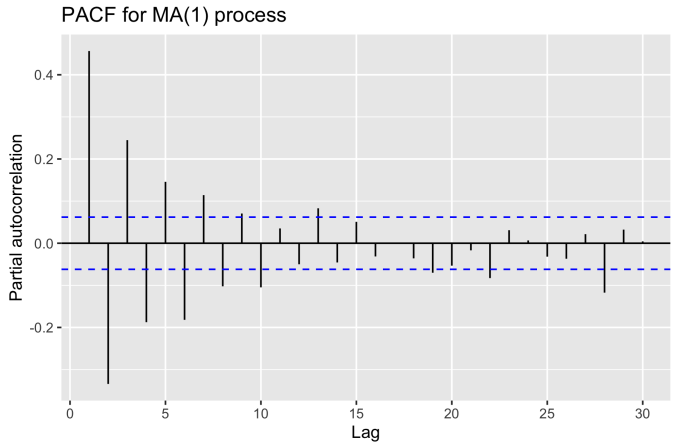
- PACF is particularly useful for identifying AR order.

PACF and ACF of AR(1) and MA(1)

- Simulate:
  - AR(1): $Y_t = 0.9Y_{t-1} + \varepsilon_t$,
  - MA(1): $Y_t = \varepsilon_t + 0.9\varepsilon_{t-1}$.
- Compute empirical PACF and ACF for each.



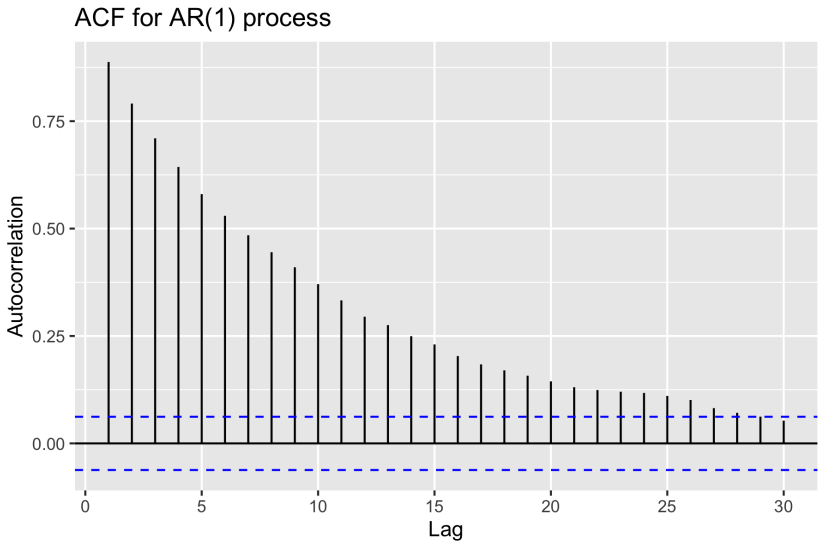
(a) PACF: AR(1)



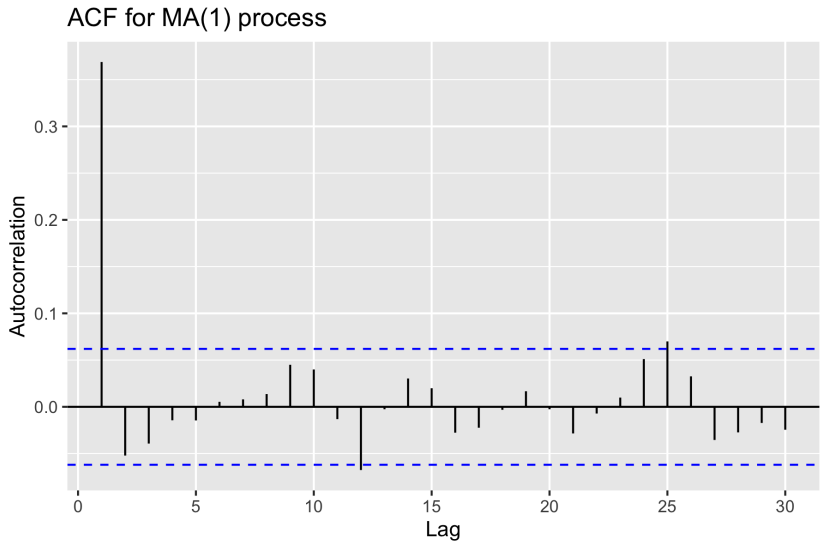
(b) PACF: MA(1)

Figure: PACFs for simulated AR(1) and MA(1) series.

ACF vs PACF: AR(1) and MA(1)



(a) ACF: AR(1)



(b) ACF: MA(1)

Figure: ACFs for simulated AR(1) and MA(1) series.

ACF vs PACF: AR(1) and MA(1)

- AR(1):
  - ACF tails off (geometric decay),
  - PACF shows a spike at lag 1 and then (theoretically) zero.
- MA(1):
  - ACF cuts off after lag 1,
  - PACF tails off.

ACF/PACF Patterns and Model Identification

- For **AR(p)**:
  - ACF: tails off,
  - PACF: cuts off after lag p.
- For **MA(q)**:
  - ACF: cuts off after lag q,
  - PACF: tails off.
- For **ARMA(p, q)**:
  - both ACF and PACF typically tail off.

Table: Heuristic patterns for ACF and PACF

Model	ACF	PACF
AR(p)	tails off	cuts off after lag p
MA(q)	cuts off	tails off
ARMA(p, q)	tails off	tails off
Unsuitable	cuts off	cuts off

Finite Samples, Significance Bands, and ARMA Orders

- **Why doesn't the AR(1) PACF literally hit zero at higher lags?**
  - Theoretical value is zero, but sample estimates are random variables.
  - For a continuous r.v., probability of being exactly zero is itself zero.
  - In practice, we look at whether sample PACF lies within $\pm$ (approx) $2/\sqrt{T}$ bands.
- **Why can't we read off ARMA(p, q) orders from ACF/PACF alone?**
  - ARMA = AR(∞) and MA(∞) views mixed together: both ACF and PACF tail off.
  - Many different (p, q) can produce similar ACF/PACF shapes.
  - We usually combine:
    - ACF/PACF patterns,
    - information criteria (AIC, BIC),
    - likelihood-based diagnostics.

Stationarity, Causality, and MA vs AR

- **MA(q)**:
  - finite linear combination of shocks,
  - automatically covariance-stationary (given finite variance).
- **AR(p)** and **ARMA(p, q)**:
  - need additional conditions on coefficients for stationarity.

Stationarity, Causality, and MA vs AR

- Example AR(1):

$$Y_t = \phi Y_{t-1} + \varepsilon_t.$$

- By repeated substitution:

$$Y_t = \phi^T Y_{t-T} + \sum_{j=0}^{T-1} \phi^j \varepsilon_{t-j}.$$

- Let $T \to \infty$ with $|\phi| < 1$ (so $\phi^T Y_{t-T} \to 0$):

$$Y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j},$$

a linear (MA(∞)) process.

AR(1) as a Linear Process: ACF

Example: AR(1) as MA(∞)

$$Y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}, \quad \mathrm{E}(\varepsilon_t) = 0.$$

AR(1) as a Linear Process: ACF

- Mean:

$$\mathrm{E}(Y_t) = \sum_{j=0}^{\infty} \phi^j \mathrm{E}(\varepsilon_{t-j}) = 0.$$

- Autocovariance (for $h \geq 0$):

$$\gamma(h) = \mathrm{Cov}(Y_{t+h}, Y_t) = \sigma^2 \sum_{k=0}^{\infty} \phi^{k+h}\phi^k = \frac{\sigma^2\phi^h}{1-\phi^2}.$$

- Hence $\gamma(-h) = \gamma(h)$ and

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h.$$

- This matches the earlier AR(1) ACF.

Causality for ARMA Processes

- Informally, a process is **causal** if the current value Y_t depends only on current and past shocks, not future ones.
- For ARMA(p, q), causality means we can write:

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t \in \mathbb{Z},$$

with $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

Causality for ARMA Processes

Definition (Causality)

An ARMA(p, q) process is *causal* if there exists a sequence $\{\psi_j\}_{j\geq 0}$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

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

- Causality rules out dependence on future shocks and ensures a Wold-type representation.

MA(1) Example and Non-Uniqueness

MA(1) process:

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{i.i.d. } WN(0, \sigma^2).$$

Autocovariance and ACF:

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2, & h = 0, \\ \theta\sigma^2, & h = 1, \\ 0, & h > 1, \end{cases} \quad \rho(h) = \begin{cases} 1, & h = 0, \\ \dfrac{\theta}{1+\theta^2}, & h = 1, \\ 0, & h > 1. \end{cases}$$

MA(1) Example and Non-Uniqueness

- Note: $\rho(1)$ is unchanged if θ is replaced by $1/\theta$.
- There are pairs (θ, σ^2) that give the same autocovariances, e.g. $(5, 1)$ and $(1/5, 25)$.
- Models

$$Y_t = \varepsilon_t + \tfrac{1}{5}\varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, 25),$$

and

$$Y_t = \tilde{\varepsilon}_t + 5\tilde{\varepsilon}_{t-1}, \quad \tilde{\varepsilon}_t \sim WN(0, 1),$$

are observationally equivalent from $\{Y_t\}$ alone.

Invertibility: Recovering Shocks from Observations

Definition (Invertibility)

An ARMA(p, q) process is **invertible** if there exists a sequence $\{\pi_j\}_{j\geq 0}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ such that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}, \quad t \in \mathbb{Z}.$$

Invertibility: Recovering Shocks from Observations

- Invertibility ensures that shocks ε_t can be expressed as a linear filter of past observations.
- It rules out non-unique parameterizations like the MA(1) example with both θ and $1/\theta$.
- Practically, we impose:
  - all roots of the MA polynomial lie outside the unit circle.
- Combined with causality (roots of AR polynomial outside unit circle), we obtain:
  - stationary, causal, invertible ARMA models with unique representations.
- From now on, unless stated otherwise, ARMA processes are assumed to be both causal and invertible.

Structural Models vs Time-Series Models

- **Structural (regression) models**:
  - Aim to capture *causal* mechanisms behind economic phenomena.
  - Based on economic theory or substantive knowledge.
  - Typical form: regression of an outcome on explanatory variables.
- **Time-series models**:
  - Focus on *temporal dependence* in a single (or several) series.
  - AR, MA, ARMA and related models used for dynamics and forecasting.
- These approaches are **not** contradictory:
  - structural models describe causal relations,
  - time-series models describe the stochastic evolution over time,
  - combining them yields richer empirical analysis.

Example: Structural Regression + MA(1) Covariate

Example

Suppose

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

with white-noise ε_t, and suppose

$$X_t = u_t + \alpha u_{t-1},$$

where u_t is white noise, independent of ε_t.

Example: Structural Regression + MA(1) Covariate

Then

$$Y_t = \beta_0 + \beta_1(u_t + \alpha u_{t-1}) + \varepsilon_t = \beta_0 + \beta_1 u_t + \alpha\beta_1 u_{t-1} + \varepsilon_t.$$

- $\mathrm{E}(Y_t) = \beta_0$.
- $\mathrm{Var}(Y_t) = \beta_1^2(1+\alpha^2)\sigma_u^2 + \sigma_\varepsilon^2$.
- $\mathrm{Cov}(Y_t, Y_{t-1}) = \beta_1^2\alpha\sigma_u^2$.
- $\mathrm{Cov}(Y_t, Y_{t-k}) = 0$ for $k \geq 2$.
- So Y_t is also an **MA(1)** process.

This shows that a structural causal relation can coexist with (and induce) a familiar time-series representation.

Table of Contents

Yule-Walker Equations for AR(p)

Recall AR(p):

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

Yule–Walker relations:

$$\gamma_0 = \phi_1 \gamma_1 + \cdots + \phi_p \gamma_p + \sigma^2,$$

$$\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_p \gamma_{k-p}, \quad k = 1, 2, \ldots$$

Yule-Walker Equations for AR(p)

Matrix form (for $k = 1, \ldots, p$):

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \ldots & \gamma(0) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{bmatrix}.$$

- Replace γ_k by sample autocovariances (or ACF) to obtain YW estimates $\widehat{\phi}_j$.

AR(2) Example via Yule-Walker

- Simulate an AR(2) process, then compute sample ACF at lags 0,1,2:

$$\widehat{\gamma}_0,\ \widehat{\gamma}_1,\ \widehat{\gamma}_2.$$

- Solve the 2x2 Yule–Walker system

$$\begin{bmatrix} \widehat{\gamma}_0 & \widehat{\gamma}_1 \\ \widehat{\gamma}_1 & \widehat{\gamma}_0 \end{bmatrix} \begin{bmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{bmatrix} = \begin{bmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_2 \end{bmatrix}.$$

- The resulting estimates $(\widehat{\phi}_1, \widehat{\phi}_2)$ are very close to the true AR(2) parameters used in the simulation.
- In practice, R's `ar.yw()` implements Yule–Walker estimation directly.

Likelihood and Log-Likelihood

Likelihood Function

Given a sample $x = (x_1, \ldots, x_n)$ from a model with parameter $\boldsymbol{\theta}$,

$$\mathcal{L}(\boldsymbol{\theta} \mid x) = f(x_1, \ldots, x_n \mid \boldsymbol{\theta}),$$

the joint density (or mass) of the sample viewed as a function of $\boldsymbol{\theta}$.

Likelihood and Log-Likelihood

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta} \mid x).$$

- Products of densities become sums of logs:

$$\log \prod_i f(x_i \mid \theta) = \sum_i \log f(x_i \mid \theta).$$

- Avoids numerical underflow (products of many small numbers).
- LLN/CLT apply directly to sums of log-densities, which is convenient for asymptotics.
- Maximizing $\ell(\theta)$ is equivalent to maximizing $\mathcal{L}(\theta)$.

MLE for Normal $N(\mu, \sigma^2)$

Assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$.

Likelihood:

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

Log-likelihood:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Differentiating and setting derivatives to zero yields:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2.$$

- $\widehat{\mu}$: unbiased and consistent for μ.
- $\widehat{\sigma}^2$: biased but consistent for σ^2.

MLE for ARMA(p, q): Setup

Consider an ARMA(p, q) process:

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

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

Parameter vector:

$$\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'.$$

Given observed data $y_1, \ldots, y_T$ and initial values for $(y_0, y_{-1}, \ldots)$ and $(\varepsilon_0, \varepsilon_{-1}, \ldots)$, we can recursively compute residuals:

$$\varepsilon_t = y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}.$$

Conditional Log-Likelihood for ARMA(p, q)

Treating initial values as given (or approximated), the conditional log-likelihood is:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log f(Y_T, \ldots, Y_1 \mid \text{initials}; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}.\end{aligned}$$

- Maximize $\ell(\boldsymbol{\theta})$ numerically over $\boldsymbol{\theta}$:
  - gradient-based methods (Newton, BFGS, etc.).
- Initial conditions (for Y_{-j}, ε_{-j}) are not observed:
  - can be set to unconditional mean / zero,
  - effect vanishes as $T \to \infty$.
- In practice, functions like `arima()` or `Arima()` in R perform MLE (or QMLE) for ARMA models.

Table of Contents

Information Criteria: AIC and BIC

For a fitted ARMA(p, q) model with MLE $\widehat{\boldsymbol{\theta}}$, log-likelihood $\mathcal{L}(\widehat{\boldsymbol{\theta}})$, and k parameters:

$$\text{AIC} = -2\,\mathcal{L}(\widehat{\boldsymbol{\theta}}) + 2k,$$

$$\text{BIC} = -2\,\mathcal{L}(\widehat{\boldsymbol{\theta}}) + k \log T,$$

where T is the sample size.

- Smaller AIC/BIC $\Rightarrow$ better trade-off between fit and complexity.
- AIC uses penalty $2k$; BIC uses stronger penalty $k \log T$.
- Both aim to prevent overfitting (too many parameters).

Principle of Parsimony (Occam's Razor)

Principle of Parsimony

Among competing models that all fit the data reasonably well, prefer the simpler one (fewer parameters, fewer assumptions).

- In statistics:
  - complexity increases risk of overfitting,
  - information criteria penalize complexity, reflecting Occam's Razor.
- Analogy with **axioms vs theorems**:
  - axioms are minimal assumptions,
  - theorems build on axioms + extra conditions,
  - axioms are more fundamental and broadly applicable.
- For ARMA models:
  - prefer low-order ARMA that fits well,
  - don't add lags unless they significantly improve fit.

Example: ARMA Order Selection in Practice

- Simulate an ARMA(1,1) process.
- Use an automatic procedure (e.g. `auto.arima()` in R) that:
  - searches over candidate p, q,
  - evaluates AIC and/or BIC for each,
  - selects the model with smallest criterion value.
- In the example, both AIC and BIC correctly selected an ARMA(1,1) specification, with estimated coefficients close to the true values.

Table of Contents

Unconditional Mean of ARMA(p, q)

ARMA(p, q):

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

Take expectations:

$$\mu = \mathrm{E}(Y_t) = c + (\phi_1 + \cdots + \phi_p)\,\mu.$$

Assuming stationarity,

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

- In practice, plug in estimates $\widehat{\phi}_j$ and $\widehat{c}$ to obtain $\widehat{\mu}$.
- This mean is used when centering the series in forecasting formulas.

One-Step-Ahead Forecasting

Write the model in deviations from mean:

$$(1 - \phi_1 L - \cdots - \phi_p L^p)(Y_t - \mu) = (1 + \theta_1 L + \cdots + \theta_q L^q)\varepsilon_t.$$

One-step-ahead forecast at time t:

$$\begin{aligned}\widehat{Y}_{t+1|t} - \mu = \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p+1} - \mu) \\ + \theta_1 \widehat{\varepsilon}_t + \cdots + \theta_q \widehat{\varepsilon}_{t-q+1},\end{aligned}$$

with forecast errors

$$\widehat{\varepsilon}_t = Y_t - \widehat{Y}_{t|t-1}.$$

- AR part uses past *observed* Y_t.
- MA part uses past *estimated residuals* $\widehat{\varepsilon}_t$.

s-Step-Ahead Forecasts and Long-Horizon Behavior

For s-step-ahead forecast:

$$\widehat{Y}_{t+s|t} - \mu = \phi_1(\widehat{Y}_{t+s-1|t} - \mu) + \cdots + \phi_p(\widehat{Y}_{t+s-p|t} - \mu) + \theta_s\widehat{\varepsilon}_t + \cdots + \theta_q\widehat{\varepsilon}_{t+s-q},$$

for $s = 1, \ldots, q$; for $s > q$, the MA terms drop out:

$$\widehat{Y}_{t+s|t} - \mu = \phi_1(\widehat{Y}_{t+s-1|t} - \mu) + \cdots + \phi_p(\widehat{Y}_{t+s-p|t} - \mu).$$

- For horizons $s \leq q$: both AR and MA parts play a role.
- For $s > q$: forecasts follow a pure AR recursion.
- Thus, MA influence fades at longer horizons; AR structure governs long-run dynamics.

Example: ARMA(1,1) Forecasting

- Simulate ARMA(1,1): e.g. $Y_t = 0.5 Y_{t-1} + \varepsilon_t + 0.3\varepsilon_{t-1}$.
- Fit an ARMA(1,1) model by MLE.
- Produce forecasts $h = 20$ steps ahead.
- Plot shows:
  - historical data,
  - point forecasts,
  - forecast intervals widening with horizon.

Example: ARMA(1,1) Forecasting

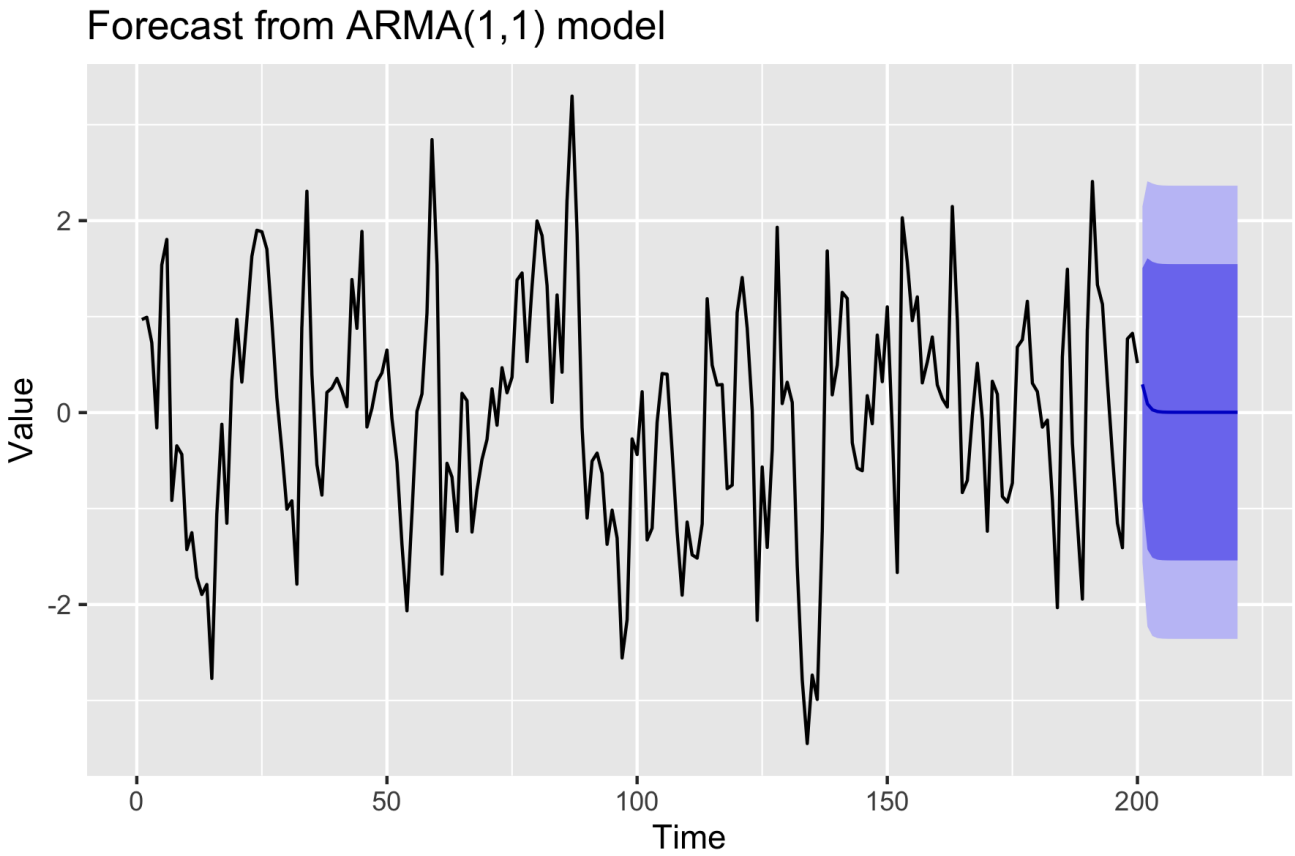


Figure: Forecast from fitted ARMA(1,1) model.

Table of Contents

Unbiasedness vs Consistency

- **Unbiasedness**:
$$\mathrm{E}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}.$$
- **Consistency**:
$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta} \quad \text{as } T \to \infty.$$
- They are distinct:
  - estimator can be unbiased but inconsistent,
  - or biased but consistent,
  - or both, or neither.

Unbiasedness vs Consistency

Example (Normal $N(\mu, \sigma^2)$):

- Sample mean $\bar{X}$:
  - unbiased and consistent for μ.
- MLE for variance:

$$\widehat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is biased but consistent, whereas

$$\widehat{\sigma}^2_{\text{unb}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased.

Score and Its Expectation

Let X have density $f(x \mid \boldsymbol{\theta})$.

Score Function

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X \mid \boldsymbol{\theta}).$$

- Under mild conditions,
$$\mathrm{E}\left[S(\boldsymbol{\theta})\right] = 0.$$
- Proof sketch:
$$\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X \mid \boldsymbol{\theta})\right] = \int \frac{\partial}{\partial \boldsymbol{\theta}} f(x \mid \boldsymbol{\theta})\, dx = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(x \mid \boldsymbol{\theta})\, dx = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = 0.$$
- Score is the gradient of the log-likelihood; its zeros are candidates for MLEs.

Fisher Information and Hessian

Fisher Information

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left[S(\boldsymbol{\theta})S(\boldsymbol{\theta})'\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X \mid \boldsymbol{\theta})\right)^2\right] \quad \text{(scalar case)}.$$

If $\log f(x \mid \boldsymbol{\theta})$ is twice differentiable and regularity conditions hold:

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log f(X \mid \boldsymbol{\theta})\right].$$

Hessian of the Log-Likelihood

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log f(X \mid \boldsymbol{\theta}).$$

- Thus $-\mathcal{I}(\theta)$ is the expected Hessian of the log-likelihood.
- Curvature of $\ell(\theta)$ near the maximum is tied to precision of the MLE.

KL Divergence and Consistency

- Define the empirical loss:

$$R_T(\widehat{\theta}, \theta) = \frac{1}{T} \sum_{i=1}^{T} \log \frac{f(X_i \mid \theta)}{f(X_i \mid \widehat{\theta})}.$$

- Population counterpart:

$$R(\widehat{\theta}, \theta) = \mathrm{E}_\theta \log \frac{f(X \mid \theta)}{f(X \mid \widehat{\theta})} = \mathrm{KL}\big(f(\cdot \mid \theta) \,\|\, f(\cdot \mid \widehat{\theta})\big),$$

  the Kullback–Leibler divergence.

- LLN implies $R_T(\tilde{\theta}, \theta) \xrightarrow{p} R(\tilde{\theta}, \theta)$ for any $\tilde{\theta}$.

KL Divergence and Consistency

Identifiability and Uniform LLN

- **Identifiability**: $\theta_1 \neq \theta_2 \Rightarrow f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$.
- Stronger condition: for every $\epsilon > 0$,

$$\inf_{\tilde{\theta}:|\tilde{\theta}-\theta|\geq\epsilon} \mathrm{KL}\big(f(\cdot \mid \theta) \,\|\, f(\cdot \mid \tilde{\theta})\big) > 0.$$

- Uniform LLN:

$$\sup_{\tilde{\theta}} \left| R_T(\tilde{\theta}, \theta) - R(\tilde{\theta}, \theta) \right| \xrightarrow{p} 0.$$

Consistency Theorem (Sketch)

Theorem (Consistency of MLE)

Under strong identifiability and uniform LLN, the MLE $\widehat{\boldsymbol{\theta}}$ is consistent:

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}.$$

Sketch of proof idea:

- By definition, MLE maximizes average log-likelihood:

$$\frac{1}{T}\sum_{i=1}^{T} \log f(X_i \mid \widehat{\theta}) \geq \frac{1}{T}\sum_{i=1}^{T} \log f(X_i \mid \theta).$$

- Equivalently, $R_T(\widehat{\theta}, \theta) \leq 0$.
- Uniform LLN implies $R_T(\tilde{\theta}, \theta)$ converges uniformly to $R(\tilde{\theta}, \theta)$.
- Strong identifiability implies $R(\tilde{\theta}, \theta) > 0$ whenever $|\tilde{\theta} - \theta| \geq \epsilon$.
- Therefore, with high probability, $\widehat{\theta}$ must lie within an ϵ-neighborhood of θ for large T.

Regularity Conditions for Asymptotic Normality

Key regularity conditions (informally):

- Parameter dimension d fixed (does not grow with T): $\theta \in \mathbb{R}^d$.
- $f(x \mid \theta)$ is sufficiently smooth in θ (e.g. three times differentiable).
- Differentiation under the integral sign is permitted:
  - support of X does not depend on θ,
  - relevant expectations of derivatives are finite.
- Identifiability holds.
- True parameter θ is an interior point of the parameter space.

Asymptotic Normality and Cramér–Rao Bound

Theorem (Asymptotic Normality of MLE)

Under regularity conditions,

$$\sqrt{T}(\widehat{\theta} - \theta) \Longrightarrow N\left(0, \mathcal{I}^{-1}(\theta)\right),$$

where $\mathcal{I}(\theta)$ is the Fisher information (per observation).

Asymptotic Normality and Cramér–Rao Bound

Cramér–Rao Lower Bound

For any unbiased estimator $\tilde{\theta}$,

$$\mathrm{Var}(\tilde{\theta}) \geq \frac{1}{T\,\mathcal{I}(\theta)}.$$

- The MLE attains this bound asymptotically, so it is **asymptotically efficient**.
- Standard errors are obtained from $\mathcal{I}^{-1}(\widehat{\theta})$ or from $-H(\widehat{\theta})^{-1}$.

Fisher Information Equality

Under regularity conditions,

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right].$$

- Left-hand side: **variance of the score**
  - measures how variable the score is,
  - high variance $\Rightarrow$ sample is informative about θ.
- Right-hand side: **negative expected Hessian**
  - measures curvature of log-likelihood around the true parameter,
  - sharper curvature $\Rightarrow$ more precise estimates.

In practice, approximate

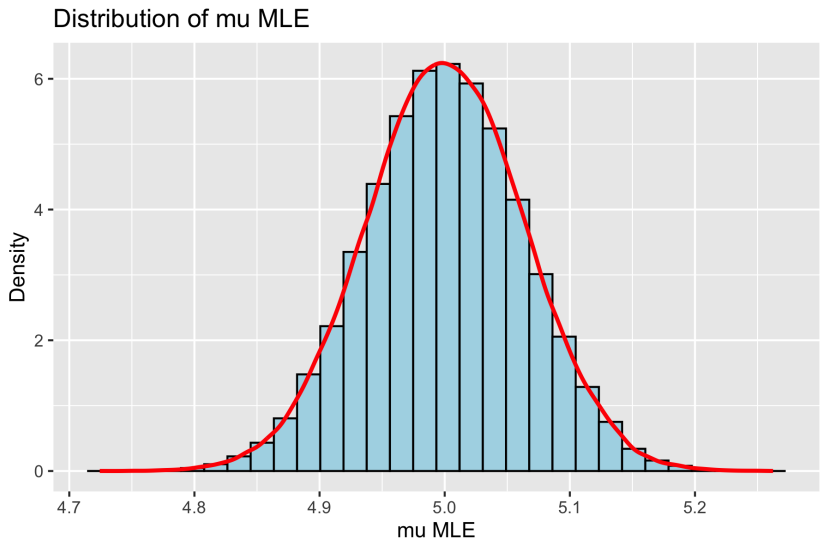
$$\widehat{\mathcal{I}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2}{\partial\theta^2}\log f(X_t \mid \theta),$$

and use $\widehat{\mathcal{I}}^{-1}(\widehat{\theta})$ as an estimate of the asymptotic covariance of $\widehat{\theta}$.

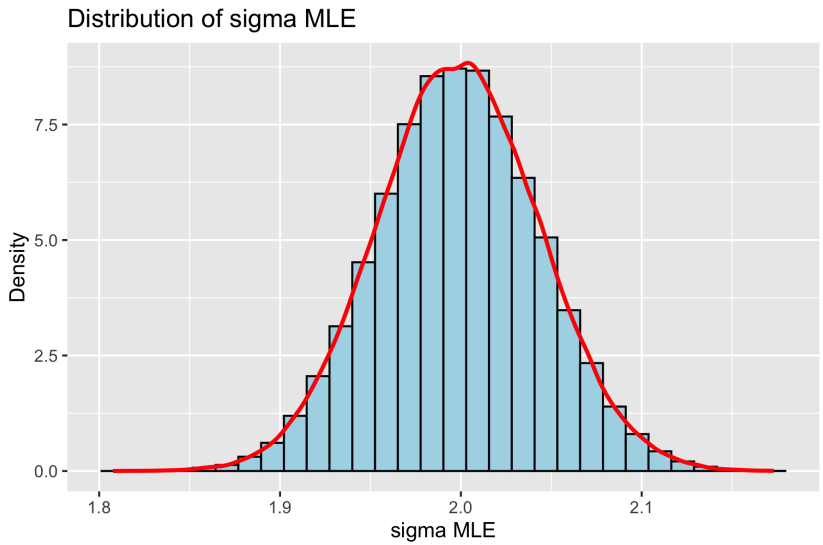
Simulation: MLE for Normal Mean and Variance

- True parameters: $\mu = 5$, $\sigma = 2$.
- For each simulation:
  - draw sample of size $n = 1000$ from $N(\mu, \sigma^2)$,
  - compute MLEs: $\widehat{\mu} = \bar{X}$, $\widehat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.
- Repeat a large number of times (e.g. 100,000).
- Empirical results:
  - average of $\widehat{\mu}$ close to 5,
  - average of $\widehat{\sigma}^2$ close to 4,
  - histograms of $\widehat{\mu}$ and $\widehat{\sigma}$ approximately normal, with spread consistent with $1/(T\mathcal{I}(\theta))$.

Empirical Distributions of MLEs



(a) Distribution of $\widehat{\mu}$



(b) Distribution of $\widehat{\sigma}$

Figure: Simulation-based distributions of the MLEs $\widehat{\mu}$ and $\widehat{\sigma}$.

- $\widehat{\mu}$ is tightly concentrated around the true mean.
- $\widehat{\sigma}$ is concentrated near the true standard deviation.
- Shapes are close to normal, illustrating asymptotic normality.

Table of Contents

Chapter Summary

- Developed a toolkit for univariate time-series analysis:
  - strict vs weak stationarity,
  - autocovariance γ_k, autocorrelation ρ_k,
  - ergodicity and mixing for LLN and CLT with dependent data.
- Wold decomposition:
  - any stationary process = deterministic part + linear filter of white noise,
  - motivates ARMA models and linear processes.
- Lag polynomials:
  - convenient representation of ARMA,
  - link AR and MA forms via invertible lag operators.

Summary (cont'd)

- ACF/PACF patterns:
  - AR(p): ACF tails off, PACF cuts off after lag p.
  - MA(q): ACF cuts off after lag q, PACF tails off.
  - ARMA(p, q): both typically tail off.
- Causality and invertibility:
  - characteristic roots outside the unit circle,
  - ensure stationary solutions and unique shock representations.
- Estimation:
  - AR by Yule–Walker,
  - ARMA by (quasi-)maximum likelihood.
- Model selection:
  - AIC/BIC trade off fit vs complexity,
  - parsimony principle guides order choice.

Summary (cont'd)

- Forecasting with ARMA:
  - one-step and multi-step forecasts via AR recursion + MA residuals,
  - MA influence fades at long horizons; AR part dominates long-run behavior.
- MLE toolkit:
  - score, Fisher information, Hessian,
  - consistency (via KL divergence and LLN),
  - asymptotic normality and efficiency (Cramér–Rao bound).
- R implementation (conceptually):
  - simulation of ARMA processes,
  - estimation (YW, MLE),
  - diagnostics (ACF, PACF),
  - forecasting and plotting,
  - simulation to verify MLE properties.

These tools prepare us to move on to more advanced models: time-series regressions, VAR/SVAR, and volatility models.

Table of Contents

References I


Bickel, Peter J and Peter Bühlmann (1997). "Closure of Linear Processes". In: *Journal of Theoretical Probability* 10, pp. 445–479.

Chapter 3 — Multivariate Linear Time Series

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Motivation: Why Multivariate Time Series?

- Many economic and financial variables move together over time:
  - GDP, inflation, interest rates,
  - equity indices, FX rates, yields, etc.
- **Univariate** models analyse each series in isolation:
  - useful but ignore cross-variable dynamics.
- **Multivariate** models (e.g. VAR) jointly model several series:
  - capture dynamic interdependencies,
  - better suited for joint forecasting, policy analysis, and causal interpretation.

Table of Contents

VAR(p) Model: Definition

Vector autoregression (VAR) generalizes AR models to multiple series.

VAR(p) model:

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \cdots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where:

- $\mathbf{y}_t$: $n \times 1$ vector of variables,
- $\mathbf{c}$: $n \times 1$ constant vector,
- Φ_j: $n \times n$ coefficient matrices, $j = 1, \ldots, p$,
- $\boldsymbol{\varepsilon}_t$: $n \times 1$ white-noise innovations:

$$\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0, \quad \mathrm{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_\tau') = \begin{cases} \boldsymbol{\Omega}, & t = \tau, \\ \mathbf{0}, & t \neq \tau, \end{cases}$$

$\boldsymbol{\Omega}$ symmetric positive definite.

Lag-Operator Representation of VAR(p)

Using lag operator L, $L\mathbf{y}_t = \mathbf{y}_{t-1}$:

$$\left(I_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p\right)\mathbf{y}_t = \mathbf{c} + \varepsilon_t,$$

or

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

with

- $\Phi(L)$ an $n \times n$ matrix polynomial in L,
- element in row i, column j:

$$\delta_{ij} - \phi_{ij}^{(1)} L - \phi_{ij}^{(2)} L^2 - \cdots - \phi_{ij}^{(p)} L^p,$$

  where $\delta_{ij} = 1$ if $i = j$, 0 otherwise.

- This compact form is convenient for deriving stability and impulse-response properties.

Companion Form: VAR(p) as VAR(1)

Following Hamilton (1994, pp. 258–259), rewrite VAR(p) as VAR(1):

$$\xi_t = \mathbf{F}\,\xi_{t-1} + \mathbf{v}_t,$$

where

$$\xi_t = \begin{bmatrix} \mathbf{y}_t - \mu \\ \mathbf{y}_{t-1} - \mu \\ \vdots \\ \mathbf{y}_{t-p+1} - \mu \end{bmatrix}, \quad \mathbf{v}_t = \begin{bmatrix} \varepsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix},$$

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

- $\mathbf{F}$ is an $(np) \times (np)$ companion matrix.
- Stationarity/stability is determined by eigenvalues of $\mathbf{F}$.

Stability Condition via Eigenvalues

From

$$\boldsymbol{\xi}_{t+s} = \mathbf{v}_{t+s} + \mathbf{F}\mathbf{v}_{t+s-1} + \cdots + \mathbf{F}^{s-1}\mathbf{v}_{t+1} + \mathbf{F}^{s}\boldsymbol{\xi}_{t},$$

we see that:

- For covariance stationarity:

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

  i.e. all eigenvalues of $\mathbf{F}$ must lie **inside** the unit circle.

- Then shocks from $\boldsymbol{\varepsilon}_t$ dissipate over time, and the system returns to its long-run equilibrium.
- If eigenvalues lie outside/ on the unit circle:
  - shocks may have permanent or explosive effects,
  - long-run forecasts can be unstable or unrealistic.

Checking eigenvalues of $\mathbf{F}$ (or equivalent characteristic roots) is crucial for VAR diagnostics.

Characteristic Polynomial and Roots

Proposition 1

The eigenvalues λ of $\mathbf{F}$ *satisfy*

$$\left|\lambda I_n - \Phi_1 \lambda^{p-1} - \Phi_2 \lambda^{p-2} - \cdots - \Phi_p\right| = 0.$$

A VAR(p) is covariance stationary when all such λ have modulus less than 1. Equivalently, if

$$\left|I_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right| = 0,$$

then all roots z must lie **outside** *the unit circle.*

- This mirrors the univariate AR case: AR roots outside the unit circle.
- Software for VAR estimation typically reports the moduli of these roots as a stability diagnostic.

Gaussian VAR(p) Specification

Consider a Gaussian VAR(p):

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \cdots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

with

$$\boldsymbol{\varepsilon}_t \sim \text{i.i.d. } N(0, \boldsymbol{\Omega}).$$

- Number of series: n.
- Sample: $T + p$ observations $(\mathbf{y}_{-p+1}, \ldots, \mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T)$.
- Treat the first p observations as given (conditioning).
- Estimate parameters: $\mathbf{c}, \Phi_1, \ldots, \Phi_p$, and $\boldsymbol{\Omega}$ via maximum likelihood.

Conditional Likelihood and Log-Likelihood

Condition on initial values $\mathbf{y}_0, \ldots, \mathbf{y}_{-p+1}$.

Conditional likelihood:

$$f_{\mathbf{Y}_T,\ldots,\mathbf{Y}_1|\mathbf{Y}_0,\ldots,\mathbf{Y}_{-p+1}}\left(\mathbf{y}_T, \ldots, \mathbf{y}_1 \mid \mathbf{y}_0, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}\right),$$

where $\boldsymbol{\theta} = (\mathbf{c}, \Phi_1, \ldots, \Phi_p, \boldsymbol{\Omega})$.

Define

$$\mathbf{x}_t = \begin{bmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix}, \quad \Pi' = \begin{bmatrix} \mathbf{c} & \Phi_1 & \Phi_2 & \cdots & \Phi_p \end{bmatrix}.$$

Then

$$\mathbf{y}_t \mid \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{-p+1} \sim N\left(\Pi'\mathbf{x}_t, \boldsymbol{\Omega}\right).$$

Log-Likelihood and MLE Formulas

Log-likelihood:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \log f\left(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \ldots; \boldsymbol{\theta}\right) \\ &= -\frac{Tn}{2}\log(2\pi) + \frac{T}{2}\log|\boldsymbol{\Omega}^{-1}| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \Pi'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \Pi'\mathbf{x}_t).\end{aligned}$$

Maximization yields:

$$\widehat{\Pi}' = \left(\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_t'\right)\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1},$$

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

with $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\Pi}'\mathbf{x}_t$.

Log-Likelihood and MLE Formulas

- For Gaussian VARs, MLE coincides with multivariate OLS:
  - regress each element of $\mathbf{y}_t$ on $\mathbf{x}_t$,
  - stack coefficients into $\widehat{\Pi}$.

Asymptotic Properties of VAR MLE (I)

Hamilton (1994, Props. 11.1, 11.2) shows:

Proposition 2

Let

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \cdots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

with $\boldsymbol{\varepsilon}_t$ i.i.d. mean 0, variance $\boldsymbol{\Omega}$, and finite fourth moments. Assume the VAR is stable (roots of $|I_n - \Phi_1 z - \cdots - \Phi_p z^p| = 0$ lie outside unit circle).
Let $k = np + 1$ and define

$$\mathbf{x}_t' = \begin{bmatrix} 1 & \mathbf{y}_{t-1}' & \cdots & \mathbf{y}_{t-p}' \end{bmatrix}.$$

Let $\widehat{\boldsymbol{\pi}}_T = \mathrm{vec}(\widehat{\Pi}_T)$ be the stacked OLS coefficients and $\boldsymbol{\pi}$ the population counterpart. Define $\widehat{\boldsymbol{\Omega}}_T$ as the sample covariance of residuals. Then:

1. $\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \xrightarrow{p} \mathbf{Q} = \mathrm{E}(\mathbf{x}_t \mathbf{x}_t')$.
2. $\widehat{\boldsymbol{\pi}}_T \xrightarrow{p} \boldsymbol{\pi}$.
3. $\widehat{\boldsymbol{\Omega}}_T \xrightarrow{p} \boldsymbol{\Omega}$.

Asymptotic Properties of VAR MLE (II)

Proposition 3

Under the same conditions,

$$\sqrt{T}(\widehat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) \Longrightarrow N\big(\mathbf{0},\ \boldsymbol{\Omega} \otimes \mathbf{Q}^{-1}\big),$$

where $\otimes$ *denotes the Kronecker product.*
Moreover, if $\boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Omega})$,

$$\sqrt{T}\begin{bmatrix} \mathrm{vec}(\widehat{\Pi}_T - \Pi) \\ \mathrm{vech}(\widehat{\boldsymbol{\Omega}}_T - \boldsymbol{\Omega}) \end{bmatrix} \Longrightarrow N\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega} \otimes \mathbf{Q}^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right),$$

where the covariance between $\widehat{\sigma}_{ij}$ *and* $\widehat{\sigma}_{lm}$ *is* $\sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl}$.

- Thus, standard OLS formulas for coefficient standard errors apply.
- Even under non-Gaussian innovations, MLE/OLS are consistent; Gaussianity mainly underpins exact finite-sample distributions.

Why Lag Order Matters

- VAR lag order p controls how much history enters each equation.
- Too many lags:
  - many parameters,
  - risk of overfitting and large estimation variance.
- Too few lags:
  - omitted dynamics,
  - autocorrelated residuals and biased inference.
- Goal: choose p to balance fit and parsimony.

Information Criteria: AIC, BIC, HQIC

For a VAR model with log-likelihood $\mathcal{L}$, k parameters, and sample size T:

$$\text{AIC} = 2k - 2\mathcal{L},$$

$$\text{BIC} = \log(T)k - 2\mathcal{L},$$

$$\text{HQIC} = -2\mathcal{L} + 2k\,\frac{\log\log T}{T}.$$

- AIC: weaker penalty on complexity; may overfit for large T.
- BIC: stronger penalty $\propto \log T$; tends to favor more parsimonious models.
- HQIC: penalty between AIC and BIC.
- In R, `VARselect()` in the `vars` package reports AIC, BIC, HQIC for candidate lag orders; choose p that minimizes a chosen criterion.

Likelihood Ratio (LR) Test for Lag Order

Suppose we test:

$$H_0 : \text{VAR with } p_0 \text{ lags} \quad \text{vs.} \quad H_1 : \text{VAR with } p_1 > p_0 \text{ lags.}$$

Log-likelihood under H_0:

$$\mathcal{L}_0^* = -\frac{Tn}{2}\log(2\pi) + \frac{T}{2}\log|\widehat{\mathbf{\Omega}}_0^{-1}| - \frac{Tn}{2}.$$

Log-likelihood under H_1:

$$\mathcal{L}_1^* = -\frac{Tn}{2}\log(2\pi) + \frac{T}{2}\log|\widehat{\mathbf{\Omega}}_1^{-1}| - \frac{Tn}{2}.$$

LR statistic:

$$2(\mathcal{L}_1^* - \mathcal{L}_0^*) = T\left[\log|\widehat{\mathbf{\Omega}}_1^{-1}| - \log|\widehat{\mathbf{\Omega}}_0^{-1}|\right].$$

Under H_0, asymptotically χ^2 with $n(p_1 - p_0)$ degrees of freedom.

LR, Score, and Wald Tests

Three key tests in MLE framework:

Likelihood Ratio (LR) Test

$$\mathrm{LR} = -2\log\left(\frac{\mathcal{L}(\hat{\theta}_0)}{\mathcal{L}(\hat{\theta})}\right) = 2\big[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\hat{\theta}_0)\big],$$

compares restricted ($\hat{\theta}_0$) vs unrestricted ($\hat{\theta}$) models.

Score (LM) Test

$$S = g(\hat{\theta}_0)'\,\mathcal{I}(\hat{\theta}_0)^{-1}\,g(\hat{\theta}_0),$$

where g is the gradient of log-likelihood at $\hat{\theta}_0$, $\mathcal{I}$ Fisher information.

- Only needs estimation under the null.

Wald Test

$$W = (\hat{\theta} - \theta_0)'\big[\mathrm{Var}(\hat{\theta})\big]^{-1}(\hat{\theta} - \theta_0),$$

tests whether $\hat{\theta}$ is close to hypothesized θ_0.

All three are asymptotically χ^2 with d.f. equal to the number of tested restrictions.

LR, Score, Wald: Graphical Interpretation

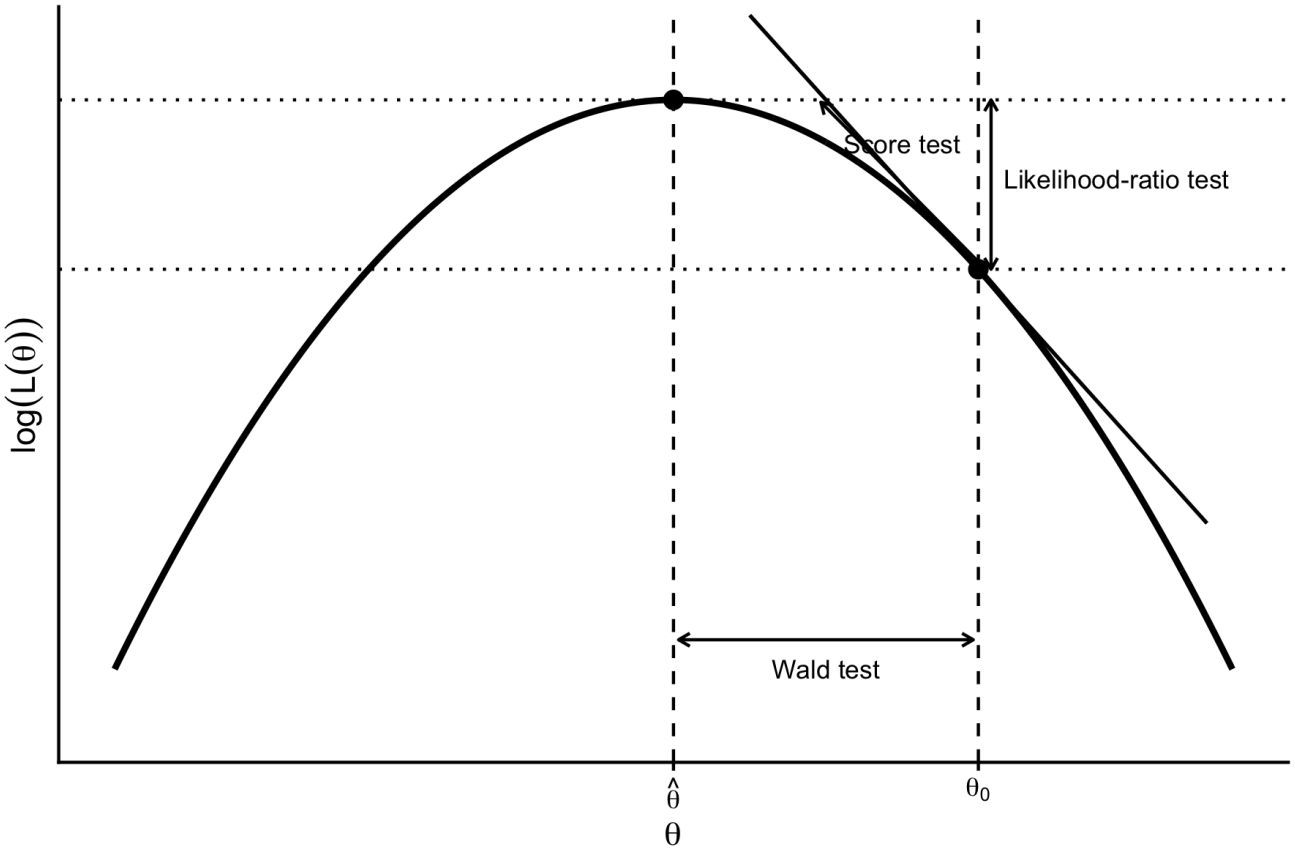


Figure: Comparative illustration of likelihood-ratio, score, and Wald tests (adapted from Fox, 1997).

LR, Score, Wald: Graphical Interpretation

- All three tests use different aspects of the same likelihood surface.
- In regular problems, they are asymptotically equivalent.
- In nonstandard cases, LR is often more robust, but Wald is simpler (only one model fit).

Granger Causality: Definition

- Granger causality asks: do past values of y help predict x, beyond what past x already does?
- Let $\mathrm{MSE}[\hat{x}_{t+s}]$ be MSE of forecasting x_{t+s}.

Definition 1

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

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

If

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

then y Granger-causes x.

Granger Causality in a Bivariate VAR(p)

Consider a bivariate VAR(p) in $(x_t, y_t)'$. If all coefficient matrices Φ_j are lower triangular, then y does not Granger-cause x:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} \phi_{11}^{(j)} & 0 \\ \phi_{21}^{(j)} & \phi_{22}^{(j)} \end{bmatrix} \begin{bmatrix} x_{t-j} \\ y_{t-j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

- Constraints:
$$\phi_{12}^{(1)} = \phi_{12}^{(2)} = \cdots = \phi_{12}^{(p)} = 0.$$
- Unconstrained model allows non-zero $\phi_{12}^{(j)}$:
$$\Phi_j = \begin{bmatrix} \phi_{11}^{(j)} & \phi_{12}^{(j)} \\ \phi_{21}^{(j)} & \phi_{22}^{(j)} \end{bmatrix}.$$
- Granger test compares constrained vs unconstrained VAR via F-test or LR test.

Does Granger Causality Imply True Causality?

- Granger causality is about *predictability*, not structural causation.
- Example (Hamilton, 1994, pp. 306–307):
  - Stock prices may Granger-cause dividends (prices predict future dividends),
  - but economically, expected dividends cause prices (forward-looking markets).

Does Granger Causality Imply True Causality?

- Another example (Lütkepohl, 2005, p. 48):

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

  - Here x_t does not Granger-cause z_t (upper-right element is zero).
  - But after a nonsingular linear transformation B, one can write a different representation where x_t appears to influence z_t.

- Conclusion:
  - absence or presence of Granger causality does not fully determine true causal relationships,
  - structural reasoning and theory remain essential.

From VAR to VMA and IRFs

For a stable VAR, we can write an infinite VMA representation:

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \Psi_1 \boldsymbol{\varepsilon}_{t-1} + \Psi_2 \boldsymbol{\varepsilon}_{t-2} + \cdots .$$

- Ψ_s are $n \times n$ moving-average coefficient matrices.
- The derivative:
$$\frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} = \Psi_s.$$
- The element (i, j) of Ψ_s is:
$$\frac{\partial y_{i,t+s}}{\partial \varepsilon_{jt}},$$
measuring effect of a one-unit shock in ε_{jt} on variable $y_{i,t+s}$.
- Plotting $\Psi_s(i, j)$ over s yields the **impulse response function** from shock j to variable i.

Orthogonalized IRFs via Shock Decomposition

- In practice, VAR residuals $\boldsymbol{\varepsilon}_t$ may be contemporaneously correlated: $\mathrm{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \boldsymbol{\Omega}$.
- Then a shock to ε_{jt} typically also affects other components at time t.
- To isolate shocks, decompose $\boldsymbol{\Omega}$ using an LDL or Cholesky factorization:

$$\boldsymbol{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}',$$

where $\mathbf{A}$ is lower triangular with ones on the diagonal, $\mathbf{D}$ diagonal with positive entries.

- Define orthogonalized shocks:

$$\mathbf{u}_t = \mathbf{A}^{-1}\boldsymbol{\varepsilon}_t,$$

so $\mathrm{E}(\mathbf{u}_t \mathbf{u}_t') = \mathbf{D}$ (uncorrelated components).

Orthogonalized IRFs (OIRFs)

Orthogonalized IRFs measure the response to a unit shock in the *orthogonal* innovations.

- Let $\mathbf{a}_j$ denote the j-th column of $\mathbf{A}$.
- A one-unit shock in u_{jt} corresponds to a particular combination of $\boldsymbol{\varepsilon}_t$, summarized by $\mathbf{a}_j$.

Orthogonalized IRF

The effect of a one-unit orthogonalized shock u_{jt} on $\mathbf{y}_{t+s}$ is:

$$\frac{\partial \widehat{\mathrm{E}}(\mathbf{y}_{t+s} \mid \text{info})}{\partial u_{jt}} = \Psi_s \mathbf{a}_j.$$

Orthogonalized IRFs (OIRFs)

- In estimation:
  - obtain $\widehat{\Phi}_1, \ldots, \widehat{\Phi}_p$,
  - compute $\widehat{\Psi}_s$ (e.g. recursively or via software),
  - estimate $\widehat{\Omega}$ from residuals and factorize: $\widehat{\Omega} = \widehat{A}\widehat{D}\widehat{A}'$,
  - orthogonalized IRFs: $\widehat{\Psi}_s \widehat{\mathbf{a}}_j$.

Case Study: Data and Context

- We model daily equity ETF returns using a VAR:
  - SPDR S&P 500 ETF (**SPY**) — U.S. large caps,
  - Invesco QQQ Trust (**QQQ**) — Nasdaq/growth,
  - iShares Russell 2000 ETF (**IWM**) — U.S. small caps,
  - iShares MSCI EAFE ETF (**EFA**) — developed ex–U.S. equities.
- Daily adjusted closing prices from Yahoo Finance for calendar year 2024:
  - in R, can be downloaded via `quantmod::getSymbols()`.
- We:
  1. compute log returns for each ETF,
  2. fit a VAR model with lag order selected by AIC,
  3. compute orthogonal impulse response functions using `vars` package,
  4. obtain confidence bands via bootstrap.

Case Study: Modelling Steps

1. **Data acquisition**
   - Download daily adjusted prices $P_t^{\text{SPY}}, P_t^{\text{QQQ}}, P_t^{\text{IWM}}, P_t^{\text{EFA}}$.
2. **Compute log returns**

$$r_t^{(i)} = \log P_t^{(i)} - \log P_{t-1}^{(i)}, \quad i \in \{\text{SPY, QQQ, IWM, EFA}\}.$$

3. **Select VAR lag order**
   - apply `VARselect` (or equivalent) with lag.max $= 10$,
   - choose lag order minimizing AIC.
4. **Estimate VAR**
   - fit VAR with the chosen lag order and constant term.
5. **Impulse response analysis**
   - compute orthogonal IRFs (Cholesky-based) with bootstrap confidence intervals,
   - examine how shocks to one ETF propagate to others.

Case Study: Example IRFs (Conceptual)

- For each ordered pair (impulse, response), we can compute a 10-day IRF:
  - e.g. response of QQQ returns to a unit shock in SPY returns,
  - response of IWM to QQQ, etc.
- Orthogonalization (e.g. Cholesky) ensures shocks are uncorrelated at impact:
  - ordering of variables matters (SPY, QQQ, IWM, EFA).
- Bootstrap confidence bands provide uncertainty intervals around IRFs.

Case Study: Example IRFs (Conceptual)

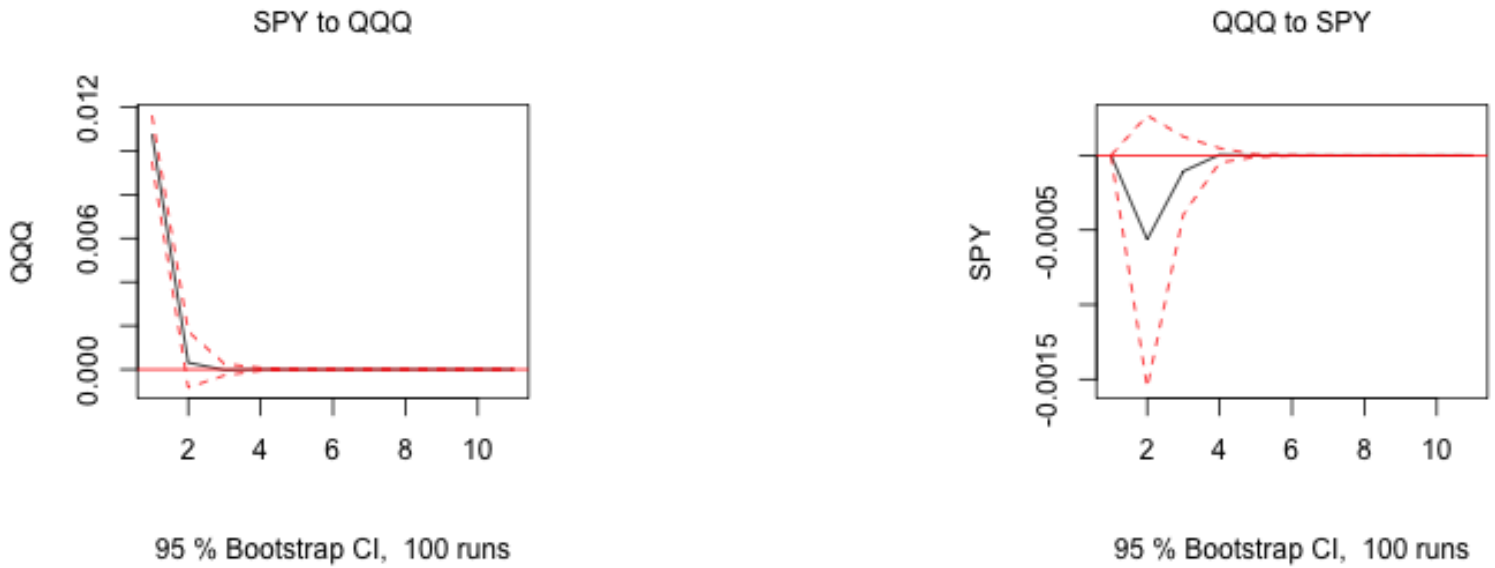


Figure: Illustrative orthogonal IRFs: SPY to QQQ (left) and QQQ to SPY (right).

Case Study: Interpretation and Takeaways

- IRFs quantify dynamic connectivity among equity segments:
  - how U.S. large-cap shocks (SPY) propagate to growth (QQQ), small caps (IWM), and international (EFA),
  - how shocks in QQQ or IWM feed back into SPY.
- VAR framework allows:
  - joint forecasting of ETF returns,
  - assessment of shock transmission,
  - Granger causality testing between segments.
- This case study demonstrates:
  - practical application of VAR modelling in finance,
  - integration of estimation, model selection, and IRF analysis,
  - how multivariate time series tools give richer insight than univariate models.

Nonstationary VAR and Cointegration: Overview

- Many macro and financial series are **nonstationary** and often exhibit unit roots.
- Vector autoregressions with unit roots can be written in **error-correction** form and may exhibit **cointegration**.
- Key ideas in this section:
  - Random walks and dimensionality (recurrence vs transience),
  - Cointegration and the Granger Representation Theorem,
  - Spurious regression and the need for unit-root / cointegration tests,
  - Engle–Granger residual-based tests and Johansen's system tests,
  - Vector error correction models (VECMs),
  - Empirical case study: income–consumption cointegration using FRED data.

Random Walk in $\mathbb{R}^n$

- Define a random walk in $\mathbb{R}^n$:

$$\mathbf{y}_t = \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

with $\boldsymbol{\varepsilon}_t$ i.i.d. mean 0 and covariance $\boldsymbol{\Omega}_\varepsilon$.

- Iterating:

$$\mathbf{y}_t = \sum_{s=1}^{t} \boldsymbol{\varepsilon}_s + \mathbf{y}_0.$$

- Conditional mean:

$$\mathrm{E}(\mathbf{y}_t \mid \mathbf{y}_0) = \mathbf{y}_0,$$

- Conditional covariance:

$$\mathrm{E}\big[(\mathbf{y}_t - \mathbf{y}_0)(\mathbf{y}_t - \mathbf{y}_0)' \mid \mathbf{y}_0\big] = t\,\boldsymbol{\Omega}_\varepsilon.$$

- Variability grows linearly in t; the process is **nonstationary**.

Recurrence vs Transience and the "Drunken Man"

- **Recurrence**: the walk returns to its starting point (or any point) infinitely often with probability 1.
- **Transience**: probability of ever returning to a given point is less than 1.
- Classic result:
  - $n = 1$ (line): random walk is **recurrent**.
  - $n = 2$ (plane): still recurrent, but returns less frequently.
  - $n \geq 3$: random walk is **transient**.

Recurrence vs Transience and the "Drunken Man"

Metaphor:

- 1D "drunken man" on a line must keep crossing any point infinitely often.
- 2D "drunken man" in a plane still returns, but with lower frequency.
- 3D "drunken bird" in space can drift off and may never come back to the starting point.
- This dimensionality effect is analogous to cointegration: in higher-dimensional systems it becomes harder to find stationary linear combinations (i.e., stationary "returns" to equilibrium).

Cointegration: Economic Motivation

- Cointegration describes **long-run equilibrium** relations among nonstationary series.
- Different from simple correlation:
  - correlation ignores time and stability,
  - cointegration focuses on common stochastic trends and equilibrium error.
- In economics:
  - long-run links between GDP, interest rates, investment, etc.,
  - consumption–income relationships, PPP, money demand, etc.
- In finance:
  - used to identify pairs or portfolios that move together in the long run,
  - key to pairs trading strategies (e.g. Gatev, Goetzmann, and Rouwenhorst, 2006).

VAR(1) Representation and Common Trends

- Consider VAR(1):
$$\mathbf{y}_t = \Phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$
- Allows eigenvalues of Φ on or outside the unit circle, e.g. $\Phi = I_n$ gives a multivariate random walk.
- Not restrictive: any VAR(p) can be written as VAR(1) in companion form.
- If Φ is symmetric, we can write $\Phi = Q\Lambda Q'$:
  - Λ diagonal with eigenvalues,
  - Q orthonormal.
- Let $\mathbf{x}_t = Q'\mathbf{y}_t$, $\boldsymbol{\varepsilon}_t^* = Q'\boldsymbol{\varepsilon}_t$; then
$$\mathbf{x}_t = \Lambda \mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t^*.$$
- Component x_{jt} is stationary iff $|\lambda_j| < 1$.

Error-Correction Representation

- Generally, define:

$$\Delta \mathbf{y}_t = \Pi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

  where $\Pi = -I_n + \Phi$.

- If Φ has eigenvalues equal to 1, Π may have reduced rank and can be written as:

$$\Pi = \alpha \beta',$$

  where α, β are $n \times r$ matrices with full column rank r.

- This is the basis for cointegration and the Granger Representation Theorem.

Granger Representation Theorem

Theorem 2 (Granger Representation Theorem)

Let $\Phi(z) = \Phi z - I$, *and assume:*

1. *All roots of* $|\Phi(z)| = 0$ *are either outside the unit circle or equal to 1.*
2. $\Phi(1) = \Pi = -I_n + \Phi$ *has reduced rank* $r < n$, *i.e.* $\Pi = \alpha\beta'$ *with* α, β $n \times r$ *full rank.*
3. $\alpha'_\perp \beta_\perp$ *has full rank* r, *where* $\alpha_\perp, \beta_\perp$ *are orthogonal complements:* $\alpha'_\perp \alpha = 0$, $\beta'_\perp \beta = 0$.

Then

$$\mathbf{y}_t = C \sum_{s=1}^{t} \varepsilon_s + (I_n - C) \sum_{i=0}^{\infty} (I_n + \alpha\beta')^i \varepsilon_{t-i} + C\mathbf{y}_0,$$

where $C = \beta_\perp (\alpha'_\perp \beta_\perp)^{-1} \alpha'_\perp$ *has rank* r.

Granger Representation Theorem

- Decomposition into:
  - nonstationary "trend" part,
  - stationary part,
  - effect of initial condition.
- Condition (i): roots at or outside unit circle, including unit roots.
- Condition (ii): reduced rank $\Pi = \alpha\beta'$ captures fewer long-run relations than variables:
  - typical cointegration structure: nonstationary series tied together by r equilibrium relations.

Interpretation of the Granger Representation

- Condition (iii): independence/regularity condition ensuring non-redundant cointegrating relations.
- The decomposition shows:
  - a finite number of common stochastic trends (driven by cumulated shocks),
  - plus transitory stationary components.
- Cointegration implies existence of stationary linear combinations $\beta' \mathbf{y}_t$.

Definition of Cointegration

Definition 3

Let $\mathbf{y}_t \in \mathbb{R}^n$ be a vector of $I(1)$ processes. If there exists $\alpha \in \mathbb{R}^n$ such that $\alpha' \mathbf{y}_t$ is $I(0)$, then $\mathbf{y}_t$ is said to be **cointegrated** with cointegrating vector α.

Simple bivariate example ($n = 2$):

- $\mathbf{y}_t = (y_{1t}, y_{2t})'$, cointegrating vector $\alpha = (1, -\beta)$.
- Then

$$u_{1t} = y_{1t} - \beta y_{2t}$$

  is stationary, so we can write:

$$y_{1t} = \beta y_{2t} + u_{1t}.$$

- Estimate β by OLS:

$$\widehat{\beta} = \left(\sum_{t=1}^{T} y_{2t}^2 \right)^{-1} \sum_{t=1}^{T} y_{2t} y_{1t}.$$

Superconsistency of Cointegration Estimators

- Suppose y_{2t} is a random walk:

$$y_{2t} = y_{2,t-1} + \varepsilon_t,$$

and u_t is stationary.

- Under regularity:

$$\frac{1}{T^2}\sum_{t=1}^{T} y_{2t}^2 \Longrightarrow \int_0^1 B_\varepsilon(s)^2\, ds,$$

$$\frac{1}{T}\sum_{t=1}^{T} y_{2t}u_t \Longrightarrow \int_0^1 B_\varepsilon(s)\, dB_u(s),$$

where B_ε, B_u are independent Brownian motions.

- Hence

$$T(\widehat{\beta} - \beta) \Longrightarrow \frac{\int_0^1 B_\varepsilon(s)\, dB_u(s)}{\int_0^1 B_\varepsilon(s)^2\, ds}.$$

- Convergence rate is T (superconsistency) rather than $\sqrt{T}$.

Superconsistency of Cointegration Estimators

Superconsistency means convergence faster than the usual $\sqrt{T}$ rate. This yields very accurate estimates in large samples, but:

- the limiting distribution is nonstandard,
- standard t-/F-tests based on normality do not apply directly,
- inference often requires special critical values or resampling methods.

Random Walk vs Brownian Motion

- **Random walk**: discrete-time process with i.i.d. steps on a space (e.g. $\mathbb{Z}$ or $\mathbb{R}$).
- **Brownian motion** (Wiener process): continuous-time limit of a random walk:
  - continuous in time and space,
  - random fluctuations without jumps.
- Random walk and Brownian motion are intimately connected:
  - Donsker's invariance principle (functional CLT): scaled partial sums of i.i.d. variables converge to Brownian motion.

Random Walk vs Brownian Motion

Historical note:

- Robert Brown (1827) observed erratic motion of pollen grains $\Rightarrow$ "Brownian motion".
- Einstein (1905) gave a physical theory linking Brownian motion to molecular collisions, supporting atomic theory.
- Norbert Wiener developed the rigorous mathematical model (Wiener process), foundational in modern probability and finance.

Spurious Regression: Random Walks

- If two unrelated random walks are regressed on each other, OLS often suggests a significant relationship.
- Example:

$$\text{RW1}_t = \sum_{s=1}^{t} \varepsilon_s, \quad \text{RW2}_t = \sum_{s=1}^{t} \eta_s,$$

  with ε_s, η_s independent.
- Regression:

$$\text{RW1}_t = \delta + \gamma \text{RW2}_t + u_t$$

  frequently yields a significant $\hat{\gamma}$, even though RW1 and RW2 are independent.
- This is **spurious regression** (Granger and Newbold, 1974).

Spurious Regression: Random Walks

Figure: Spurious regression between two independent random walks

Spurious Regression with Deterministic Trends

- Consider
$$y_t = \alpha + \beta t + \varepsilon_t, \quad x_t = \alpha^* + \beta^* t + \varepsilon_t^*,$$
with independent noise terms.
- Regressing y_t on x_t:
$$y_t = \delta + \gamma x_t + u_t,$$
yields $\gamma = \beta/\beta^*$.
- The regression indicates a "significant" relation driven solely by shared trends, not by a true causal link.
- Remedy:
  - check stationarity and cointegration,
  - regress on differenced series if appropriate,
  - or explicitly model trend and error-correction terms.

Engle–Granger Residual-Based Test

- Engle and Granger (1987) proposed a residual-based test for cointegration.
- For $\mathbf{y}_t \in \mathbb{R}^n$, regress

$$y_{1t} = \delta + \pi' \mathbf{y}_{2t} + \varepsilon_t, \tag{1}$$

  where y_{1t} is the first element, $\mathbf{y}_{2t}$ collects the remaining $n-1$.
- Under $r = 1$ and y_{1t} in the cointegrating relation, this regression represents the cointegration equation.
- Under H_0 of no cointegration ($r = 0$), regression is spurious; usual t-/F-statistics diverge.
- Nonetheless, residuals $\hat{\varepsilon}_t$ are informative:
  - if $\hat{\varepsilon}_t$ is $I(0)$, variables are cointegrated,
  - if $\hat{\varepsilon}_t$ is $I(1)$, no cointegration.

ADF Test on Cointegration Residuals

- Let x_t be OLS residuals from the cointegration regression:

$$x_t = y_{1t} - \hat{\delta} - \hat{\pi}' \mathbf{y}_{2t}.$$

- Apply an ADF test:

$$\Delta x_t = (\theta - 1)x_{t-1} + c_1 \Delta x_{t-1} + \cdots + c_p \Delta x_{t-p} + \text{error}, \tag{2}$$

- Differences from usual univariate ADF:
  - regression (**??**) includes constant, so x_t has mean zero $\Rightarrow$ no constant in ADF,
  - time trend typically excluded, as trend is already captured in regressors,
  - number of lags p must increase with T but slower than $T^{1/3}$.
- Critical values depend on the number of regressors and whether they have drift:
  - Phillips (1990) and later tables (Fuller, 1996; Phillips and Ouliaris, 1990).

Residual-Based ADF Critical Values

- Case 1: no drift in $\Delta \mathbf{y}_{2t}$ and Δy_{1t}:
  - residual-based ADF critical values given by Phillips (1990, Table IIb),
  - depend on the number of regressors (excluding constant).

Table: Critical values for ADF on residuals (no drift in regressors).

# regressors	1%	2.5%	5%	10%
1	-3.96	-3.64	-3.37	-3.07
2	-4.31	-4.02	-3.77	-3.45
3	-4.73	-4.37	-4.11	-3.83
4	-5.07	-4.71	-4.45	-4.16
5	-5.28	-4.98	-4.71	-4.43

Source: Phillips (1990), Table IIb (also Hamilton, 1994).

Cases with Drift and Trend

- Case 2: some regressors have drift, $\mathrm{E}(\Delta \mathbf{y}_{2t}) \neq 0$:
  - linear trends from different regressors can be combined into a single time trend,
  - residual-based ADF critical values correspond to regressions with time trend and one fewer $I(1)$ regressor.
- Case 3: $\mathrm{E}(\Delta \mathbf{y}_{2t}) = 0$ but $\mathrm{E}(\Delta y_{1t}) \neq 0$:
  - include time in the cointegration regression to remove linear trend from residuals,
  - ADF critical values as in Phillips (1990, Table IIc) / Fuller (1996) with adjusted number of regressors.
- Key point:
  - residual-based ADF tests on cointegration residuals use **nonstandard** critical values,
  - these depend on deterministic structure (constant, trend) and dimension n.

Phillips' Triangular System

- Suppose $\mathbf{y}_t = (\mathbf{y}_{1t}', \mathbf{y}_{2t}')'$ where both components are $I(1)$.
- Model:

$$\mathbf{y}_{1t} = B\mathbf{y}_{2t} + \mathbf{u}_{1t}, \quad \mathbf{y}_{2t} = \mathbf{y}_{2,t-1} + \mathbf{u}_{2t},$$

  with $\mathbf{u}_t = (\mathbf{u}_{1t}', \mathbf{u}_{2t}')'$ stationary.
- Then

$$\mathbf{u}_{1t} = \mathbf{y}_{1t} - B\mathbf{y}_{2t}$$

  is stationary, giving n_1 cointegrating relations:

$$\mathbf{y}_{1t} - B\mathbf{y}_{2t}.$$

- B encodes the long-run (cointegrating) relationship.

Vector Error Correction Representation

Take first differences:

$$\Delta \mathbf{y}_t = -\begin{bmatrix} I_{n_1} & B \\ 0 & I_{n_2} \end{bmatrix} \mathbf{y}_{t-1} + \mathbf{v}_t = -E\Pi \mathbf{y}_{t-1} + \mathbf{v}_t,$$

where

$$E = \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix}, \quad \Pi = (I, -B), \quad \mathbf{v}_t = \begin{bmatrix} I_{n_1} & B \\ 0 & I_{n_2} \end{bmatrix} \mathbf{u}_t.$$

- Error-correction term: $\Pi \mathbf{y}_{t-1}$ measures deviation from long-run equilibrium.
- Negative sign means the correction at time t moves the system back towards equilibrium.
- This is a special case of the general VECM structure:

$$\Delta \mathbf{y}_t = \Gamma_1 \Delta \mathbf{y}_{t-1} + \cdots + \Gamma_{p-1} \Delta \mathbf{y}_{t-p+1} + \Pi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$

Conditional Mean and Inference on B

Assuming $\mathbf{v}_t \sim i.i.d.\ N(0, \mathbf{\Omega})$, the conditional mean of $\mathbf{y}_{1t}$ given $\mathbf{y}_{2t}$ is:

$$\mathbf{y}_{1t} = B\mathbf{y}_{2,t-1} + C\Delta\mathbf{y}_{2t} + \mathbf{v}_{1\cdot 2,t},$$

where

$$C = \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-1}, \quad \mathbf{v}_{1\cdot 2,t} = \mathbf{v}_{1t} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-1}\mathbf{v}_{2t}.$$

- Under Gaussianity, MLE of B equals the OLS estimator from regression of $\mathbf{y}_{1t}$ on $(\mathbf{y}_{2,t-1}, \Delta\mathbf{y}_{2t})$.
- Using a functional CLT, one can derive the limiting distribution:

$$T(\widehat{B} - B) \Longrightarrow N(0, V)$$

  with V a mixed normal variance matrix (Phillips, 1991).
- Independence between certain Brownian components allows standard χ^2-based Wald tests on B in this triangular system setting.

Johansen's System Approach

- Johansen (1988, 1991, 1995) developed a full-system method to:
  - test for the number of cointegrating relations,
  - estimate cointegrating vectors and adjustment coefficients.
- Start from VAR(p) in levels:

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \cdots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

- Rewrite as VECM:

$$\Delta \mathbf{y}_t = \Gamma_1 \Delta \mathbf{y}_{t-1} + \cdots + \Gamma_{p-1} \Delta \mathbf{y}_{t-p+1} + \Pi \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where

$$\Gamma_i = \Phi_1 + \cdots + \Phi_i - I_n, \quad \Pi = \Phi_1 + \cdots + \Phi_p - I_n.$$

Rank of Π and Cointegration

- Rank of Π characterizes cointegration:
  - $\text{rank}(\Pi) = 0$: no cointegration, all series are $I(1)$ but no stationary linear combinations.
  - $\text{rank}(\Pi) = n$: all series stationary; cointegration not needed.
  - $0 < \text{rank}(\Pi) = r < n$:
    - r cointegrating relations, $\Pi = \alpha\beta'$,
    - $\beta'\mathbf{y}_t$ are $I(0)$,
    - α contains adjustment (loading) coefficients.
- Johansen's tests:
  - **Trace test**: $H_0 : \text{rank}(\Pi) \leq r$ vs $H_1 : \text{rank}(\Pi) > r$.
  - **Maximum eigenvalue test**: $H_0 : \text{rank}(\Pi) = r$ vs $H_1 : \text{rank}(\Pi) = r + 1$.

Johansen Tests: Implementation Sketch

- Let:
  - R_{0t}: residuals from regressing $\Delta \mathbf{y}_t$ on $\Delta \mathbf{y}_{t-1}, \ldots, \Delta \mathbf{y}_{t-p+1}$,
  - R_{kt}: residuals from regressing $\mathbf{y}_{t-p}$ on same lags.
- Define sample covariance matrices:

$$S_{00} = \frac{1}{T}\sum R_{0t}R_{0t}', \quad S_{0k} = \frac{1}{T}\sum R_{0t}R_{kt}', \quad S_{k0} = S_{0k}', \quad S_{kk} = \frac{1}{T}\sum R_{kt}R_{kt}'.$$

- Canonical correlations $\widehat{\lambda}_i$ are obtained from

$$\left|\lambda S_{kk} - S_{0k}S_{00}^{-1}S_{k0}\right| = 0,$$

  ordered $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$.
- Trace statistic:

$$\mathrm{LR}_{\mathrm{trace}}(r) = -T\sum_{i=r+1}^{n}\log(1-\widehat{\lambda}_i).$$

- Maximum eigenvalue statistic:

$$\mathrm{LR}_{\mathrm{max}}(r) = -T\log(1-\widehat{\lambda}_{r+1}).$$

Johansen Test Distributions

- Under H_0, Johansen's LR statistics converge to functionals of Brownian motion:

$$\mathcal{J} = \mathrm{tr}\left(\int_0^1 dBB' \left(\int_0^1 BB' du \right)^{-1} \int_0^1 BdB' \right),$$

  with B an $(n-r)$-dimensional standard Brownian motion.

- Distributions are **non-Gaussian** but tabulated.
- Johansen (1991) extended results to include:
  - intercepts and trends,
  - seasonal terms, etc.
- In practice:
  - use software (e.g. R's `urca::ca.jo`) which reports test statistics and critical values.

FRED Case Study: Data Description

- Data from FRED (St. Louis Fed):
  - **PCEC96**: real personal consumption expenditures,
  - **DSPIC96**: real disposable personal income.
- Monthly data from 1960 onward.
- Process:
  1. Download series using FRED's API (e.g. via `fredr` package in R).
  2. Merge and take logs:

$$\ell C_t = \log(\text{PCEC96}_t), \quad \ell Y_t^d = \log(\text{DSPIC96}_t).$$

  3. Remove missing values, convert to a monthly `ts` object.
  4. Plot log levels to visualise co-movement.

FRED Case Study: Unit-Root Diagnostics

- ADF tests on log levels with linear trend:
  - $\tau_3 \approx -3.74$ for $\log C_t$,
  - $\tau_3 \approx -3.43$ for $\log Y_t^d$,
  - borderline vs 5% critical values $\Rightarrow$ nonstationarity is plausible.
- ADF tests on first differences:
  - $\tau_2 \approx -13.25$ for $\Delta \log C_t$,
  - $\tau_2 \approx -3.96$ for $\Delta \log Y_t^d$,
  - strongly reject unit root in differences.
- Conclusion:
  - both log real consumption and log real disposable income behave as $I(1)$ processes.

FRED Case Study: Engle–Granger and Phillips–Ouliaris Tests

Engle–Granger (two-step):

- Step 1: regress $\log C_t$ on $\log Y_t^d$:

$$\log C_t = \delta + \beta \log Y_t^d + u_t.$$

- OLS yields:

$$\widehat{\beta} \approx 0.895, \quad R^2 \approx 0.92,$$

- Step 2: ADF on residual $\hat{u}_t$ (no constant, no trend): $\tau_1 \approx -3.19$, more negative than the 5% critical value.
- $\Rightarrow$ residuals appear stationary $\Rightarrow$ evidence of cointegration.

Phillips–Ouliaris:

- P_z statistic ≈ 130.6, far above critical values,
- strongly rejects the null of no cointegration.

FRED Case Study: Johansen Tests and VECM

- Choose VAR lag order in levels (e.g. via BIC or AIC), then apply Johansen tests with trend in the cointegration space.
- Trace test:
  - rejects $r = 0$ at 5%,
  - does not reject $r \leq 1$,
  - suggests **rank** $r = 1$.
- Maximum eigenvalue test yields similar conclusion.
- Estimate VECM, extract cointegrating vector β (normalized on ℓC):

$$\beta' = (\ 1,\ \beta_2,\ \beta_3\ ),$$

  e.g. $\beta_2 \approx 0.635$, $\beta_3 \approx -0.0037$ (trend term).
- Adjustment (loading) coefficients:
  - income equation: significant negative coefficient on ECM (≈ -0.06),
  - consumption equation: insignificant loading,
  - suggests income adjusts to restore equilibrium while consumption is weakly exogenous.

Log Consumption and Log Income: Visual Evidence

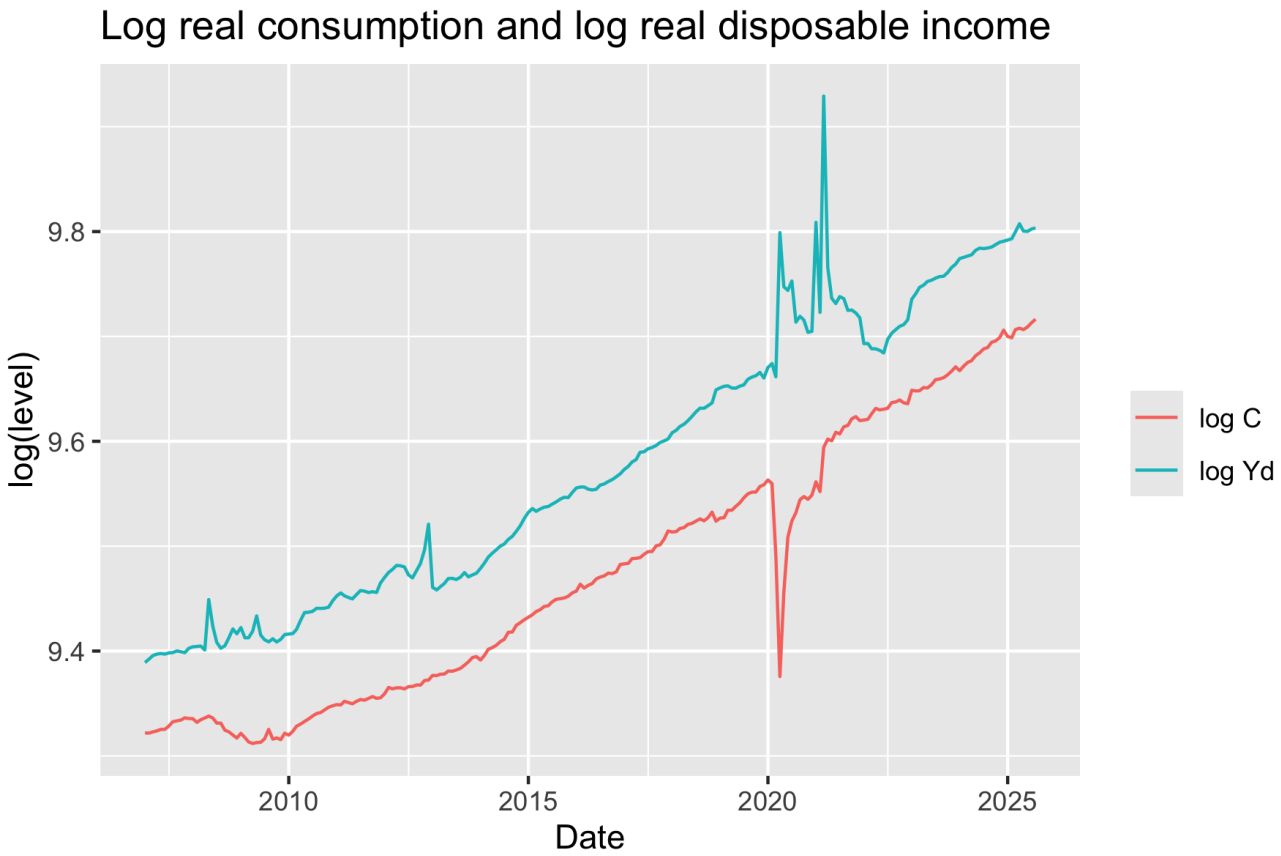


Figure: Log real consumption and log real disposable income

Log Consumption and Log Income: Visual Evidence

- Both $\log C_t$ and $\log Y_t^d$ exhibit strong upward trends and close co-movement.
- Notable divergence around 2020:
  - log consumption falls during COVID-19 restrictions,
  - log disposable income spikes due to fiscal transfers.
- After the acute COVID phase, the series realign, consistent with a stable long-run cointegrating relation.

Spurious Regression Reminder

- The case study also includes a regression of one simulated random walk on another:

$$\text{RW1}_t = \delta + \gamma \text{RW2}_t + u_t,$$

  with RW1 and RW2 independent.

- The regression nevertheless yields a statistically significant $\hat{\gamma}$:
  - classic spurious regression phenomenon,
  - underscores the necessity of:
    - testing for unit roots,
    - testing for cointegration,
    - using appropriate models (e.g. VECM) before interpreting "long-run" regressions.

Table of Contents

From VAR to VARMA

- We extend the VAR framework to include moving–average components, in direct analogy with univariate ARMA models.
- Let

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

  denote an n–dimensional time series.
- We will:
  - revisit VAR models and their covariance/spectral structure,
  - introduce vector MA and general VARMA models,
  - define and interpret impulse response functions (IRFs),
  - discuss estimation and inference for VAR / VARMA models,
  - and outline practical computation strategies (including VAR approximations).

Definition of VAR(p)

Definition 4 (Vector autoregression)

A *vector autoregression of order* p, VAR(p), is defined by

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where

- $\mathbf{c} \in \mathbb{R}^n$ is a constant (intercept) vector,
- $\mathbf{\Phi}_j$ are $(n \times n)$ coefficient matrices, $j = 1, \ldots, p$,
- $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})'$ is an innovation vector with

$$\mathrm{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}, \qquad \mathrm{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \mathbf{\Omega},$$

- $\mathbf{\Omega}$ is a positive definite $(n \times n)$ matrix,
- for most of this section we assume $\{\boldsymbol{\varepsilon}_t\}$ is i.i.d.

Given initial values $\mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{1-p}$, this defines $\mathbf{y}_t$ for $t \geq 1$.

Stationarity and Types of Dependence in VAR

- We focus on covariance–stationary VAR(p), where (4) holds for all $t \in \mathbb{Z}$ and the unconditional moments of $\mathbf{y}_t$ do not depend on t.
- VAR models capture two forms of dependence:
  1. **Contemporaneous dependence** via $\mathbf{\Omega}$:
     - innovations ε_{it} and ε_{jt} may be contemporaneously correlated.
  2. **Dynamic dependence** via $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p$:
     - each y_{it} can react to lagged values of all y_{jt}, $j = 1, \ldots, n$.

Standardization and Lag-Polynomial Notation

- If $\mathbf{y}_t$ is stationary with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{y}}$, consider the standardized process:

$$\mathbf{y}_t^* = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}(\mathbf{y}_t - \boldsymbol{\mu}).$$

- Then $\mathbf{y}_t^*$ is also VAR(p) with:

$$\boldsymbol{\Phi}_j^* = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\,\boldsymbol{\Phi}_j\,\boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}, \quad \boldsymbol{\Omega}^* = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}.$$

- Introduce the matrix lag polynomial

$$\boldsymbol{\Phi}(L) \equiv I_n - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p,$$

so (4) becomes

$$\boldsymbol{\Phi}(L)\,\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t.$$

Componentwise VAR Equations

- For the first component y_{1t}, (4) reads:

$$y_{1t} = c_1 + \sum_{j=1}^{n} (\mathbf{\Phi}_1)_{1j} y_{j,t-1} + \cdots + \sum_{j=1}^{n} (\mathbf{\Phi}_p)_{1j} y_{j,t-p} + \varepsilon_{1t}.$$

- This shows y_{1t} depends on the full p-period history of all n series.
- If $(\mathbf{\Phi}_k)_{1j} = 0$ for all k and all $j \neq 1$, the first equation collapses to a univariate AR(p).
- In general, each equation is a linear regression of y_{it} on lags of all n variables, but regressors are themselves endogenous and dynamically related.

VAR(1): Definition and MA Representation

- Any VAR(p) can be written as a VAR(1) in companion form, but to understand basic properties we study VAR(1) directly:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}\,\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

with $\mathbf{A}$ $(n \times n)$.

- Iterating backward:

$$\begin{aligned}\mathbf{y}_t &= \mathbf{c} + \boldsymbol{\varepsilon}_t + \mathbf{A}(\mathbf{c} + \boldsymbol{\varepsilon}_{t-1}) + \mathbf{A}^2(\mathbf{c} + \boldsymbol{\varepsilon}_{t-2}) + \cdots \\ &= (I_n + \mathbf{A} + \mathbf{A}^2 + \cdots)\mathbf{c} + \sum_{j=0}^{\infty} \mathbf{A}^j \boldsymbol{\varepsilon}_{t-j}.\end{aligned}$$

- If eigenvalues of $\mathbf{A}$ lie strictly inside the unit circle, $\sum_{j=0}^{\infty} \mathbf{A}^j = (I_n - \mathbf{A})^{-1}$.
- Then

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \mathbf{A}^j \boldsymbol{\varepsilon}_{t-j}, \quad \boldsymbol{\mu} = (I_n - \mathbf{A})^{-1}\mathbf{c}.$$

VAR(1): Autocovariances

- Define the lag–k autocovariance matrix:

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

- From (76) and stationarity:

$$\mathbf{\Gamma}(0) = \mathbf{A}\,\mathbf{\Gamma}(0)\,\mathbf{A}' + \mathbf{\Omega},$$

  a matrix Sylvester equation.

- Iterating:

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

  consistent with the MA(∞) representation (76).

VAR(1): Vectorized Autocovariance and $\Gamma(k)$

- Use $\mathrm{vec}(\mathbf{A}\boldsymbol{\Gamma}(0)\mathbf{A}') = (\mathbf{A} \otimes \mathbf{A})\mathrm{vec}(\boldsymbol{\Gamma}(0))$ to write

$$\mathrm{vec}(\boldsymbol{\Gamma}(0)) = (\mathbf{A} \otimes \mathbf{A})\mathrm{vec}(\boldsymbol{\Gamma}(0)) + \mathrm{vec}(\boldsymbol{\Omega}),$$

  so

$$\mathrm{vec}(\boldsymbol{\Gamma}(0)) = \left(I_{n^2} - \mathbf{A} \otimes \mathbf{A}\right)^{-1}\mathrm{vec}(\boldsymbol{\Omega}),$$

  well-defined if eigenvalues of $\mathbf{A}$ are inside the unit circle.

- For $k \geq 1$:

$$\boldsymbol{\Gamma}(k) = \boldsymbol{\Gamma}(0)\left(\mathbf{A}^k\right)', \quad \boldsymbol{\Gamma}(-k) = \mathbf{A}^k\boldsymbol{\Gamma}(0).$$

VAR(1): Spectral Density and Long-Run Covariance

- Spectral density matrix:

$$f(\lambda) = \frac{1}{2\pi}\left(I_n - \mathbf{A}e^{i\lambda}\right)^{*-1}\mathbf{\Omega}\left(I_n - \mathbf{A}e^{i\lambda}\right)^{-1}, \quad \lambda \in [-\pi, \pi].$$

  where $*$ denotes conjugate transpose.

- Long–run covariance:

$$2\pi f(0) = (I_n - \mathbf{A})^{*-1}\mathbf{\Omega}(I_n - \mathbf{A})^{-1}.$$

- Autocorrelation matrices $\mathbf{R}(k)$ usually have no closed-form except in special cases (e.g. diagonal $\mathbf{A}$ or $\mathbf{\Omega} \propto I_n$).

Marginal Dynamics from VAR to ARMA

- Question: how complex can the univariate dynamics of a single component y_{it} be when $\mathbf{y}_t$ follows a VAR?
- Any univariate AR(p)

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

can be embedded in a VAR(1) via:

$$\mathbf{x}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \quad \mathbf{u}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \end{bmatrix}, \quad \mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t.$$

VAR(p) Implies High-Order ARMA Marginals

- Conversely, if $\mathbf{y}_t$ follows VAR(p) with n components, then (under mild conditions) each y_{it} can be represented as univariate ARMA(p^*, q^*) with

$$p^* = np, \quad q^* = (n-1)p,$$

  possibly with reduced orders in special cases.

- When all $\mathbf{\Phi}_j$ are diagonal:
  - each marginal reduces to a univariate AR(p).
- For a bivariate VAR(1), each component is typically ARMA(2,1) (with cross–equation restrictions).
- VAR is a parsimonious way of encoding a collection of univariate ARMA processes that share structure across equations.

Definition of VMA(q)

Definition 5 (Vector moving average model)

A vector moving average of order q, VMA(q), is

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}_1 \boldsymbol{\varepsilon}_{t-1} - \cdots - \boldsymbol{\Theta}_q \boldsymbol{\varepsilon}_{t-q},$$

where $\boldsymbol{\varepsilon}_t$ is i.i.d. with mean zero and covariance $\boldsymbol{\Omega}$, and $\boldsymbol{\Theta}_j$ are $(n \times n)$ coefficient matrices.

Introduce the lag polynomial

$$\mathbf{B}(L) = I_n - \boldsymbol{\Theta}_1 L - \cdots - \boldsymbol{\Theta}_q L^q,$$

so

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}(L)\, \boldsymbol{\varepsilon}_t.$$

Stationarity and Invertibility of VMA(q)

- For any $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_q$, VMA($q$) is (weakly) stationary.
- **Invertibility** requires:

$$\det\left(\mathbf{B}(z)\right) \neq 0 \quad \forall\, |z| \leq 1.$$

- Under invertibility, we can write:

$$\mathbf{B}(L)^{-1}(\mathbf{y}_t - \boldsymbol{\mu}) = \boldsymbol{\varepsilon}_t,$$

giving an infinite-order VAR representation.

VMA(1): Autocovariances and Spectrum

- VMA(1):

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1},$$

  so $\mathbf{B}(L) = I_n - \boldsymbol{\Theta}L$.

- Invertible if eigenvalues of $\boldsymbol{\Theta}$ have modulus < 1.
- Lag–k autocovariances:

$$\boldsymbol{\Gamma}(k) = \begin{cases} \boldsymbol{\Omega} + \boldsymbol{\Theta}\boldsymbol{\Omega}\boldsymbol{\Theta}', & k = 0, \\ -\boldsymbol{\Theta}\boldsymbol{\Omega}, & k = 1, \\ -\boldsymbol{\Omega}\boldsymbol{\Theta}', & k = -1, \\ \mathbf{0}, & |k| > 1. \end{cases}$$

- Spectral density:

$$f(\lambda) = \frac{1}{2\pi}\left(I_n - \boldsymbol{\Theta}e^{i\lambda}\right)^* \boldsymbol{\Omega}\left(I_n - \boldsymbol{\Theta}e^{i\lambda}\right), \quad \lambda \in [-\pi, \pi].$$

Example: Common-Factor VMA(1)

Example 6 (A common–factor VMA(1))

Let $\mathbf{A}$ be an $(n \times n)$ matrix and define independent univariate MA(1) processes

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

with ε_{it} independent across i and t, variance σ_ε^2. Stack

$$\mathbf{x}_t = (x_{1t}, \ldots, x_{nt})', \quad \mathbf{y}_t = \mathbf{A}\mathbf{x}_t.$$

Then $\mathbf{y}_t$ is VMA(1) with

$$\mathbf{\Gamma}(0) = \sigma_\varepsilon^2 (1 + \theta^2)\, \mathbf{A}\mathbf{A}', \qquad \mathbf{\Gamma}(1) = -\theta \sigma_\varepsilon^2\, \mathbf{A}\mathbf{A}'.$$

Thus, for $|k| \geq 2$, $\mathbf{R}(k) = \mathbf{0}$, and

$$\mathbf{R}(1) = -\frac{\theta}{1 + \theta^2}\, \mathrm{diag}(\mathbf{A}\mathbf{A}')^{-1/2} \mathbf{A}\mathbf{A}' \,\mathrm{diag}(\mathbf{A}\mathbf{A}')^{-1/2}.$$

Example: Two-Dimensional VMA(1)

Example 7 (A simple two–dimensional VMA(1))

Let

$$y_{1t} = \varepsilon_{1t} + \varepsilon_{2,t-1}, \qquad y_{2t} = \varepsilon_{2t} + \theta\varepsilon_{2,t-1},$$

with ε_{jt} i.i.d. mean zero, unit variance, and $|\theta| < 1$.
Then:

$$\Gamma_{11}(1) = 0, \quad \Gamma_{21}(1) = 1, \quad \Gamma_{22}(1) = \theta,$$

so

$$R_{21}(1) = \frac{1}{\sqrt{2(1+\theta^2)}}, \qquad R_{22}(1) = \frac{\theta}{1+\theta^2}.$$

Hence the cross–autocorrelation $R_{21}(1)$ can exceed the own–lag autocorrelation $R_{22}(1)$.

VARMA(p, q) and Infinite-Order Forms

- General vector ARMA:

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

  where:

  - $\mathbf{\Phi}(L)$ as in (74),
  - $\mathbf{B}(L)$ as in (82).

- Under stability and invertibility of $\mathbf{\Phi}(z)$ and $\mathbf{B}(z)$:

  - VAR(∞) form:

$$\mathbf{C}(L)\mathbf{y}_t = \mathbf{c}' + \varepsilon_t,$$

  - VMA(∞) form:

$$\mathbf{y}_t = \mathbf{c}'' + \mathbf{D}(L)\varepsilon_t,$$

    where $\mathbf{C}(L)$ and $\mathbf{D}(L)$ are convergent matrix power series in L.

Impulse Response Matrices

- For a VARMA model under invertibility:

$$\mathbf{y}_t - \boldsymbol{\mu} = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \, \boldsymbol{\varepsilon}_{t-k},$$

  where $\boldsymbol{\Psi}_k$ are $(n \times n)$ **impulse response matrices**.

- The (i, j) element:

$$l_{ij}(k) = \frac{\partial y_{i,t+k}}{\partial \varepsilon_{jt}} = (\boldsymbol{\Psi}_k)_{ij},$$

  quantifies the response of y_i at horizon k to a unit shock in innovation j at time t.

- In univariate ARMA, IRFs reduce to the MA coefficients $\{\psi_k\}$.

- For VAR(1), $\boldsymbol{\Psi}_k = \mathbf{A}^k$.

IRFs for VAR(p): Recursion

- For VAR(p):

$$\mathbf{y}_t - \boldsymbol{\mu} = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\varepsilon}_{t-k},$$

with recursion

$$\boldsymbol{\Psi}_0 = I_n, \qquad \boldsymbol{\Psi}_k = \sum_{i=1}^{\min\{p,k\}} \boldsymbol{\Psi}_{k-i}\, \boldsymbol{\Phi}_i, \quad k \geq 1,$$

using $\boldsymbol{\Phi}_i = 0$ for $i > p$.

- Alternatively, transform VAR(p) to companion VAR(1) with matrix $\mathbf{F}$; then blocks of $\mathbf{F}^k$ give IRFs.

Standardized and Orthogonalized Shocks

- Often we want IRFs with respect to shocks that are:
  - standardized, or
  - orthogonal (uncorrelated).
- **Standardized shocks**:

$$\varepsilon_t = \mathbf{D}\mathbf{z}_t, \quad \mathbf{D} = \operatorname{diag}(\mathbf{\Omega})^{1/2},$$

where $\mathbf{z}_t$ has independent unit–variance components. Then

$$\frac{\partial y_{i,t+k}}{\partial z_{jt}} = (\mathbf{\Psi}_k \mathbf{D})_{ij}.$$

- **Orthogonal shocks**:

$$\varepsilon_t = \mathbf{P}\mathbf{w}_t, \quad \mathbf{\Omega} = \mathbf{P}\mathbf{P}',$$

where $\mathbf{P}$ is a Cholesky factor, $\mathrm{E}(\mathbf{w}_t \mathbf{w}_t') = I_n$. Then

$$\frac{\partial y_{i,t+k}}{\partial w_{jt}} = (\mathbf{\Psi}_k \mathbf{P})_{ij}.$$

Functionals of IRFs and Connectedness

- Beyond raw IRFs, we often consider functionals:
  - full trajectory $\{\Psi_{ij,k}\}_{k=0}^{K}$,
  - horizon $k_{\max} = \arg\max_{0 \le k \le K} |\Psi_{ij,k}|$,
  - contribution of shocks in j to variances of i.
- Diebold–Yilmaz (2014) define a connectedness measure:

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

  where:
  - e_i is the ith unit vector,
  - σ_{jj} is the jth diagonal of Ω.
- d_{ij}^{K} measures the share of forecast error variance of y_i due to shocks in y_j over horizon K.

Multivariate Yule–Walker Equations

Assume $\mathrm{E}(\mathbf{y}_t) = \mathbf{0}$ for simplicity. For VAR(p):

$$\mathbf{y}_t = \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

the Yule–Walker equations are:

$$\mathbf{\Gamma}(k) = \mathbf{\Phi}_1 \mathbf{\Gamma}(k-1) + \cdots + \mathbf{\Phi}_p \mathbf{\Gamma}(k-p), \quad k = 1, \ldots, p,$$

with $\mathbf{\Gamma}(k) = \mathrm{E}(\mathbf{y}_t \mathbf{y}'_{t-k})$ and $\mathbf{\Gamma}(-k) = \mathbf{\Gamma}(k)'$.
Vectorizing:

$$\gamma(k) = (\mathbf{\Gamma}(k-1)' \otimes I_n)\mathbf{a}_1 + \cdots + (\mathbf{\Gamma}(k-p)' \otimes I_n)\mathbf{a}_p,$$

where $\mathbf{a}_j = \mathrm{vec}(\mathbf{\Phi}_j)$.

Yule–Walker System and Practical Considerations

- Collecting equations for $k = 1, \ldots, p$ yields a linear system for

$$\mathbf{a} = (\mathbf{a}_1', \ldots, \mathbf{a}_p')'.$$

- Replace $\mathbf{\Gamma}(k)$ by sample autocovariances $\widehat{\mathbf{\Gamma}}(k)$ to get YW estimates.
- In practice, for moderate / large n and p, system can be high–dimensional:
  - iterative solvers (e.g. Gauss–Seidel) or regularization may be helpful.
  - More common is to use OLS / ML estimators described next.

OLS (Regression) Estimation of VAR

- For each $i = 1, \ldots, n$:

$$y_{it} = c_i + \sum_{j=1}^{n} \sum_{\ell=1}^{p} (\mathbf{\Phi}_\ell)_{ij} y_{j,t-\ell} + \varepsilon_{it}.$$

- Stack observations for $t = p+1, \ldots, T$:

$$\mathbf{X} = [\mathbf{1}_{T-p}, \ \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-p}], \quad \mathbf{Y}_i = (y_{i,p+1}, \ldots, y_{i,T})'.$$

- Equation can be written:

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{e}_i,$$

  where $\boldsymbol{\beta}_i$ contains c_i and row i of $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p$.

- Under $E(\boldsymbol{\varepsilon}_t \mid \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots) = \mathbf{0}$, OLS is consistent:

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

System Representation and Residual Covariance

- Stack all equations:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where

$$\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n], \quad \mathbf{B} = [\beta_1, \ldots, \beta_n],$$

and $\mathbf{E}$ stacks residuals.

- System is SUR with identical regressors in each equation, so OLS = GLS = conditional MLE under Gaussianity.
- Residuals and innovation covariance:

$$\widehat{\varepsilon}_t = \mathbf{y}_t - \widehat{\mathbf{c}} - \sum_{\ell=1}^{p} \widehat{\mathbf{\Phi}}_\ell \mathbf{y}_{t-\ell},$$

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

Asymptotic Distribution of VAR Coefficients

- Under stability and i.i.d. shocks with covariance $\mathbf{\Omega}$:

$$\mathbf{\Gamma}(0) = \mathrm{Var}(\mathbf{y}_t).$$

- Let

$$\mathbf{a} = \mathrm{vec}(\mathbf{\Phi}), \quad \widehat{\mathbf{a}} = \mathrm{vec}(\widehat{\mathbf{\Phi}}).$$

- Then:

$$\sqrt{T}\,(\widehat{\mathbf{a}} - \mathbf{a}) \Longrightarrow N(\mathbf{0}, \mathbf{\Upsilon}), \quad \mathbf{\Upsilon} = \mathbf{\Gamma}(0)^{-1} \otimes \mathbf{\Omega}.$$

- Intercepts:

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

- $\widehat{\mathbf{c}}$ and $\widehat{\mathbf{\Phi}}$ are asymptotically correlated as in multivariate regression.

Gaussian Likelihood and Conditional MLE

- If $\varepsilon_t \sim i.i.d.\ N(\mathbf{0}, \mathbf{\Omega})$, the conditional log-likelihood (conditioning on first p observations) is:

$$\mathcal{L}_C(\boldsymbol{\theta}) = -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\det(\mathbf{\Omega}) - \frac{1}{2}\sum_{t=p+1}^{T}(\mathbf{y}_t - \mathbf{c} - \sum_{\ell=1}^{p}\mathbf{\Phi}_\ell \mathbf{y}_{t-\ell})'\mathbf{\Omega}^{-1}(\cdots),$$

  where $\boldsymbol{\theta}$ collects all parameters.

- Maximizing in $(\mathbf{c}, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p)$ for fixed $\mathbf{\Omega}$ yields OLS estimators.
- Maximizing in $\mathbf{\Omega}$ yields $\widehat{\mathbf{\Omega}}$.
- Thus OLS $\equiv$ conditional Gaussian MLE.

Goodness-of-Fit and Model Selection

- For each equation i, define:

$$R_i^2 = 1 - \frac{\sum_{t=p+1}^{T} \widehat{\varepsilon}_{it}^2}{\sum_{t=p+1}^{T} (y_{it} - \bar{y}_i)^2}, \quad \bar{y}_i = \frac{1}{T-p} \sum_{t=p+1}^{T} y_{it}.$$

- Adjusted R_i^2 accounts for number of regressors:

$$R_i^{*2} = 1 - (1 - R_i^2) \frac{T-p}{T-np-1}.$$

- System measure:

$$R_{\text{system}}^2 = 1 - \frac{\sum_{i,t} \widehat{\varepsilon}_{it}^2}{\sum_{i,t} (y_{it} - \bar{y}_i)^2}.$$

- Model selection via AIC/BIC:

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

Linear Hypotheses: Wald and LR Tests

- Many hypotheses can be written as $H_0 : \mathbf{Ra} = \mathbf{r}$, where:

$$\mathbf{a} = \mathrm{vec}(\mathbf{\Phi}), \quad \mathbf{R} \in \mathbb{R}^{q \times n^2 p}, \quad \mathbf{r} \in \mathbb{R}^q.$$

- **Wald statistic**:

$$W = T(\mathbf{R}\widehat{\mathbf{a}} - \mathbf{r})' \left[\mathbf{R}(\widehat{\mathbf{\Gamma}}(0)^{-1} \otimes \widehat{\mathbf{\Omega}})\mathbf{R}'\right]^{-1} (\mathbf{R}\widehat{\mathbf{a}} - \mathbf{r}),$$

  asymptotically χ^2_q under H_0.

- **Likelihood Ratio**:

$$\mathrm{LR} = T(\log\det(\widetilde{\mathbf{\Omega}}) - \log\det(\widehat{\mathbf{\Omega}})),$$

  where $\widetilde{\mathbf{\Omega}}$ is covariance under H_0. Also χ^2_q asymptotically.

Robust Inference Under MDS Shocks

- If ε_t is a martingale difference sequence with time–varying conditional covariance, OLS remains consistent but asymptotic variance changes.
- Let:

$$\mathbf{z}_t = \mathrm{vec}\big(\varepsilon_t(\mathbf{y}_{t-1} - \mu)'\big) = ((\mathbf{y}_{t-1} - \mu) \otimes I_n)\varepsilon_t,$$

and

$$\Lambda = \lim_{T\to\infty} \frac{1}{T} \sum E(\mathbf{z}_t \mathbf{z}_t') = E\big[(\mathbf{y}_{t-1} - \mu)(\mathbf{y}_{t-1} - \mu)' \otimes E(\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1})\big].$$

- Then

$$\Upsilon = (\Gamma(0) \otimes I_n)^{-1} \Lambda (\Gamma(0) \otimes I_n)^{-1}.$$

- Estimated by a "sandwich" (robust) covariance matrix; yields heteroskedasticity–robust tests.

Delta Method for IRFs

- IRFs $\mathbf{\Psi}_k$ are smooth functions of parameters $\boldsymbol{\theta}$.
- Let
$$\boldsymbol{\iota}(k) = \mathrm{vec}(\mathbf{\Psi}_k), \quad \boldsymbol{\iota} = [\,\boldsymbol{\iota}(1)'; \ldots; \boldsymbol{\iota}(K)'\,]'.$$
- Then $\boldsymbol{\iota} = f(\boldsymbol{\theta})$ for smooth f.
- Delta method:
$$\sqrt{T}(\widehat{\boldsymbol{\iota}} - \boldsymbol{\iota}) \Longrightarrow N\big(\mathbf{0}, \mathbf{V}_I(\boldsymbol{\theta})\big),$$
where $\mathbf{V}_I$ constructed from Jacobian of f and asymptotic covariance of $\widehat{\boldsymbol{\theta}}$.
- Plug–in: $\widehat{\mathbf{V}}_I = \mathbf{V}_I(\widehat{\boldsymbol{\theta}})$.

Pointwise and Simultaneous Confidence Bands

- **Pointwise intervals** for each (i, j, k):

$$\widehat{\Psi}_{ij,k} \pm z_{\alpha/2}\,\text{stderr}_{ij,k},$$

where $\text{stderr}^2_{ij,k}$ is the corresponding diagonal of $\widehat{\mathbf{V}}_l$.

- **Simultaneous bands**: approximate distribution of

$$\max_{i,j,k} \frac{\sqrt{T}|\widehat{\Psi}_{ij,k} - \Psi_{ij,k}|}{\text{stderr}_{ij,k}}$$

by maximum of $q = n^2K$ independent $N(0,1)$; cdf $\Phi(x)^q$.

- Solve

$$(1 - 2\Phi(-w_{\alpha/2}))^q = 1 - \alpha,$$

and use

$$\widehat{\Psi}_{ij,k} \pm w_{\alpha/2}\,\text{stderr}_{ij,k}.$$

- Simultaneous bands are wider but provide uniform coverage.

Bootstrap and Simulation for IRFs

Residual bootstrap:

1. Estimate VAR(p) $\Rightarrow$ $\widehat{\mathbf{\Phi}}_\ell$, $\widehat{\mathbf{c}}$, $\widehat{\mathbf{\Omega}}$, and residuals $\widehat{\varepsilon}_t$.
2. Draw ε_t^* by resampling (or parametric resampling from $N(0, \widehat{\mathbf{\Omega}})$).
3. Generate bootstrap series $\mathbf{y}_t^*$ recursively:

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

4. Re-estimate VAR and IRFs on $\{\mathbf{y}_t^*\}$ to get $\widehat{\mathbf{\Psi}}_k^*$.
5. Repeat many times; use empirical distribution of $\widehat{\mathbf{\Psi}}_k^*$ to form CIs.

Matrix-normal simulation: for VAR(1), use asymptotic distribution of $\widehat{\mathbf{A}}$ to simulate $\widehat{\mathbf{A}}^k$ directly.

Local Projections

- **Local projections** (Jordà, 2005):
  - For each horizon k, run a separate regression:

$$\mathbf{y}_{t+k} = \text{function of } \mathbf{y}_t, \mathbf{y}_{t-1}, \cdots + \text{error}.$$

  - IRF at horizon k is coefficient on contemporaneous shock/variable.
- Pros:
  - robust to VAR misspecification,
  - flexible to include nonlinearities or time variation.
- Cons:
  - less efficient than VAR-based IRFs if VAR is correctly specified,
  - can be noisy at long horizons.

Challenges of VMA and VARMA Estimation

- Direct estimation of VMA/VARMA is considerably more complex than VAR:
  - matrix quadratic equations,
  - non-convex likelihood surfaces,
  - identification and invertibility constraints.
- Example: VMA(1):

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1},$$

with

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Omega} + \boldsymbol{\Theta}\boldsymbol{\Omega}\boldsymbol{\Theta}', \quad \boldsymbol{\Gamma}(1) = -\boldsymbol{\Theta}\boldsymbol{\Omega},$$

implies a matrix quadratic in $\mathbf{P} = \boldsymbol{\Omega}^{-1}$:

$$\boldsymbol{\Gamma}(1)\mathbf{P}\boldsymbol{\Gamma}(1)'\mathbf{P} - \boldsymbol{\Gamma}(0)\mathbf{P} + I_n = 0,$$

with no closed-form solution in general.

VAR Approximation Strategy

- Practical strategy: approximate VARMA with higher-order VAR:
  - VAR is easy to estimate by OLS / ML,
  - with enough lags, VAR can approximate VARMA arbitrarily well,
  - especially effective with appropriate lag selection and diagnostics.
- Direct VARMA estimation:
  - used when dimension is small and theory strongly dictates specific ARMA structure,
  - may involve specialized algorithms (e.g. state-space methods, EM, Hannan–Rissanen–type procedures).

R Implementation: Overview

- To complement the theory, one can implement VAR, VMA, and VARMA computations in R.
- The illustrative script:
  - uses FRED data to build a bivariate VAR for $(\log C_t, \log Y_t^d)'$,
  - shows how to select lag order via `VARselect`,
  - estimates VAR(p) and extracts $\widehat{\mathbf{\Phi}}_\ell$, $\widehat{\mathbf{\Omega}}$,
  - computes theoretical and empirical covariance matrices for VAR(1) as in (78),
  - computes and plots impulse responses via `irf()` and via the recursion (89).
- (The detailed R code is omitted here but follows directly from the formulas in this section.)

Connectedness and Spillovers in R

- Using the estimated IRFs and $\widehat{\mathbf{\Omega}}$, one can implement the Diebold–Yilmaz connectedness measure:

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

- For the bivariate macro VAR, this yields a 2×2 spillover table:
  - how much shocks in income contribute to consumption variance, and vice versa.
- These computations demonstrate how the theoretical IRF-based measures directly map into empirical connectedness diagnostics.

VMA(1) Example in R

- The script simulates the two–dimensional VMA(1) system of Example 7:

$$y_{1t} = \varepsilon_{1t} + \varepsilon_{2,t-1}, \quad y_{2t} = \varepsilon_{2t} + \theta\varepsilon_{2,t-1},$$

  and computes sample $\widehat{\Gamma}(1)$ and $\widehat{R}(1)$.
- Empirical values are compared with theoretical:

$$\Gamma_{11}(1) = 0,\ \Gamma_{21}(1) = 1,\ \Gamma_{22}(1) = \theta, \quad R_{21}(1) = \frac{1}{\sqrt{2(1+\theta^2)}},\ R_{22}(1) = \frac{\theta}{1+\theta^2}.$$

- This illustrates how cross–autocorrelations can be larger than own–lag autocorrelations in multivariate systems.

VARMA Simulation and VAR Approximation

- The R script uses `MTS::VARMAsim` to simulate a bivariate VARMA(1,1) model with specified AR and MA matrices and innovation covariance.
- It then estimates a VAR(4) on the simulated data:
  - illustrating how a higher-order VAR can approximate underlying VARMA dynamics,
  - allowing straightforward estimation of orthogonalized IRFs.
- This reflects the common practice recommended in the text: use reasonably high-order VARs as flexible approximations to VARMA processes in applied work.

Table of Contents

Reduced-Form VARs and Their Limitations

- Up to now we used *reduced-form* VARs:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\ (0, \mathbf{\Omega}).$$

- Reduced-form parameters $\left(\mathbf{c}, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p, \mathbf{\Omega}\right)$ are estimated directly from data (OLS / ML).
- As Hamilton (1994, p. 324) notes, such VARs:
  - "make no use of prior theoretical ideas about how these variables are expected to be related,"
  - cannot be used directly to test structural theories or interpret shocks as economically meaningful.
- In many applications we care about the *causal effect* of a specific shock:

$$\frac{\partial \mathbf{y}_{t+h}}{\partial \varepsilon_t^r}, \quad h = 0, 1, 2, \ldots,$$

  e.g. the dynamic effect of a monetary policy shock on macro variables.
- Structural VAR (SVAR) models distinguish *structural shocks* from reduced-form innovations.

Structural VAR(p) and Its Components

- Structural VAR(p):

$$\mathbf{A}_0\mathbf{y}_t = \mathbf{c}^* + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t,$$

  where:

  - $\mathbf{y}_t \in \mathbb{R}^n$: endogenous variables,
  - $\mathbf{A}_0$ $(n \times n)$ nonsingular: contemporaneous relations,
  - $\mathbf{A}_j$ $(n \times n)$: lagged structural coefficients,
  - $\mathbf{c}^* \in \mathbb{R}^n$: structural intercept,
  - $\mathbf{u}_t$: structural shocks, $i.i.d.$ with

$$\mathrm{E}(\mathbf{u}_t) = 0, \quad \mathrm{E}(\mathbf{u}_t\mathbf{u}_t') = \mathbf{\Sigma}_u.$$

    Often $\mathbf{\Sigma}_u$ is taken diagonal or I_n.

- Lag-operator form:

$$\mathbf{A}(L)\mathbf{y}_t = \mathbf{c}^* + \mathbf{u}_t, \quad \mathbf{A}(L) = \mathbf{A}_0 - \mathbf{A}_1 L - \cdots - \mathbf{A}_p L^p.$$

Reduced-Form VAR from Structural VAR

- Premultiply (113) by $\mathbf{A}_0^{-1}$:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

  where:

$$\mathbf{c} = \mathbf{A}_0^{-1}\mathbf{c}^*, \quad \mathbf{\Phi}_j = \mathbf{A}_0^{-1}\mathbf{A}_j, \quad \boldsymbol{\varepsilon}_t = \mathbf{A}_0^{-1}\mathbf{u}_t.$$

- Reduced-form innovations:

$$\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0, \quad \mathrm{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \mathbf{\Omega} = \mathbf{A}_0^{-1}\mathbf{\Sigma}_u \mathbf{A}_0^{-1\prime}.$$

- Thus (114) has the same form as our Gaussian VAR(p: $\left(\mathbf{c}, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p, \mathbf{\Omega}\right)$ are identifiable and estimable.

Parameter Counting and Identification Problem

- Structural parameters:

$$\left(\mathbf{c}^*, \mathbf{A}_0, \ldots, \mathbf{A}_p, \mathbf{\Sigma}_u\right):$$

total number of unknowns:

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

- Reduced-form parameters:

$$\left(\mathbf{c}, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p, \mathbf{\Omega}\right)$$

have

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

unknowns.

Parameter Counting and Identification Problem

- Short by n^2 restrictions: **structural parameters are not uniquely determined** by reduced-form VAR alone.
- Common normalization: $\mathbf{\Sigma}_u = I_n$ (orthogonal shocks, unit variance):
  - imposes $n(n+1)/2$ restrictions,
  - still leaves $\frac{n(n-1)}{2}$ free parameters that must be fixed via economic/statistical assumptions (zero / sign restrictions, etc.).

Short-Run Identification via Recursive Systems

- A common short-run strategy: impose zero restrictions on $\mathbf{A}_0$.
- **Recursive / triangular** system: $\mathbf{A}_0$ lower triangular:

$$\mathbf{A}_0 = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ a_{n1} & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix},$$

often with $a_{ii} = 1$.

Short-Run Identification via Recursive Systems

- Interpretation:
  - variable 1 does not respond contemporaneously to any others,
  - variable 2 responds contemporaneously only to variable 1,
  - variable k responds contemporaneously to variables $1, \ldots, k$, but not to $k+1, \ldots, n$,
  - all variables can react with lags via $\mathbf{A}_1, \ldots, \mathbf{A}_p$.
- Lower triangular $\mathbf{A}_0$ has $n(n-1)/2$ free off-diagonal elements: exactly the number of restrictions needed when $\boldsymbol{\Sigma}_u = I_n$ (order condition).

Example: Recursive Monetary SVAR

Example 8 (Recursive monetary SVAR)

Let

$$\mathbf{y}_t = \begin{bmatrix} p_t \\ gdp_t \\ m_t \\ i_t \end{bmatrix},$$

with p_t log price level, gdp_t log real GDP, m_t log money, i_t short-term interest rate.

One possible recursive $\mathbf{A}_0$:

$$\mathbf{A}_0 = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

Example: Recursive Monetary SVAR

- Interpretation:
  - prices do not respond contemporaneously to output, money, or interest rate shocks;
  - output does not respond contemporaneously to money/interest shocks;
  - money responds contemporaneously to prices and output, but not rates;
  - the policy rate responds contemporaneously to all variables.
- Plausibility must be argued using economic/institutional knowledge.

Long-Run Identification via Structural VMA Form

- Another approach: use **long-run** restrictions.
- Structural VMA representation:

$$\mathbf{y}_t = \sum_{j=0}^{\infty} \mathbf{D}_j \mathbf{u}_{t-j} = \mathbf{D}(L)\mathbf{u}_t,$$

with $\mathbf{D}_0 = \mathbf{A}_0^{-1}$.

- **Long-run impact matrix**:

$$\mathbf{D}(1) = \sum_{j=0}^{\infty} \mathbf{D}_j.$$

- Imposing that certain shocks have *no long-run effect* on selected variables:

$$D_{ij}(1) = 0 \quad \text{for some } (i, j),$$

yields identifying restrictions.

Example: Blanchard–Quah Long-Run Restrictions

- Blanchard & Quah (1989): two variables (e.g. output and unemployment) driven by supply and demand shocks.
- Key identifying assumption:
  - demand shocks have no long-run effect on unemployment.
- Implemented as zero restriction on relevant element of $\mathbf{D}(1)$.
- With two variables and two shocks, one long-run zero restriction is enough to identify both shocks.
- Long-run restrictions are widely used in macro VARs (e.g. technology shocks vs demand shocks).

Other Identification Strategies

- Beyond short- and long-run zero restrictions, several strategies exist:
  - **Sign restrictions**: constrain the sign of selected impulse responses over certain horizons.
  - **Heteroskedasticity-based identification**: exploit changes in volatility (e.g. pre/post-crisis regimes) to disentangle shocks.
  - **Narrative / external instruments**: use external information (e.g. high-frequency surprises, narrative shocks) as proxies for structural shocks.
- These methods often relax exact zero constraints while still achieving meaningful structural identification.

A-, B-, and AB-Model Parameterizations

- Reduced-form VAR(p):

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

  with $\boldsymbol{\varepsilon}_t \sim (0, \mathbf{\Omega})$.

- Suppose there are structural shocks $\mathbf{u}_t \sim (0, I_n)$ such that

$$\boldsymbol{\varepsilon}_t = \mathbf{B}\mathbf{u}_t, \quad \Rightarrow \quad \mathbf{\Omega} = \mathbf{B}\mathbf{B}'.$$

- Different parameterizations:
  - **A-model**: restrictions on $\mathbf{A}$ and diagonal covariance $\mathbf{\Lambda}$,
  - **B-model**: restrictions on $\mathbf{B}$ (impact matrix),
  - **AB-model**: restrictions on both $\mathbf{A}$ and $\mathbf{B}$.

A-Model (Contemporaneous Structure)

- A-model structural form:

$$\mathbf{A}\mathbf{y}_t = \mathbf{c}^* + \mathbf{A}_1^*\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p^*\mathbf{y}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim (0, \mathbf{\Lambda}),$$

  with $\mathbf{\Lambda}$ diagonal.

- $\mathbf{A}$ encodes contemporaneous interactions:
  - often lower triangular with ones on the diagonal,
  - zero restrictions on $\mathbf{A}$ achieve identification (plus diagonal $\mathbf{\Lambda}$).
- Recursive systems are special cases of A-models.

B-Model (Impact Matrix on Shocks)

- B-model:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \mathbf{B}\mathbf{u}_t, \quad \mathbf{u}_t \sim (0, I_n),$$

  so $\mathbf{\Omega} = \mathbf{B}\mathbf{B}'$.

- Restrictions imposed directly on $\mathbf{B}$:
  - e.g. lower triangular with positive diagonal (Cholesky),
  - or zero restrictions matching economic theory.
- The LDL factorization $\mathbf{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}'$ corresponds to a B-model with $\mathbf{B} = \mathbf{A}\mathbf{D}^{1/2}$.

AB-Model (Combined Restrictions)

- AB-model:

$$\mathbf{A}\mathbf{y}_t = \mathbf{c}^* + \mathbf{A}_1^*\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p^*\mathbf{y}_{t-p} + \mathbf{B}\mathbf{u}_t, \quad \mathbf{u}_t \sim (0, I_n),$$

  so

$$\boldsymbol{\varepsilon}_t = \mathbf{A}^{-1}\mathbf{B}\mathbf{u}_t, \quad \boldsymbol{\Omega} = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}'\mathbf{A}^{-1\prime}.$$

- Restrictions may be placed on both $\mathbf{A}$ and $\mathbf{B}$:
  - typically requires more restrictions than A- or B-model alone,
  - in practice often one of $\mathbf{A}$ or $\mathbf{B}$ is set to I_n to reduce to A- or B-model.

Example: Kilian (2009) Oil Market SVAR

Kilian (2009) estimate an SVAR for the global oil market to disentangle different types of oil price shocks.

Let

$$\mathbf{z}_t = \begin{bmatrix} \Delta \text{prod}_t \\ \text{rea}_t \\ rpo_t \end{bmatrix},$$

where:

- Δprod_t: growth rate of global crude oil production,
- rea_t: real economic activity index,
- rpo_t: real price of oil.

Example: Kilian (2009) Oil Market SVAR

Structural VAR(24):

$$\mathbf{A}_0 \mathbf{z}_t = \boldsymbol{\alpha} + \sum_{i=1}^{24} \mathbf{A}_i \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_t,$$

with three structural shocks:

- oil supply shock,
- aggregate demand shock,
- oil-market specific (precautionary) demand shock.

Recursive Identification in Kilian's Model

- Inverse of $\mathbf{A}_0$ assumed recursive (lower triangular):

$$\mathbf{e}_t = \begin{bmatrix} e_t^{\Delta \text{prod}} \\ e_t^{\text{rea}} \\ e_t^{rpo} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \varepsilon_t^{\text{oil supply}} \\ \varepsilon_t^{\text{agg demand}} \\ \varepsilon_t^{\text{oil-specific}} \end{bmatrix}.$$

- Recursive structure:
  - oil supply shock affects all variables contemporaneously,
  - aggregate demand shock affects rea and rpo contemporaneously, but not prod,
  - oil-specific demand shock affects only rpo contemporaneously.
- With these restrictions, Kilian shows that:
  - oil-price movements driven by global demand have different macro effects than those driven by supply or precautionary demand.

Estimating A- and B-Models in R

- The `vars` package in R implements SVAR estimation via `SVAR()`.
- Workflow:
  1. Simulate or estimate a reduced-form VAR:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

  2. Specify an **A-model**: supply an `Amat` with fixed entries (including zeros and ones) and `NA` for parameters to estimate; usually lower triangular with ones on diagonal.
  3. Or specify a **B-model**: supply `Bmat` (impact matrix) with restrictions (e.g. lower triangular, positive diagonal).
  4. Call `SVAR(varModelEstimate, Amat = A_matrix)` or `SVAR(varModelEstimate, Bmat = B_matrix)`.

SVAR Impulse Responses in R

- `SVAR` then:
  - estimates the structural matrices consistent with the reduced form,
  - yields structural shocks and their covariances.
- Once an SVAR is estimated (A- or B-model), impulse responses are computed with `irf()` as before:
  - e.g. `irf(SVAR_A_Model, n.ahead=10, boot=TRUE)`.
- These IRFs are responses to *identified structural shocks* rather than arbitrary innovations.
- Interpretation:
  - e.g. "What is the response of output and inflation to a one-standard-deviation monetary policy shock?",
  - IRFs can be used for policy analysis, shock decomposition, and testing structural theories.

Table of Contents

Summary: VAR and VARMA Models

- We developed a comprehensive treatment of multivariate linear time series:
  - Vector Autoregressive (VAR) models,
  - Vector Moving Average (VMA) and VARMA models.
- In VAR models:
  - each variable depends on its own lags and lags of other variables,
  - stationarity ensures constant means/covariances,
  - impulse response functions (IRFs) are key to understanding dynamics.
- VMA models:
  - represent current values as linear combinations of past shocks,
  - capture short-run propagation of innovations.
- VARMA models:
  - combine AR and MA components for richer dynamics,
  - often approximated in practice by higher-order VARs.

Summary: Structural VARs and Empirical Implementation

- Structural VARs (SVARs) add economic content:
  - impose contemporaneous (A-model) and/or impact (B-model) restrictions,
  - identify structural shocks with economic meaning.
- Identification strategies:
  - short-run zero restrictions (recursive systems),
  - long-run restrictions (e.g. Blanchard–Quah),
  - sign restrictions, heteroskedasticity, narrative/external instruments.

Summary: Structural VARs and Empirical Implementation

- Empirical tools:
  - `vars` and `MTS` packages in R for estimating VAR/VARMA/SVAR,
  - diagnostic checks, lag-order selection, and robustness (e.g. robust standard errors, bootstrap),
  - IRFs, variance decompositions, and causality tests for interpretation.
- Overall:
  - multivariate time series models provide a powerful framework to analyse dynamic interactions,
  - structural identification bridges statistical models with economic theory and policy analysis.

Table of Contents

References I

 Kilian, Lutz (2009). "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market". In: *American Economic Review* 99.3, pp. 1053–1069.

Chapter 4 — Volatility Models

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Volatility and Volatility Clustering

Volatility—the conditional variability of a series around its conditional mean—is a central object in econometrics.

Empirical features:

- **Volatility clustering**: large movements in returns tend to be followed by further large movements (of either sign), while quiet periods persist.
- Observed in financial returns, macro growth rates, inflation, exchange rates, commodity prices, etc.
- This behavior is incompatible with homoskedastic disturbances assumed in basic linear models.

Volatility and Volatility Clustering

Volatility models:

- Allow the conditional variance of the error to evolve over time as a function of past shocks and past volatility.
- ARCH/GARCH families introduce *conditional heteroskedasticity* in a structured way, capturing persistence and clustering.
- Widely used in:
  - financial econometrics (risk, VaR, option pricing),
  - macro uncertainty (inflation variability, policy uncertainty),
  - multivariate extensions (joint dynamics of variances and covariances).

Conditional Mean vs Conditional Variance

In earlier chapters, ARMA models were used for **conditional means**:

$$Y_t = \mu_t + \varepsilon_t, \qquad \mu_t = f(\mathcal{F}_{t-1}).$$

Assumption in standard ARMA:

$$\varepsilon_t \sim i.i.d.(0, \sigma^2).$$

For a basic MA(1),

$$Y_t = \mu + \varepsilon_t + \alpha\varepsilon_{t-1},$$

we derived

$$\mathrm{E}(Y_t) = \mu, \quad \mathrm{Var}(Y_t) = (1+\alpha^2)\sigma^2, \quad \mathrm{Cov}(Y_t, Y_{t-1}) = \alpha\sigma^2.$$

Conditional Mean vs Conditional Variance

These results require only that ε_t has:

$$\mathrm{E}(\varepsilon_t) = 0, \tag{1}$$

$$\mathrm{E}(\varepsilon_t^2) = \sigma^2, \tag{2}$$

$$\mathrm{E}(\varepsilon_t \varepsilon_{t-k}) = 0 \quad (k \neq 0). \tag{3}$$

Processes satisfying (1)–(3) are **white noise**. This allows *conditional* variance to depend on the past even if the *unconditional* variance is constant.

White Noise, MDS, and i.i.d. Disturbances

Classes of stationary disturbance processes:

- **i.i.d.** processes:

$$\varepsilon_t \text{ independent across } t, \ \varepsilon_t \overset{d}{=} \varepsilon_1.$$

- **Martingale difference sequence (MDS)**:

$$\mathrm{E}(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0$$

but allows dependence in higher moments (e.g. conditional variance).

- **White noise**: no linear predictability,

$$\rho(\varepsilon_t, \varepsilon_s) = 0, \quad t \neq s,$$

but can be non-Gaussian or conditionally heteroskedastic.

White Noise, MDS, and i.i.d. Disturbances

We often write $\varepsilon_t \sim WN(0, \sigma^2)$ for a mean-zero white noise process with constant unconditional variance. ARCH/GARCH models focus on the *conditional* variance,

$$\mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2,$$

while ARMA models capture conditional mean dynamics.

Table of Contents

Motivation and Definition of ARCH

The Autoregressive Conditional Heteroskedasticity (ARCH) model posits that current volatility depends on past squared residuals. It is designed to capture volatility clustering.

Robert F. Engle introduced the ARCH model in 1982 and received the 2003 Nobel Prize in Economics for this work. His contributions also include cointegration, ACD, CAViaR, and DCC models. He has held positions at MIT, UCSD, and NYU Stern, and is a fellow of many leading societies.

Motivation and Definition of ARCH

Simple ARCH idea (in Engle's notation):

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + u_t,$$

with u_t white noise. This is an AR(1) for ε_t^2.

More commonly we define volatility via the *conditional variance*:

$$\sigma_t^2 \equiv \mathrm{E}(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2.$$

This is an **ARCH(1)** process.

ARCH(1): Constraints and Unconditional Variance

ARCH(1):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2, \tag{4}$$

with:

- $\mathcal{F}_{t-1}$ = information up to $t-1$.
- Need $\sigma_t^2 \geq 0$ for all t: hence

$$\alpha_0 \geq 0, \quad \alpha_1 \geq 0.$$

ARCH(1): Constraints and Unconditional Variance

Unconditional variance:

$$\sigma^2 = \mathrm{E}(\varepsilon_t^2) = \mathrm{E}(\sigma_t^2) = \alpha_0 + \alpha_1 \mathrm{E}(\varepsilon_{t-1}^2) = \alpha_0 + \alpha_1 \sigma^2.$$

Solving:

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1}, \quad \text{if } 0 \leq \alpha_1 < 1.$$

So:

- a stationary solution exists if $0 \leq \alpha_1 < 1$,
- unconditional variance is finite and independent of t,
- but conditional variance σ_t^2 varies over time and exhibits clustering.

ARCH(p) and Interpretation

ARCH(p) generalization:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2.$$

As AR(p) for ε_t^2:

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + u_t,$$

with u_t white noise.

ARCH(p) and Interpretation

Constraints:

- Non-negativity: $\alpha_0, \alpha_1, \ldots, \alpha_p \geq 0$ to ensure $\sigma_t^2 \geq 0$.
- Weak stationarity: $\alpha_1 + \cdots + \alpha_p < 1$.

Unconditional variance:

$$\sigma^2 = \frac{\alpha_0}{1 - (\alpha_1 + \cdots + \alpha_p)}.$$

Interpretation:

- α_j measures the impact of a shock j periods ago on current volatility.
- In ARCH(p), shocks older than p periods have no direct effect.

Testing for ARCH Effects

Idea: if ε_t has ARCH-type dynamics, then ε_t^2 should display serial correlation.

Ljung–Box on squared residuals:

$$Q = T(T+2)\sum_{k=1}^{h}\frac{\widehat{\rho}_k^2}{T-k},$$

where $\widehat{\rho}_k$ is the sample autocorrelation of $\widehat{\varepsilon}_t^2$ at lag k.

- Large $Q \Rightarrow$ reject H_0 of no serial correlation (no ARCH).
- Applied both to the squared residuals of Y_t and to residuals from an estimated ARCH(p) model.

Practical workflow:

1. Fit a conditional mean model (e.g. ARMA) and obtain residuals $\widehat{\varepsilon}_t$.
2. Test for serial correlation in $\widehat{\varepsilon}_t^2$ using Ljung–Box or Engle's LM test.
3. If rejected, specify and estimate an ARCH/GARCH model for volatility.

Stationarity of the ARCH Process

Let us revisit ARCH(1):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2.$$

Assume weak stationarity and $\mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty$. Then

$$\sigma^2 = \mathrm{E}(\sigma_t^2) = \alpha_0 + \alpha_1 \mathrm{E}(\varepsilon_{t-1}^2) = \alpha_0 + \alpha_1 \sigma^2.$$

If $|\alpha_1| < 1$, then

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1}.$$

If $\alpha_1 > 1$ the solution is negative, which is impossible for a variance.

Stationarity of the ARCH Process

Asymptotic stationarity: starting from some initial σ_1^2,

$$\mathrm{E}(\sigma_t^2) = \alpha_0 \sum_{j=0}^{t-2} \alpha_1^j + \alpha_1^{t-1}\mathrm{E}(\sigma_1^2) \longrightarrow \frac{\alpha_0}{1-\alpha_1}, \quad t \to \infty.$$

So for $|\alpha_1| < 1$, the process is *asymptotically weakly stationary* regardless of $\mathrm{E}(\sigma_1^2)$, and exactly stationary if $\mathrm{E}(\sigma_1^2)$ is set to the fixed point.

Table of Contents

From ARCH to GARCH

ARCH models often require high orders p to capture long–lasting volatility.
GARCH introduces lagged volatility terms to achieve parsimony.

Tim Bollerslev generalized ARCH to GARCH in 1986. He is a leading financial econometrician, with work on GARCH, realized volatility, and more. GARCH is one of the most widely used models for time-varying volatility.

From ARCH to GARCH

General GARCH(p, q):

$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^{p} \alpha_k \varepsilon_{t-k}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 = \alpha_0 + \alpha(L)\varepsilon_{t-1}^2 + \beta(L)\sigma_{t-1}^2.$$

Simplest and most common: GARCH(1,1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \tag{5}$$

GARCH(1,1) is very parsimonious and empirically effective.

GARCH(1,1): Constraints and Unconditional Variance

GARCH(1,1):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Non-negativity:

$$\alpha_0 \geq 0, \quad \alpha_1 \geq 0, \quad \beta_1 \geq 0$$

to ensure $\sigma_t^2 > 0$.

GARCH(1,1): Constraints and Unconditional Variance

Unconditional variance: Assuming weak stationarity and $\mathrm{E}(\varepsilon_t^2) = \sigma^2$,

$$\sigma^2 = \mathrm{E}(\sigma_t^2) = \alpha_0 + \alpha_1\sigma^2 + \beta_1\sigma^2 \Rightarrow \sigma^2 = \frac{\alpha_0}{1 - (\alpha_1 + \beta_1)}.$$

- Stationary solution exists only if $\alpha_1 + \beta_1 < 1$.
- α_1 measures the short–run response to shocks,
- β_1 measures persistence in volatility.

GARCH(1,1) as Infinite ARCH

Iterating (5):

$$\begin{aligned}\sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ &= \alpha_0(1 + \beta_1 + \beta_1^2 + \cdots) + \alpha_1\left(\varepsilon_{t-1}^2 + \beta_1 \varepsilon_{t-2}^2 + \beta_1^2 \varepsilon_{t-3}^2 + \cdots\right) \\ &= \frac{\alpha_0}{1-\beta_1} + \alpha_1 \sum_{j=1}^{\infty} \beta_1^{j-1} \varepsilon_{t-j}^2.\end{aligned}$$

GARCH(1,1) as Infinite ARCH

Implications:

- GARCH(1,1) is equivalent to an ARCH(∞) with geometrically decaying coefficients.
- The impact of past shocks on current volatility decays over time at rate β_1^{j-1}.
- GARCH models provide a parsimonious alternative to high-order ARCH.

Table of Contents

Weak Stationarity of GARCH(p, q)

For GARCH(p, q),

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2.$$

Weak stationarity condition:

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1.$$

Under this condition:

$$\sigma^2 = \mathrm{Var}(\varepsilon_t) = \mathrm{E}(\sigma_t^2) = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j - \sum_{k=1}^{q} \beta_k}.$$

Weak Stationarity of GARCH(p, q)

For GARCH(1,1),

$$\varepsilon_t = \sigma_t \nu_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

with ν_t i.i.d. mean zero, variance 1. Weak stationarity requires $\alpha_1 + \beta_1 < 1$.

Strict Stationarity of GARCH(1,1)

Strict stationarity for GARCH(1,1) requires conditions on the entire distribution of ν_t, not just its moments.

Theorem 1 (Nelson (1990))

Consider

$$\varepsilon_t = \sigma_t \nu_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

with ν_t i.i.d. non-degenerate, and $\beta_1 \geq 0$, $\alpha_0, \alpha_1 > 0$. If

$$\mathrm{E}\big[\log(\beta_1 + \alpha_1 \nu_t^2)\big] < 0,$$

then:

1. *The process*

$$_u\sigma_t^2 = \alpha_0 \left(1 + \sum_{j=1}^{\infty} \prod_{i=1}^{j} (\alpha_1 \nu_{t-i}^2 + \beta_1)\right)$$

is strictly stationary;

Strict Stationarity of GARCH(1,1) (cont.)

Theorem 1 (Nelson (1990), continued)

2. ${}_u\sigma_t^2 \in [\alpha_0/(1-\beta_1), \infty)$;
3. $\sigma_t^2 - {}_u\sigma_t^2 \to 0$ *almost surely as* $t \to \infty$;
4. *The distribution of* σ_t^2 *converges to that of* ${}_u\sigma_t^2$ *as* $t \to \infty$.

Strict vs Weak Stationarity and Heavy Tails

If $\mathrm{E}(\nu_t^2) = 1$, then by Jensen's inequality:

$$\mathrm{E}[\log(\beta_1 + \alpha_1 \nu_t^2)] < \log\left(\mathrm{E}[\beta_1 + \alpha_1 \nu_t^2]\right) = \log(\alpha_1 + \beta_1).$$

Thus the strict stationarity condition may hold even when $\alpha_1 + \beta_1 \geq 1$.

Implications:

- GARCH can be strictly stationary without being weakly stationary.
- Such processes have infinite unconditional variance but well-defined stationary distributions for σ_t^2.
- Alongside Cauchy examples (strictly but not weakly stationary), GARCH provides another class with this property.

Stochastic volatility (SV) models go further by making volatility itself a latent stochastic process with its own innovations, offering additional flexibility.

Table of Contents

Motivation for Asymmetry

Standard ARCH/GARCH models are **symmetric**:

- they depend on the magnitude of shocks ε_t^2,
- but treat positive and negative shocks of the same size identically.

Empirical evidence:

- In equity markets, negative shocks ("bad news") typically increase future volatility more than positive shocks of the same magnitude:
  - risk-averse investors react more strongly to losses,
  - short-selling constraints and market microstructure can amplify volatility after bad news.
- This is known as the *leverage effect* or *asymmetric volatility*.

Asymmetric GARCH models incorporate the *sign* of shocks, not just their squared magnitude.

EGARCH and GJR-GARCH

EGARCH(1,1) (Nelson 1991):

$$\log \sigma_t^2 = \omega + \beta \log \sigma_{t-1}^2 + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha \frac{|\varepsilon_{t-1}|}{\sigma_{t-1}}.$$

- Log transform guarantees $\sigma_t^2 > 0$, no sign constraints on parameters.
- $\gamma \neq 0$ introduces asymmetry:
  - if $\gamma < 0$, negative shocks increase volatility more than positive shocks.

EGARCH and GJR-GARCH

GJR-GARCH (1,1) (Glosten, Jagannathan, and Runkle 1993):

$$\sigma_t^2 = \omega + \beta_1 \sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 1(y_{t-1} < 0).$$

- Asymmetry enters via $1(y_{t-1} < 0)$:
  - if $\delta > 0$, negative returns increase volatility more than positive ones.

News Impact Curve and Other Asymmetric Models

News impact curve: plots the impact of shocks of different signs and magnitudes on future volatility, holding past volatility fixed.

- Highlights asymmetry: bad news (negative returns) often has a larger effect than good news.

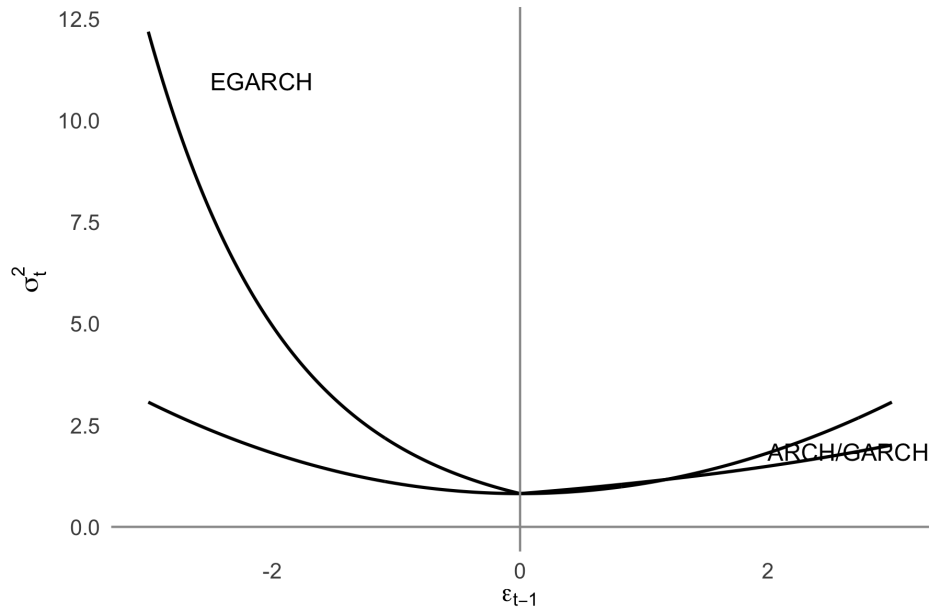


Figure: News Impact Curve

News Impact Curve and Other Asymmetric Models

Other asymmetric volatility models:

- Threshold GARCH (TGARCH) (Glosten, Jagannathan, and Runkle 1993):

$$\varepsilon_t = \nu_t \sigma_t, \quad \sigma_t^2 = \psi + \beta \sigma_{t-1}^2 + \alpha_1 \varepsilon_{t-1}^2 1(\nu_{t-1} < 0) + \alpha_2 \varepsilon_{t-1}^2 1(\nu_{t-1} \geq 0).$$

- Threshold ARCH (TARCH) (Zakoian 1994):

$$\sigma_t = \psi + \beta \sigma_{t-1} + \alpha_1 |\varepsilon_{t-1}| 1(\nu_{t-1} > 0) + \alpha_2 |\varepsilon_{t-1}| 1(\nu_{t-1} \leq 0).$$

- Markov Regime-Switching GARCH (Hamilton and Susmel 1994):

$$\sigma_t^2 = \psi(S_t) + \beta(S_t) \sigma_{t-1}^2 + \alpha(S_t) \varepsilon_{t-1}^2, \quad S_t \text{ Markov chain.}$$

Markov Regime-Switching GARCH

Specification:

$$\begin{cases} \varepsilon_t = \nu_t \sigma_t \\ \sigma_t^2 = \psi(S_t) + \beta(S_t)\sigma_{t-1}^2 + \alpha(S_t)\varepsilon_{t-1}^2 \\ \{\nu_t\} \sim i.i.d.\ (0,1) \end{cases}$$

with S_t a latent Markov chain (e.g. $S_t \in \{0,1\}$) and transition probabilities

$$P(S_t = 1 \mid S_{t-1} = 0) = p_{01}, \quad P(S_t = 0 \mid S_{t-1} = 1) = p_{10}.$$

Markov Regime-Switching GARCH

Interpretation:

- Different regimes (e.g. high vs low volatility) have different (ψ, α, β).
- Volatility is typically higher in recession regimes than in expansions (empirical evidence: Schwert (1989), Hamilton and Lin (1996), McQueen and Vorkink (1993)).
- If the shape of the innovation distribution also depends on S_t, the innovation sequence forms an MDS rather than an i.i.d. process.

Table of Contents

IGARCH(1,1) and Unit-Root Volatility

Start from GARCH(1,1):

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2.$$

Weak stationarity (finite unconditional variance) requires:

$$\omega > 0,\ \beta \geq 0,\ \alpha \geq 0,\ \beta + \alpha < 1.$$

IGARCH(1,1) and Unit-Root Volatility

IGARCH(1,1):

- Integrated GARCH: $\beta + \alpha = 1$.
- Volatility process has a *unit root*:

$$\sigma_t^2 = \omega + \sigma_{t-1}^2 - \beta u_{t-1}, \quad u_t = \varepsilon_t^2 - \sigma_t^2.$$

- By recursion:

$$\sigma_t^2 = \omega t + \sigma_0^2 - \beta \sum_{\ell=1}^{t-1} u_{t-\ell}.$$

Conditional expectation given $\mathcal{F}_0$:

$$\mathrm{E}(\sigma_t^2 \mid \mathcal{F}_0) = \omega t + \sigma_0^2,$$

so shocks to volatility have *permanent* effects.

IGARCH and Infinite Variance

For IGARCH(1,1):

$$\mathrm{E}(\sigma_t^2) = \infty$$

under weak stationarity conditions, so unconditional variance $\mathrm{Var}(\varepsilon_t)$ is not finite.

Implications:

- Unconditional variance not well-defined $\Rightarrow$ problematic for return series, where long-run volatility is finite in practice.
- Ignores mean-reversion in volatility; all shocks persist forever.
- Theoretically, IGARCH behavior may come from unmodelled structural breaks or time-varying intercepts.

IGARCH and Infinite Variance

However, strict stationarity may still hold under:

$$\mathrm{E}[\ln(\beta + \alpha u_t^2)] < 0.$$

By Jensen,

$$\mathrm{E}[\ln(\beta + \alpha u_t^2)] \leq \ln(\beta + \alpha \mathrm{E}(u_t^2)) = \ln(\beta + \alpha) = 0.$$

So for some α, β with $\alpha + \beta > 1$, strict stationarity is still possible.

RiskMetrics EWMA as IGARCH Special Case

J.P. Morgan's RiskMetrics (1997) uses

$$\begin{cases} \varepsilon_t = \nu_t \sigma_t, \\ \sigma_t^2 = (1-\beta)\sum_{j=1}^{\infty} \beta^{j-1} \varepsilon_{t-j}^2, \quad \beta \in (0,1), \\ \{\nu_t\} \sim i.i.d.\ N(0,1) \end{cases}$$

which can be written recursively as

$$\sigma_t^2 = \beta \sigma_{t-1}^2 + (1-\beta)\varepsilon_{t-1}^2.$$

RiskMetrics EWMA as IGARCH Special Case

Interpretation:

- This is an IGARCH(1,1) with $\omega = 0$, $\alpha = 1 - \beta$, $\beta = \beta$.
- For daily financial data, $\beta \approx 0.94$ is often used.
- Exponentially weighted moving average (EWMA) is attractive for risk management (VaR, etc.) due to simplicity and responsiveness.

Table of Contents

Likelihood for GARCH(1,1)

Consider

$$\varepsilon_t = \nu_t \sigma_t, \quad \sigma_t^2(\theta) = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2(\theta),$$

with parameter vector $\theta = (\omega, \beta, \alpha)'$.

Often ε_t are residuals from a conditional mean model.
Assume $\nu_t \sim N(0,1)$.

Conditional log-likelihood:

$$\mathcal{L}(\theta) = \sum_{t=2}^{T} \mathcal{L}_t(\theta), \quad \mathcal{L}_t(\theta) = -\frac{1}{2} \log \sigma_t^2(\theta) - \frac{1}{2} \left(\frac{\varepsilon_t}{\sigma_t(\theta)} \right)^2.$$

MLE $\widehat{\theta}$ maximizes $\mathcal{L}(\theta)$ over the parameter space Θ.

$\sigma_t^2(\theta)$ is computed recursively from an initial value σ_1^2.

Initialization and Scores

Initialization options for $\sigma_1^2(\theta)$:

1. Use unconditional variance: $\sigma_1^2(\theta) = \frac{\omega}{1-\alpha-\beta}$.
2. Use empirical variance: $\sigma_1^2 = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2$.
3. Set $\sigma_1^2 = \varepsilon_1^2$.
4. Treat σ_1^2 as an additional parameter.

Initialization and Scores

Score and Hessian:

$$\frac{\partial \mathcal{L}_t}{\partial \theta}(\theta) = -\frac{1}{2}\left(\varepsilon_t^2(\theta) - 1\right)\frac{\partial \log \sigma_t^2(\theta)}{\partial \theta}, \quad \varepsilon_t^2(\theta) = \frac{\varepsilon_t^2}{\sigma_t^2(\theta)}.$$

$$\frac{\partial \log \sigma_t^2(\theta)}{\partial \theta} = \frac{1}{\sigma_t^2(\theta)}\frac{\partial \sigma_t^2(\theta)}{\partial \theta}.$$

Partial derivatives:

$$\begin{aligned}
\frac{\partial \sigma_t^2}{\partial \omega} &= 1 + \beta \frac{\partial \sigma_{t-1}^2}{\partial \omega}, \\
\frac{\partial \sigma_t^2}{\partial \beta} &= \sigma_{t-1}^2(\theta) + \beta \frac{\partial \sigma_{t-1}^2}{\partial \beta}, \\
\frac{\partial \sigma_t^2}{\partial \alpha} &= \varepsilon_{t-1}^2 + \beta \frac{\partial \sigma_{t-1}^2}{\partial \alpha}.
\end{aligned}$$

QMLE and Asymptotic Theory

Under strong conditions (e.g. ν_t Gaussian), MLE is:

- consistent,
- asymptotically normal,
- efficient (attains Cramér–Rao bound).

Under weaker assumptions (e.g. correct mean and variance but non-Gaussian ν_t), the **quasi-MLE (QMLE)** is:

- still consistent and asymptotically normal,
- but not (necessarily) efficient w.r.t. the true distribution.

QMLE and Asymptotic Theory

Theorem 2 (Bollerslev and Wooldridge (1992) Bollerslev and Wooldridge 1992)

Under suitable regularity conditions,

$$T^{1/2}(\widehat{\theta} - \theta_0) \Longrightarrow N\big(0,\ \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}\big),$$

where

$$\mathcal{J} = \mathrm{E}\left(\frac{\partial^2 \mathcal{L}_t(\theta_0)}{\partial\theta\partial\theta'}\right), \quad \mathcal{I} = \mathrm{E}\left(\frac{\partial \mathcal{L}_t(\theta_0)}{\partial\theta}\frac{\partial \mathcal{L}_t(\theta_0)}{\partial\theta'}\right).$$

Information Matrices and Robust Variance

If ν_t is i.i.d., then

$$\mathcal{I} = \frac{1}{4}\mathrm{E}\big[(\nu_t^2 - 1)^2\big]\mathrm{E}\left[\frac{\partial \log \sigma_t^2(\theta_0)}{\partial \theta}\frac{\partial \log \sigma_t^2(\theta_0)}{\partial \theta'}\right],$$

$$\mathcal{J} = \frac{1}{2}\mathrm{E}\left[\frac{\partial \log \sigma_t^2(\theta_0)}{\partial \theta}\frac{\partial \log \sigma_t^2(\theta_0)}{\partial \theta'}\right].$$

Thus,

$$\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1} = \mathrm{E}\big[(\nu_t^2 - 1)^2\big]\left[\mathrm{E}\left(\frac{\partial \log \sigma_t^2}{\partial \theta}\frac{\partial \log \sigma_t^2}{\partial \theta'}\right)\right]^{-1}.$$

Information Matrices and Robust Variance

- If $\nu_t \sim N(0,1)$, then $\mathrm{E}[(\nu_t^2 - 1)^2] = 2$.
- If non-Gaussian, the asymptotic variance is multiplied by $\mathrm{E}[(\nu_t^2 - 1)^2]$.
- In general MDS settings, robust "sandwich" estimator

$$\widehat{\mathcal{J}}^{-1}\widehat{\mathcal{I}}\widehat{\mathcal{J}}^{-1}$$

  can be used.

- For linear constraints $R\theta = r$, the Wald statistic

$$W = T(R\widehat{\theta} - r)'[R\widehat{\mathcal{J}}^{-1}\widehat{\mathcal{I}}\widehat{\mathcal{J}}^{-1}R']^{-1}(R\widehat{\theta} - r)$$

is asymptotically χ^2_q.

Table of Contents

GARCH vs Asymmetric GARCH for SPY Returns

The SPDR S&P 500 ETF (SPY) daily log returns for 2024 can be modeled with:

- a symmetric GARCH(1,1),
- asymmetric models: EGARCH(1,1), GJR-GARCH(1,1), TGARCH(1,1),
- and a two-regime Markov-switching mean model.

Key GARCH(1,1) results:

- $\alpha_1 \approx 0.126$, $\beta_1 \approx 0.812$, $\alpha_1 + \beta_1 \approx 0.94$:
  - strongly persistent but mean-reverting volatility.
- Weighted Ljung–Box and ARCH LM tests on squared residuals show remaining dependence at longer lags:
  - symmetric GARCH(1,1) does not fully capture volatility dynamics.

Asymmetric Specifications for SPY

EGARCH(1,1) estimates:

- Leverage parameter $\gamma_1 \approx -0.045$ (negative and significant):
  - negative shocks increase future volatility more than positive ones.
- Information criteria:

$$\text{AIC} \approx -6.75, \quad \text{BIC} \approx -6.71$$

  which improve on symmetric GARCH.
- Diagnostics (Ljung–Box, ARCH LM on squared residuals) show no remaining ARCH effects.

Asymmetric Specifications for SPY

GJR-GARCH and TGARCH:

- GJR-GARCH: $\alpha_1 \approx 0$, $\gamma_1 \approx 0.214$ (significant):
  - volatility reacts mainly to negative innovations.
- TGARCH: both α_1 and threshold parameter $\eta_{11} \approx 1$ significant:
  - strong asymmetric news impact.
- Their AIC/BIC improve on symmetric GARCH but are slightly worse than EGARCH.

Diagnostics and Markov Switching

Diagnostics:

- Nyblom stability test:
  - EGARCH parameters relatively stable (joint statistic below critical values),
  - symmetric GARCH, GJR-GARCH, TGARCH show evidence of instability.
- Adjusted Pearson GOF tests reject exact normality for all models:
  - suggests using heavy-tailed innovations (e.g. Student-t).

Diagnostics and Markov Switching

Markov switching model:

- Two regimes:
  - Regime 1: mean $\approx -0.29\%$ per day, sd $\approx 2.4\%$ (high-volatility, negative-mean regime),
  - Regime 2: mean $\approx 0.13\%$ per day, sd $\approx 0.7\%$ (low-volatility, positive-mean regime).
- Highly persistent regimes:

$$P(S_t = 1 \mid S_{t-1} = 1) \approx 0.813, \quad P(S_t = 2 \mid S_{t-1} = 2) \approx 0.977.$$

- Suggests SPY alternates between extended calm periods and shorter turbulent episodes.

Table of Contents

Testing for Volatility Clustering

Before specifying ARCH/GARCH models, we must test for conditional heteroskedasticity.

Set-up:

$$Y_t = \mu(\mathcal{F}_{t-1}, \theta_0) + \varepsilon_t,$$

with $\varepsilon_t = Y_t - \mu(\mathcal{F}_{t-1}, \theta_0)$.

Hypotheses:

$$H_0 : \mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_0^2 \quad \text{(homoskedastic)},$$

$$H_1 : \mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) \neq \sigma^2 \quad \forall \sigma^2 > 0.$$

Correct specification of $\mu(\mathcal{F}_{t-1}, \theta)$ is essential:

- Misspecified mean $\Rightarrow$ residuals contain omitted dynamics $\Rightarrow$ apparent ARCH even if true innovations are homoskedastic.

Engle's Lagrange Multiplier Test

Given residuals $\widehat{\varepsilon}_t$ from the estimated mean model:

LM test steps:

1. Compute
$$\widehat{\varepsilon}_t = Y_t - \mu(\mathcal{F}_{t-1}, \widehat{\theta}).$$
2. Estimate auxiliary regression (ARCH(p) test):
$$\widehat{\varepsilon}_t^2 = \phi_0 + \sum_{j=1}^{p} \phi_j \widehat{\varepsilon}_{t-j}^2 + u_t.$$
3. Let R^2 be the R^2 from this regression, sample size T:
$$\mathrm{LM} = TR^2 \quad \text{or} \quad (T-p)R^2.$$

Engle's Lagrange Multiplier Test

Under H_0 (no ARCH effects):

$$TR^2 \Longrightarrow \chi^2_p, \quad T \to \infty.$$

- Estimation error in $\widehat{\theta}$ does not affect the asymptotic distribution.

McLeod–Li Portmanteau Test

McLeod–Li statistic:

$$\mathrm{ML}(p) = T \sum_{j=1}^{p} \widehat{\rho}_2^2(j),$$

where $\widehat{\rho}_2(j)$ is the sample autocorrelation of $\widehat{\varepsilon}_t^2$ at lag j.

Under H_0:

$$\mathrm{ML}(p) \Longrightarrow \chi_p^2, \quad T \to \infty.$$

- Asymptotically equivalent to Engle's LM statistic.
- Estimation uncertainty in θ does not affect the limit distribution.
- Intuition: $\widehat{\varepsilon}_t^2 - \varepsilon_t^2 = O_P(T^{-1})$, negligible for large T.

McLeod–Li Portmanteau Test

Remark (lag choice): using many lags p increases power against more persistent ARCH, but can reduce power in finite samples and consume degrees of freedom. It is common to report both low-order LM tests and portmanteau tests for robustness.

Case Study: Annual GDP vs Monthly Industrial Production

Annual U.S. GDP growth (1960–2023):

- Fit AR(1) mean model for annual GDP growth.
- Residuals $\widehat{\varepsilon}_t$ show:

$$\text{LM statistic} \approx 0.06, \quad \text{ML}(p) \approx 5.24, \ p = 4.$$

- 5% critical value of $\chi^2_4 \approx 9.49$:
  - do not reject H_0 of conditional homoskedasticity,
  - little evidence of ARCH effects in annual GDP growth.

Case Study: Annual GDP vs Monthly Industrial Production

Monthly U.S. industrial production growth:

- Fit AR(1) mean model for monthly IP growth.
- With $p = 12$ lags (one year):

$$\text{LM statistic} \approx 5.35, \quad \text{ML}(p) \approx 579.16.$$

- 5% critical value of $\chi^2_{12} \approx 21$:
  - very strong rejection of H_0,
  - pronounced ARCH effects and volatility clustering.

Interpretation of Macro Diagnostic Results

Annual GDP growth:

- Lower frequency, smoother series, shorter sample.
- Little evidence of conditional heteroskedasticity:
  - homoskedastic models may be adequate for volatility at this frequency.

Monthly industrial production:

- Higher frequency, more granular business-cycle and shock information.
- Strong volatility clustering:
  - ARCH/GARCH-type models appropriate to capture time-varying uncertainty.
  - Important implications for forecast intervals and risk assessment around macro aggregates.

Table of Contents

Motivation for Multivariate Volatility Modelling

So far we considered volatility series-by-series. In practice:

- Asset returns move together:
  - market stress events affect many assets simultaneously,
  - portfolios depend on joint dynamics of variances and covariances.
- Portfolio risk, VaR, hedging all depend on the full conditional covariance matrix.

Motivation for Multivariate Volatility Modelling

Multivariate volatility models describe:

$$\mathbf{\Sigma}_t = \mathrm{Var}(\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}),$$

where $\boldsymbol{\varepsilon}_t$ is a vector of innovations.

We will:

- define the multivariate set-up and conditional covariance matrix,
- introduce EWMA and MGARCH models,
- discuss correlation-based and Cholesky-based parameterizations,
- and illustrate with ETF return data.

Basic Setup and Notation

Let $\boldsymbol{r}_t$ be a k-dimensional vector of returns:

$$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t,$$

with

$$\boldsymbol{\mu}_t = \mathrm{E}(\boldsymbol{r}_t \mid \mathcal{F}_{t-1}), \quad \boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{kt})'.$$

Basic Setup and Notation

Conditional covariance matrix:

$$\boldsymbol{\Sigma}_t = \mathrm{Var}(\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}).$$

- symmetric, positive–definite,
- $k(k+1)/2$ distinct elements (variances and covariances).

We often assume:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{z}_t,$$

with $\boldsymbol{z}_t$ i.i.d. mean zero, identity covariance (e.g. multivariate normal or t).

Exponentially Weighted Covariance Matrices

Sample covariance:

$$\widehat{\Sigma} = \frac{1}{T-1} \sum_{t=1}^{T-1} \varepsilon_t \varepsilon_t'$$

assigns equal weight to each past observation.

EWMA covariance: choose $\lambda \in (0,1)$ and define

$$\widehat{\Sigma}_t = \frac{1-\lambda}{1-\lambda^{t-1}} \sum_{j=1}^{t-1} \lambda^{j-1} \varepsilon_{t-j} \varepsilon_{t-j}'.$$

For large t (so $\lambda^{t-1} \approx 0$),

$$\widehat{\Sigma}_t = (1-\lambda)\, \varepsilon_{t-1} \varepsilon_{t-1}' + \lambda\, \widehat{\Sigma}_{t-1}.$$

Exponentially Weighted Covariance Matrices

Properties:

- straightforward to compute,
- positive–definite when initialized suitably,
- all elements share the same decay factor λ,
- multivariate analogue of EWMA/IGARCH for univariate volatility.

EWMA Case Study: SPY, QQQ, EFA

Using daily log returns for SPY, QQQ, and EFA (S&P 500, Nasdaq 100, MSCI EAFE):

EWMA covariance:

$$\widehat{\Sigma}_T = \begin{pmatrix} 0.6809 & 0.9191 & 0.4892 \\ 0.9191 & 1.3184 & 0.6087 \\ 0.4892 & 0.6087 & 0.5578 \end{pmatrix}$$

(in squared percentage returns).

EWMA Case Study: SPY, QQQ, EFA

EWMA correlation:

$$\widehat{R}_T = \begin{pmatrix} 1.0000 & 0.9701 & 0.7938 \\ 0.9701 & 1.0000 & 0.7098 \\ 0.7938 & 0.7098 & 1.0000 \end{pmatrix}.$$

Interpretation:

- SPY and QQQ are almost perfectly correlated (≈ 0.97),
- both are strongly correlated with EFA (0.71–0.79),
- volatilities differ but are all sizable, reflecting global equity risk.

Multivariate GARCH: DVEC and BEKK

Multivariate GARCH extends univariate GARCH recursions to $\boldsymbol{\Sigma}_t$. Directly modeling all elements (VEC model) is often over–parameterized.

Diagonal VEC (DVEC) model:

$$\boldsymbol{\Sigma}_t = \boldsymbol{C} + \sum_{i=1}^{m} \boldsymbol{A}_i \odot \left(\boldsymbol{\varepsilon}_{t-i}\boldsymbol{\varepsilon}_{t-i}'\right) + \sum_{j=1}^{s} \boldsymbol{B}_j \odot \boldsymbol{\Sigma}_{t-j},$$

where $\odot$ is the Hadamard (elementwise) product, and $\boldsymbol{C}, \boldsymbol{A}_i, \boldsymbol{B}_j$ are symmetric.

- Each element of $\Sigma_{ij,t}$ depends on its own past and product of shocks.
- Does not automatically guarantee positive-definite $\boldsymbol{\Sigma}_t$.

Multivariate GARCH: DVEC and BEKK

BEKK(1,1) model:

$$\boldsymbol{\Sigma}_t = \boldsymbol{C}\boldsymbol{C}' + \boldsymbol{A}'\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{\Sigma}_{t-1}\boldsymbol{B},$$

with $\boldsymbol{C}$ lower triangular.

- Ensures positive–definiteness,
- allows rich volatility spillovers via off–diagonals of $\boldsymbol{A}, \boldsymbol{B}$,
- can be parameter-heavy for large k.

BEKK Case Study: SPY and TLT

For a bivariate series of SPY (equity) and TLT (long-term U.S. Treasury bond) demeaned returns, a symmetric BEKK(1,1) fit yields:

$$\boldsymbol{C} = \begin{pmatrix} 0.2029 & 0 \\ -0.1178 & 0.0727 \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} 0.4130 & -0.0334 \\ -0.0099 & 0.2432 \end{pmatrix},$$

$$\boldsymbol{G} = \begin{pmatrix} 0.8961 & 0.0162 \\ 0.0237 & 0.9594 \end{pmatrix}.$$

BEKK Case Study: SPY and TLT

Interpretation:

- $\boldsymbol{CC}'$ governs baseline variances/covariances.
- $\boldsymbol{A}$: strong own-shock effects (diagonal), weak cross-shock effects (off-diagonal).
- $\boldsymbol{G}$: high persistence in both volatility series (diagonals close to 1), small but nonzero cross–effects.
- Volatility is thus very persistent and primarily driven by own past, with modest spillovers.

Correlation-Based Parameterizations

Factor $\boldsymbol{\Sigma}_t$ as:

$$\boldsymbol{\Sigma}_t = \boldsymbol{D}_t \boldsymbol{R}_t \boldsymbol{D}_t,$$

where

$$\boldsymbol{D}_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{kt}), \quad \boldsymbol{R}_t : \text{correlation matrix.}$$

Advantages:

- Separate modelling of marginal volatilities σ_{it}^2 (via univariate GARCH) and correlations $\boldsymbol{R}_t$.
- Reduces dimensionality and improves interpretability:
  - volatilities vs correlation dynamics.

Correlation-Based Parameterizations

Constant-correlation GARCH:

$$\boldsymbol{R}_t = \boldsymbol{R}, \quad \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt} \mid \mathcal{F}_{t-1}) = R_{ij}\sigma_{it}\sigma_{jt}.$$

- Captures time-varying volatilities, but correlations fixed.

Dynamic Conditional Correlation (DCC)

DCC model:

- Standardize innovations:
$$\boldsymbol{z}_t = \boldsymbol{D}_t^{-1}\boldsymbol{\varepsilon}_t.$$
- Recursion for $\boldsymbol{Q}_t$:
$$\boldsymbol{Q}_t = (1-\alpha-\beta)\bar{\boldsymbol{Q}} + \alpha\, \boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}' + \beta\, \boldsymbol{Q}_{t-1},$$
where $\bar{\boldsymbol{Q}}$ is the unconditional covariance of $\boldsymbol{z}_t$.
- Correlations:
$$\boldsymbol{R}_t = \boldsymbol{J}_t^{-1}\boldsymbol{Q}_t\boldsymbol{J}_t^{-1}, \quad \boldsymbol{J}_t = \operatorname{diag}\left(q_{11,t}^{1/2}, \ldots, q_{kk,t}^{1/2}\right).$$

Attributes:

- Just two parameters (α, β) govern correlation dynamics.
- Combined with univariate GARCH for σ_{it}^2, yields flexible MGARCH with modest parameter count.
- Widely used for portfolio risk and contagion analysis.

Cholesky Decomposition and Orthogonal Shocks

Since $\boldsymbol{\Sigma}_t$ is positive–definite, there exists a unique lower–triangular $\boldsymbol{L}_t$ with ones on the diagonal and a diagonal matrix $\boldsymbol{G}_t$ with positive entries such that:

$$\boldsymbol{\Sigma}_t = \boldsymbol{L}_t \boldsymbol{G}_t \boldsymbol{L}_t'.$$

Define

$$\boldsymbol{b}_t = \boldsymbol{L}_t^{-1} \boldsymbol{\varepsilon}_t.$$

Then:

- $\mathrm{Var}(\boldsymbol{b}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{G}_t$ is diagonal,
- $\boldsymbol{b}_t$ are conditionally uncorrelated "orthogonal shocks",
- the off-diagonal elements of $\boldsymbol{L}_t$ encode contemporaneous relationships among shocks.

Cholesky-based parameterizations:

- allow straightforward enforcement of positive–definiteness,
- can be combined with univariate volatility models for the diagonal elements of $\boldsymbol{G}_t$.

Constant- and Dynamic-Correlation GARCH: ETF Case Study

Using daily log returns for SPY, QQQ, and EFA:

Constant-correlation GARCH:

$$\widehat{\Sigma}_T^{\text{const}} = \begin{pmatrix} 1.1026 & 1.2621 & 0.6830 \\ 1.2621 & 1.7442 & 0.7478 \\ 0.6830 & 0.7478 & 0.6448 \end{pmatrix}, \quad \widehat{R}_T^{\text{const}} = \begin{pmatrix} 1.0000 & 0.9101 & 0.8100 \\ 0.9101 & 1.0000 & 0.7052 \\ 0.8100 & 0.7052 & 1.0000 \end{pmatrix}.$$

DCC-GARCH(1,1):

$$\widehat{\Sigma}_T^{\text{DCC}} = \begin{pmatrix} 1.1026 & 1.2881 & 0.6834 \\ 1.2881 & 1.7442 & 0.7606 \\ 0.6834 & 0.7606 & 0.6448 \end{pmatrix}, \quad \widehat{R}_T^{\text{DCC}} = \begin{pmatrix} 1.0000 & 0.9288 & 0.8105 \\ 0.9288 & 1.0000 & 0.7172 \\ 0.8105 & 0.7172 & 1.0000 \end{pmatrix}.$$

Interpretation of Multivariate Case Study

Constant-correlation GARCH:

- Strongly positive correlations, especially between SPY and QQQ (≈ 0.91),
- substantial, though lower, correlations with EFA (≈ 0.71–0.81),
- volatilities differ but are in a similar range.

DCC-GARCH:

- Similar marginal volatilities but slightly higher correlations, especially SPY–QQQ (≈ 0.93),
- allows correlations to adjust over time in response to shocks,
- captures dynamic comovement beyond what constant correlation can.

Interpretation of Multivariate Case Study

Cholesky orthogonal shocks:

- For the last day, transformed shocks b_T might look like:

$$b_T \approx (-1.14,\ -0.02,\ 0.13)'.$$

- The first orthogonal factor dominates (common equity component), others are relatively small.

Table of Contents

Summary of Volatility Models

- **ARCH(p)**:
  - models conditional variance as a function of past squared shocks,
  - captures volatility clustering,
  - requires non-negativity and stationarity conditions for finite variance.
- **GARCH(p, q)**:
  - adds lagged volatility terms,
  - GARCH(1,1) is equivalent to ARCH(∞) with geometric decay,
  - parsimonious and widely used in finance.
- **IGARCH**:
  - unit root in volatility, infinite unconditional variance,
  - past shocks have permanent effects,
  - relates to EWMA volatility (RiskMetrics).
- **Asymmetric models** (EGARCH, GJR, TGARCH, regime-switching GARCH):
  - allow negative and positive shocks to have different impacts,
  - important for modelling leverage effects.

Summary of Estimation, Diagnostics and Multivariate Models

- **Estimation**:
  - MLE and QMLE for GARCH models,
  - asymptotic normality with sandwich (robust) variance,
  - Wald tests under linear constraints.
- **Diagnostics**:
  - Engle's LM and McLeod–Li tests for ARCH effects,
  - importance of correctly specifying the conditional mean,
  - macro case studies highlight frequency-specific volatility behavior.
- **Multivariate volatility**:
  - models the conditional covariance matrix,
  - EWMA, DVEC, BEKK, correlation-based (constant and DCC),
  - Cholesky decomposition and orthogonal shocks,
  - ETF case studies illustrate implementation and interpretation.

Table of Contents

References I

- Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31.3, pp. 307–327.
- Bollerslev, Tim and Jeffrey M Wooldridge (1992). "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances". In: *Econometric reviews* 11.2, pp. 143–172.
- Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica*, pp. 987–1007.
- Glosten, Lawrence R, Ravi Jagannathan, and David E Runkle (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks". In: *The journal of finance* 48.5, pp. 1779–1801.
- Hamilton, James D and Raul Susmel (1994). "Autoregressive conditional heteroskedasticity and changes in regime". In: *Journal of Econometrics* 64.1-2, pp. 307–333.

References II

- Nelson, Daniel B (1990). "ARCH models as diffusion approximations". In: *Journal of Econometrics* 45.1-2, pp. 7–38.
- — (1991). "Conditional heteroskedasticity in asset returns: A new approach". In: *Econometrica* 59.2, pp. 347–370.
- Zakoian, Jean-Michel (1994). "Threshold heteroskedastic models". In: *Journal of Economic Dynamics and Control* 18.5, pp. 931–955.

Chapter 5 — Nonparametric Methods

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Introduction to Nonparametric Methods

Nonparametric methods are a powerful class of statistical techniques that impose *few* assumptions on the underlying distribution.

Key features:

- Do not assume a specific parametric form (e.g. normality).
- Flexible enough to handle:
  - skewness,
  - multimodality,
  - heteroskedasticity,
  - nonlinear relationships.
- Adapt to the *local structure* of the data.

In econometrics:

- Useful when model specification is difficult.
- Robust to misspecification of functional form.
- Capture complex dependencies and nonlinearities often present in economic data.

Scope of this Chapter

This chapter focuses on:

- **Kernel smoothing methods**, including:
  - kernel density estimation,
  - nonparametric regression.
- **Key issues:**
  - bandwidth selection,
  - bias–variance trade-off,
  - choice of kernel and its impact.
- **Applications:**
  - uncovering underlying data structure,
  - econometric examples with real data.

We start with nonparametric density estimators, then extend to multivariate density estimation and discuss implications for econometric modelling.

Table of Contents

Kernel Smoothing for Density Estimation

Kernel smoothing is a widely used nonparametric technique for estimating probability density functions (p.d.f.s).

Advantages:

- Adapts to local features of the data.
- Handles:
  - skewness,
  - multimodality,
  - heteroskedasticity.
- Provides a smoothed estimate of the entire distribution.

Basic idea: Each observation X_t contributes to the estimate via a *kernel* function $K(\cdot)$ centered at X_t, with a smoothing parameter (bandwidth) h controlling how widely this contribution is spread.

Kernels and Bandwidth

Kernel function $K(\cdot)$:

- Assigns weights to data points in a neighborhood of x.
- Common choices:
  - Gaussian kernel,
  - Epanechnikov kernel,
  - Uniform, Quartic, etc.
- Typically:
  - nonnegative,
  - integrates to 1,
  - symmetric about 0.

Kernels and Bandwidth

Bandwidth h:

- Controls degree of smoothing:
  - small h: captures fine detail but high variance (undersmoothing),
  - large h: smoother but higher bias (oversmoothing).
- Selected via:
  - rules of thumb (e.g. Silverman's rule),
  - cross-validation,
  - plug-in methods.

Second-Order and Higher-Order Kernels

Definition (Second-order kernel)

A second-order (regular) kernel $K(\cdot)$ is a symmetric p.d.f. satisfying:

- $\int_{-\infty}^{\infty} K(u)\,du = 1;$
- $\int_{-\infty}^{\infty} uK(u)\,du = 0;$
- $\int_{-\infty}^{\infty} u^2K(u)\,du = C_K < \infty;$
- $\int_{-\infty}^{\infty} K^2(u)\,du = D_K < \infty.$

Second-Order and Higher-Order Kernels

Definition (q-th order kernel)

A kernel $K(\cdot)$ is of order q if:

- $\int K(u)\,du = 1$;
- $\int u^j K(u)\,du = 0$ for $1 \leq j \leq q-1$;
- $\int u^q K(u)\,du < \infty$;
- $\int K^2(u)\,du < \infty$.

Higher-Order Kernels and Bias

Higher-order kernels ($q \geq 2$):

- Reduce bias by matching higher-order moments.
- Better capture higher-order features of $g(x)$ (e.g. curvature).
- Useful when density has complex local structures (e.g. sudden market shifts).

Trade-off:

- As q increases:
  - bias decreases,
  - variance can increase, especially in small samples.
- Higher-order kernels often require larger samples for stable estimation.

Common Second-Order Kernels

Some widely used kernels:

- **Uniform kernel**

$$K(u) = \frac{1}{2}\,\mathbf{1}(|u| \leq 1).$$

- **Gaussian kernel**

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad -\infty < u < \infty.$$

- **Epanechnikov kernel**

$$K(u) = \frac{3}{4}(1-u^2)\,\mathbf{1}(|u| \leq 1).$$

- **Quartic kernel**

$$K(u) = \frac{15}{16}(1-u^2)^2\,\mathbf{1}(|u| \leq 1).$$

Kernel Functions: Visualization

Only the Gaussian kernel has unbounded support; the others are compactly supported on $[-1, 1]$. The Epanechnikov kernel minimizes MSE for a given bandwidth.

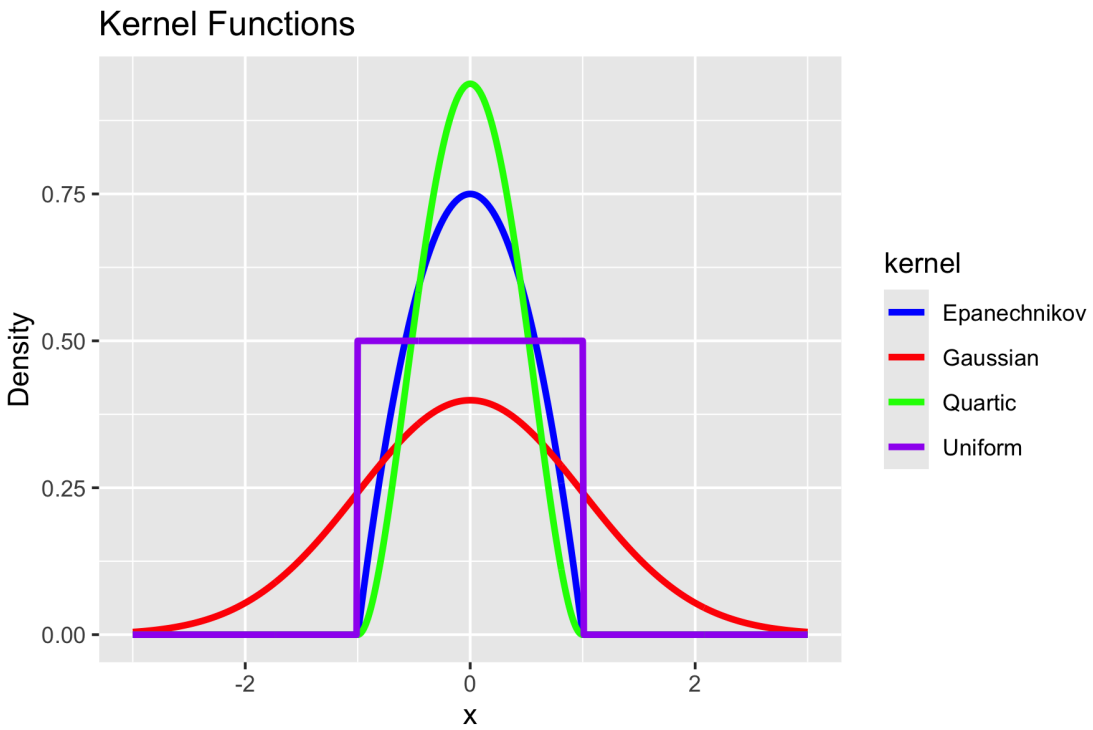


Figure 1: Uniform, Gaussian, Epanechnikov, and Quartic kernel functions

Kernel Functions: Visualization

Comments:

- Uniform kernel: flat weights within its support.
- Other kernels: bell-shaped, decreasing weights with distance from zero.
- In practice, estimation accuracy is much more sensitive to *bandwidth* than to kernel choice.

Example: Histogram as Kernel Estimator

Example (Histogram)

If $K(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1)$, then

$$\widehat{g}(x) = \frac{1}{2hT} \sum_{t=1}^{T} \mathbf{1}(|x - X_t| \leq h).$$

Example: Histogram as Kernel Estimator

Interpretation:

- Numerator counts number of observations in $[x-h, x+h]$.
- Denominator $2hT$ approximates total mass in that interval.
- As $T \to \infty$ and $h \to 0$,

$$T^{-1} \sum_{t=1}^{T} \mathbf{1}(|x - X_t| \leq h) \to \mathbb{P}(x - h \leq X_t \leq x + h) \approx 2hg(x).$$

- Histogram is a special case of kernel density estimator with uniform kernel and bin width $2h$.

FTSE 100: Histogram and Kernel Density

We can visualize FTSE 100 daily log returns with a histogram and an overlaid kernel density:

- Download daily FTSE prices from Yahoo Finance (e.g. from 2015).
- Compute daily log returns (in %).
- Plot histogram of returns and overlay kernel density estimate.

FTSE 100: Histogram and Kernel Density

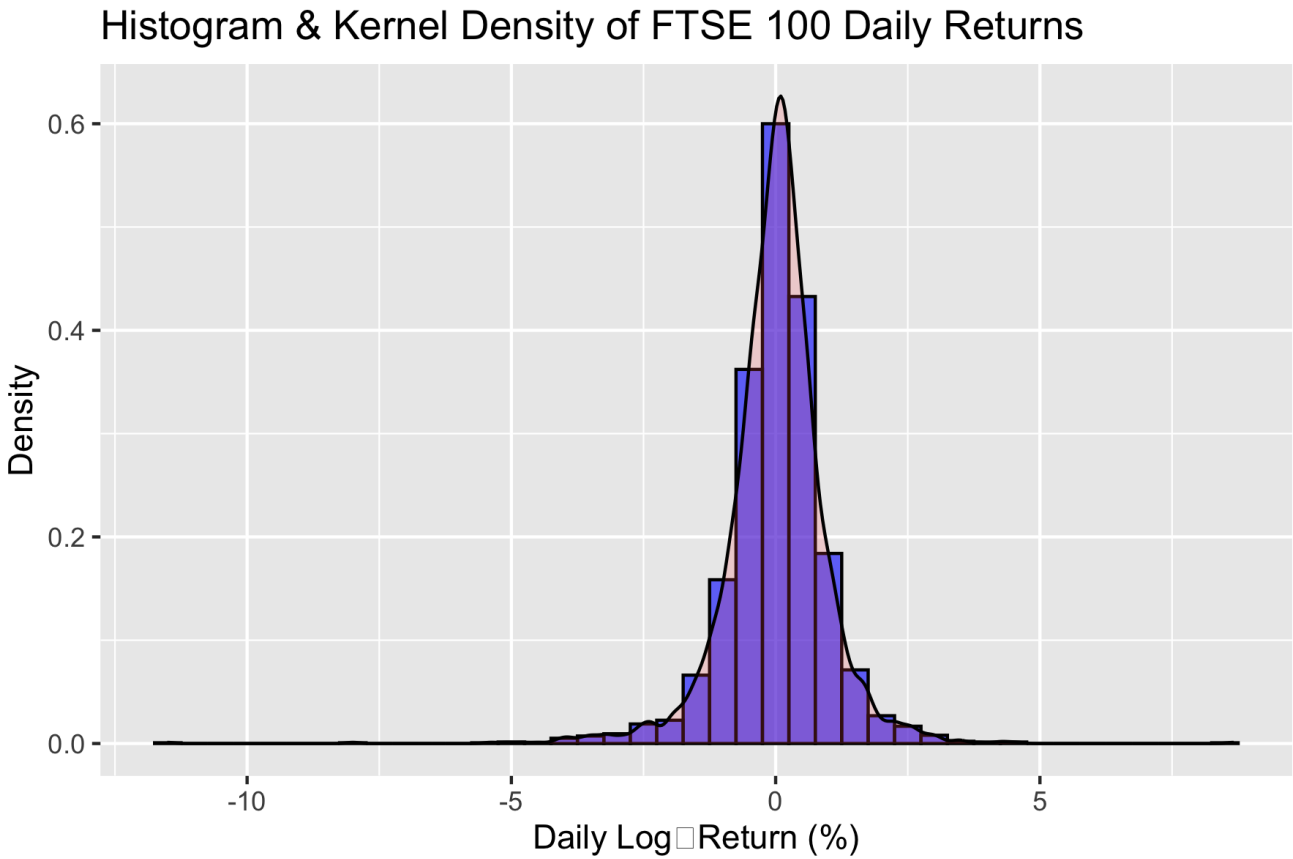


Figure 2: Probability density and histogram of FTSE 100 index returns

Interpretation: FTSE 100 Returns

From Figure 2:

- The return distribution appears **asymmetric**:
  - often negatively skewed due to sharp market declines.
- Tails are **heavier** than those of a Gaussian:
  - higher probability of extreme losses or gains,
  - consistent with large market moves and "black swan" events.
- Kernel density estimation helps reveal these non-Gaussian features, which parametric normal assumptions would obscure.

Smoothness Assumptions on the Density

To analyze bias and variance, we impose regularity on the p.d.f. $g(x)$.

Assumption A.1 (Smoothness of g)

1. $\{X_t\}$ *is strictly stationary with marginal density* $g(x)$.
2. $g(x)$ *is twice continuously differentiable on a bounded support* $[a, b]$.
3. $g''(\cdot)$ *is Lipschitz on* $[a, b]$*: there exists* $C < \infty$ *such that*

$$|g''(x_1) - g''(x_2)| \leq C|x_1 - x_2|, \quad \forall x_1, x_2 \in [a, b].$$

Derivatives at boundaries are defined via one-sided limits, e.g.

$$g'(a) = \lim_{x \to 0^+} \frac{g(a+x) - g(a)}{x}, \quad g'(b) = \lim_{x \to 0^-} \frac{g(b+x) - g(b)}{x}.$$

Lipschitz Continuity and Kernel Support

Definition (Lipschitz continuity)

A function f is Lipschitz if there exists $C \geq 0$ such that

$$|f(x) - f(y)| \leq C|x - y|, \quad \forall x, y \text{ in its domain.}$$

Lipschitz continuity:

- Stronger than continuity.
- Binds the function's slope and forbids "vertical" jumps.
- Crucial for controlling remainder terms in Taylor expansions under integration.

Lipschitz Continuity and Kernel Support

Assumption A.2 (Second-order kernel with bounded support)

$K(u)$ is a positive second-order kernel with support on $[-1, 1]$.

Bounded support simplifies derivations and is common in practice; asymptotics can also be developed for unbounded kernels (e.g. Gaussian).

Error Decomposition and MSE

For interior $x \in [a+h, b-h]$ and kernel estimator $\widehat{g}(x)$:

$$\widehat{g}(x) - g(x) = \underbrace{\mathrm{E}\widehat{g}(x) - g(x)}_{\text{bias}} + \underbrace{\widehat{g}(x) - \mathrm{E}\widehat{g}(x)}_{\text{variance term}}.$$

Thus the mean squared error (MSE) is

$$\mathrm{MSE}[\widehat{g}(x)] = \big(\mathrm{E}\widehat{g}(x) - g(x)\big)^2 + \mathrm{Var}[\widehat{g}(x)] = \mathrm{Bias}^2 + \mathrm{Var}[\widehat{g}(x)].$$

If $\widehat{g}(x)$ is consistent, then both bias and variance $\to 0$ as $T \to \infty$ under appropriate conditions. We now derive their leading orders.

Bias of the Kernel Density Estimator (1)

For an interior point $x \in [a+h, b-h]$:

$$\begin{aligned}
\mathrm{E}[\widehat{g}(x)] - g(x) &= \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[K_h(x - X_t)] - g(x) \\
&= \mathrm{E}[K_h(x - X_t)] - g(x) \quad \text{(i.d.)} \\
&= \int_a^b \frac{1}{h} K\Big(\frac{x-y}{h}\Big) g(y)\, dy - g(x).
\end{aligned}$$

Change variable $u = (y - x)/h$:

$$\begin{aligned}
\mathrm{E}[\widehat{g}(x)] - g(x) &= \int_{(a-x)/h}^{(b-x)/h} K(u)\, g(x + hu)\, du - g(x) \\
&= \int_{-1}^{1} K(u)\, g(x + hu)\, du - g(x),
\end{aligned}$$

using the bounded support $[-1, 1]$.

Bias of the Kernel Density Estimator (2)

Expand $g(x+hu)$ around x:

$$g(x+hu) = g(x) + hug'(x) + \frac{1}{2}h^2u^2g''(x+\lambda hu),$$

for some $\lambda \in (0,1)$.
Using second-order kernel properties:

$$\int K(u)\,du = 1, \quad \int uK(u)\,du = 0, \quad \int u^2K(u)\,du = C_K,$$

we get:

$$\begin{aligned}\mathrm{E}[\widehat{g}(x)] - g(x) &= \int_{-1}^{1} K(u)\big[g(x+hu) - g(x)\big]\,du \\ &= \frac{1}{2}h^2\int_{-1}^{1} u^2K(u)g''(x+\lambda hu)\,du.\end{aligned}$$

Bias of the Kernel Density Estimator (2)

Split this as

$$\frac{1}{2}h^2 C_K g''(x) + \frac{1}{2}h^2 \int_{-1}^{1} [g''(x + \lambda h u) - g''(x)] u^2 K(u)\, du.$$

The second term is $o(h^2)$ by dominated convergence and Lipschitz continuity of g''.

Conclusion:

$$\mathrm{E}[\widehat{g}(x)] - g(x) = \frac{1}{2}h^2 C_K g''(x) + o(h^2).$$

Lebesgue Dominated Convergence: Intuition

Theorem (Lebesgue Dominated Convergence Theorem (LDCT).)

If $f_n(x) \to f(x)$ pointwise a.e., and $|f_n(x)| \leq g(x)$ for all n with g integrable, then

$$\lim_{n\to\infty} \int f_n(x)\,dx = \int f(x)\,dx.$$

In our bias derivation, the relevant term is

$$\left[g''(x+\lambda hu) - g''(x)\right] u^2 K(u).$$

Lebesgue Dominated Convergence: Intuition

Why LDCT applies:

- Lipschitz continuity of $g'' \Rightarrow g''(x+\lambda hu) - g''(x) \to 0$ as $h \to 0$.
- $u^2 K(u)$ is integrable and provides a dominating function.

Thus we can interchange limit and integration:

$$\int_{-1}^{1} [g''(x+\lambda hu) - g''(x)] u^2 K(u)\, du \to 0 \quad (h \to 0).$$

Bias Summary and Consistency

For interior x,

$$\text{Bias}[\widehat{g}(x)] = \mathrm{E}[\widehat{g}(x)] - g(x) = \frac{1}{2}h^2 C_K g''(x) + o(h^2).$$

Implications:

- Bias is of order h^2.
- To have bias $\to 0$, we must take $h \to 0$ as $T \to \infty$.

The same leading bias result holds for weakly dependent data (e.g. α-mixing processes) under suitable conditions. We used the i.i.d. assumption for simplicity but it can be relaxed.
Boundary regions $[a, a+h]$ and $[b-h, b]$ require special treatment due to lack of symmetry and can exhibit larger bias ("boundary problem").

Variance of the Kernel Density Estimator

Assume i.i.d. sample $\{X_t\}_{t=1}^T$.

Assumption A.3 (i.i.d. observations)

$\{X_t\}_{t=1}^T$ are i.i.d. with density g.

Define

$$Z_t(x) = K_h(x - X_t) - \mathrm{E}[K_h(x - X_t)],$$

so $\mathrm{E}[Z_t] = 0$ and $\{Z_t\}$ are i.i.d.
Then

$$\mathrm{Var}[\widehat{g}(x)] = \mathrm{E}\big[\widehat{g}(x) - \mathrm{E}\widehat{g}(x)\big]^2 = \frac{1}{T^2}\sum_{t=1}^{T}\mathrm{Var}(Z_t)$$

$$= \frac{1}{T}\mathrm{Var}(Z_1).$$

Variance Derivation and Order

Compute

$$\mathrm{Var}(Z_1) = \mathrm{E}[K_h^2(x - X_1)] - \big(\mathrm{E}[K_h(x - X_1)]\big)^2.$$

$$\mathrm{E}[K_h^2(x - X_1)] = \int_a^b \frac{1}{h^2} K^2\left(\frac{x - y}{h}\right) g(y)\, dy.$$

Using $u = (x - y)/h$,

$$\mathrm{E}[K_h^2(x - X_1)] = \frac{1}{h}\int_{-1}^{1} K^2(u) g(x - hu)\, du = \frac{1}{h} g(x) D_K + o(h^{-1}),$$

where $D_K = \int K^2(u)\, du$.
The squared mean term is $O(1)$, so

$$\mathrm{Var}[\widehat{g}(x)] = \frac{1}{Th} g(x) D_K + o\big((Th)^{-1}\big).$$

Order: variance is $O((Th)^{-1})$ where Th approximates the effective number of observations in $[x - h, x + h]$.

Dependence and Variance

Even if $\{X_t\}$ are dependent (e.g. α-mixing), under appropriate conditions the asymptotic variance of $\widehat{g}(x)$ remains the same as in the i.i.d. case.

Intuition (Hart, 1996):

- The estimator at x uses points in the local window $[x-h, x+h]$.
- These observations are often far apart in time, hence their dependence weakens.
- Locally, the sample behaves nearly i.i.d., so asymptotic variance is unaffected.

Thus, serial dependence under standard mixing conditions has little impact on the leading variance term of kernel density estimators.

MSE and Optimal Bandwidth

$$\begin{aligned}\mathrm{MSE}[\widehat{g}(x)] &= \mathrm{Var}[\widehat{g}(x)] + \mathrm{Bias}^2[\widehat{g}(x)] \\ &= \frac{1}{Th} g(x) D_K + \frac{1}{4} h^4 [g''(x)]^2 C_K^2 + o\big(T^{-1}h^{-1} + h^4\big).\end{aligned}$$

So

$$\mathrm{MSE}[\widehat{g}(x)] = O\big(T^{-1}h^{-1} + h^4\big).$$

By Chebyshev's inequality:

$$\widehat{g}(x) - g(x) = O_P\big(T^{-1/2}h^{-1/2} + h^2\big).$$

MSE and Optimal Bandwidth

Consistency conditions:

- $h \to 0$,
- $Th \to \infty$,

so both bias and variance vanish as $T \to \infty$.

Relative MSE and Sparse Regions

For $g(x) > 0$, define relative MSE:

$$\begin{aligned}\text{MSE}\left[\frac{\widehat{g}(x)}{g(x)}\right] &= \frac{\text{MSE}[\widehat{g}(x)]}{g^2(x)} \\ &= \frac{1}{Thg(x)} D_K + \frac{1}{4} h^4 \left(\frac{g''(x)}{g(x)}\right)^2 C_K^2 + o(T^{-1}h^{-1} + h^4).\end{aligned}$$

Implications:

- When $g(x)$ is small (sparse data), relative variance is large:
  - density estimates less reliable in the tails.
- When $g''(x)$ is large (rapid curvature), bias is large:
  - local kernel windows may not capture sharp features accurately.

Relative MSE and Sparse Regions

Optimal bandwidth:

$$h_0 = \left[\frac{D_K}{C_K^2} \cdot \frac{1/g(x)}{\left(g''(x)/g(x)\right)^2} \right]^{1/8} T^{-1/5}.$$

With h_0, the optimal rate is

$$\widehat{g}(x) - g(x) = O_P(T^{-2/5}),$$

slower than the parametric $T^{-1/2}$.

Multivariate Kernel Density Estimator

Consider d-dimensional $X_t = (X_{1t}, \ldots, X_{dt})'$ with density $f(x)$.

Estimator:

$$\widehat{f}(x) = \frac{1}{T}\sum_{t=1}^{T}\prod_{i=1}^{d} K_h(x_i - X_{it}) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{K}_h(x - X_t),$$

where $\mathcal{K}_h$ is the product kernel.
Assume x is interior: $x_i \in [a_i + h, b_i - h]$.

Multivariate Kernel Density Estimator

Bias:

$$\begin{aligned}\mathrm{E}[\widehat{f}(x)] - f(x) &= \int \prod_{i=1}^{d} \frac{1}{h} K\left(\frac{x_i - y_i}{h}\right) f(y)\, dy - f(x) \\ &= \int_{-1}^{1} \cdots \int_{-1}^{1} \prod_{i=1}^{d} K(u_i) f(x + hu)\, du - f(x).\end{aligned}$$

A Taylor expansion yields:

$$\mathrm{E}[\widehat{f}(x)] - f(x) = \frac{1}{2} h^2 C_K \sum_{i=1}^{d} \frac{\partial^2 f(x)}{\partial x_i^2} + o(h^2) = O(h^2).$$

Variance and MSE in d Dimensions

Define centered kernel:

$$Z_t(x) = \mathcal{K}_h(x - X_t) - \mathrm{E}[\mathcal{K}_h(x - X_t)].$$

If $\{X_t\}$ is i.i.d., then $\{Z_t(x)\}$ is i.i.d. with mean zero.

$$\mathrm{Var}[\widehat{f}(x)] = \mathrm{E}[\widehat{f}(x) - \mathrm{E}\widehat{f}(x)]^2 = \frac{1}{T}\mathrm{Var}[Z_t(x)] = \frac{1}{Th^d} f(x) D_K^d + o(T^{-1}h^{-d}),$$

where $D_K = \int K^2(u)\, du$.

MSE:

$$\mathrm{MSE}[\widehat{f}(x)] = \frac{1}{Th^d} f(x) D_K^d + \frac{1}{4} C_K^2 h^4 \left(\sum_{i=1}^d f_{ii}(x) \right)^2 + o(T^{-1}h^{-d} + h^4).$$

$$\mathrm{MSE} = O(T^{-1}h^{-d} + h^4).$$

Optimal Bandwidth and Curse of Dimensionality

Optimal bandwidth:

$$h_0 = \left[\frac{dD_K^2}{C_K^2} \cdot \frac{1/f(x)}{\left(\sum_{i=1}^{d} f_{ii}(x)/f(x) \right)^2} \right]^{1/(d+4)} T^{-1/(d+4)}.$$

With h_0,

$$\text{MSE}[\widehat{f}(x)] \propto T^{-4/(4+d)}.$$

Convergence rates:

- $d = 1$: $T^{-4/6}$,
- $d = 2$: $T^{-4/6} = T^{-2/3}$,
- $d = 3$: $T^{-4/7}$,
- As d increases, $T^{-4/(4+d)}$ gets slower.

Optimal Bandwidth and Curse of Dimensionality

This is the **curse of dimensionality**:

- For large d, very large T is needed for accurate estimation.
- In practice, nonparametric density estimation is rarely used beyond $d \approx 5$ in economics and finance.

Mitigating the Curse of Dimensionality

Common strategies:

- **Independence:**

$$f(x) = \prod_{i=1}^{d} g_i(x_i)$$

  if $X_1, \ldots, X_d$ are independent.

- **Markov assumption:**

$$f(X_t \mid \mathcal{F}_{t-1}) = f(X_t \mid X_{t-1}) = \frac{f(X_t, X_{t-1})}{g(X_{t-1})},$$

  which reduces effective dimensionality to 2.

- **Projection pursuit:** assume $f(x)$ depends on low-dimensional linear combinations of x, estimate these projections and then apply 1D or low-D nonparametric methods.

Nonparametric methods are often combined with dimension-reduction devices to remain feasible in empirical applications.

Applications in Financial Econometrics

Kernel-based density estimation is widely used:

- Ait-Sahalia (1996): marginal kernel density estimators $\widehat{g}(x)$ to test diffusion models for short-term interest rates.
- Hong and Li (2005): nonparametric joint density estimators $\widehat{f_j}(x, y)$ to test continuous-time models, including affine term structure models.

Applications in Financial Econometrics

Affine term structure models:

- Model the short rate as

$$r(t) = a + b'X(t),$$

  where $X(t)$ is a vector of factors following stochastic differential equations.
- Called "affine" because yields and bond prices are linear (plus constant) in $X(t)$.
- Provide tractable closed-form solutions for bond prices and derivatives, widely used in interest-rate modelling and risk management.

Multivariate KDE: FTSE and DAX Returns

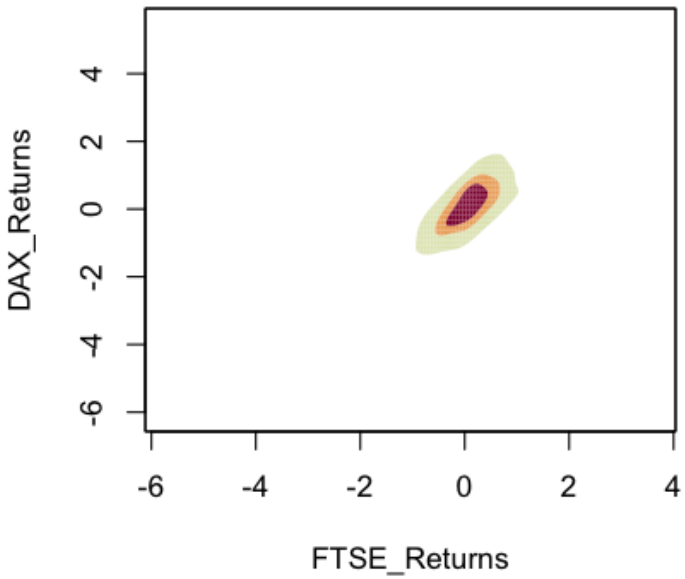
We can examine the joint distribution of FTSE 100 and DAX daily log returns:

Steps:

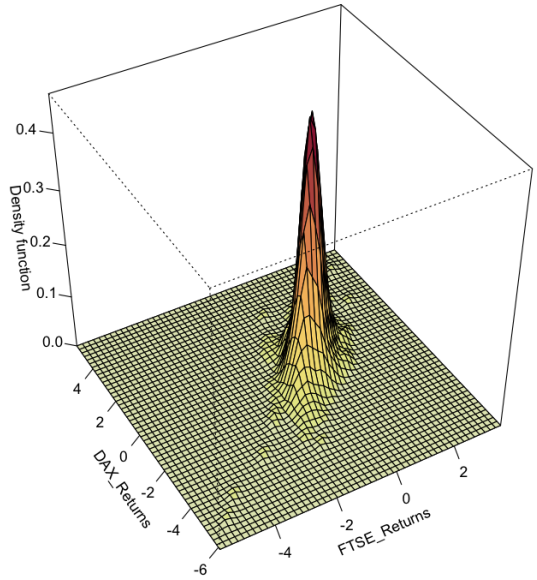
- Download FTSE and DAX daily prices (e.g. from Yahoo).
- Compute daily log returns (in %).
- Form a bivariate sample $(\text{FTSE}_t, \text{DAX}_t)$.
- Use a multivariate kernel estimator (e.g. with the `ks` package) to estimate the joint density $f(x, y)$.
- Visualize via:
  - 2D filled contour plot,
  - 3D perspective plot.

Note: Base R plotting functions are adequate for 2D/3D density plots and are simpler for static PNG export than some `ggplot2` 3D options.

FTSE and DAX: 2D and 3D Kernel Density



(a) 2D density (filled contour)



(b) 3D density (perspective)

Figure 3: Kernel density estimation of FTSE and DAX daily returns

FTSE and DAX: 2D and 3D Kernel Density

These plots provide insights into:

- the joint shape of FTSE and DAX return distributions,
- dependence structure (e.g. concentration along diagonal),
- tail behavior and potential nonlinear co-movement.

Table of Contents

Why Nonparametric Regression?

Compared to traditional parametric regression, nonparametric regression:

Advantages:

- **No pre-specified functional form**:
  - Does not assume linearity or a fixed polynomial degree.
  - Can capture complex, unknown patterns.
- **Reduced specification bias**:
  - Fewer structural assumptions.
  - Robust to nonlinearity, heterogeneity, complex dependencies.
- **Data-driven**:
  - Driven by local averaging or local polynomial fitting.
  - Particularly useful for exploratory data analysis.

We will discuss:

- Nadaraya–Watson estimator (local constant),
- Local polynomial estimators (local linear, quadratic, etc.).

Illustrative Example: Linear vs Nonparametric Fit

We simulate nonlinear data:

$$y = \sin(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Two models:

- Linear regression: $y = \beta_0 + \beta_1 x + u$.
- Nonparametric regression: LOESS (locally weighted polynomial, degree 2).

Findings (Figure 4):

- Linear fit (blue) misses the nonlinear $\sin(x)$ shape.
- LOESS fit (red) tracks the true nonlinear pattern much better.
- Linear residuals show systematic patterns.
- LOESS residuals are more randomly scattered.

Linear vs Nonparametric Regression: Plots

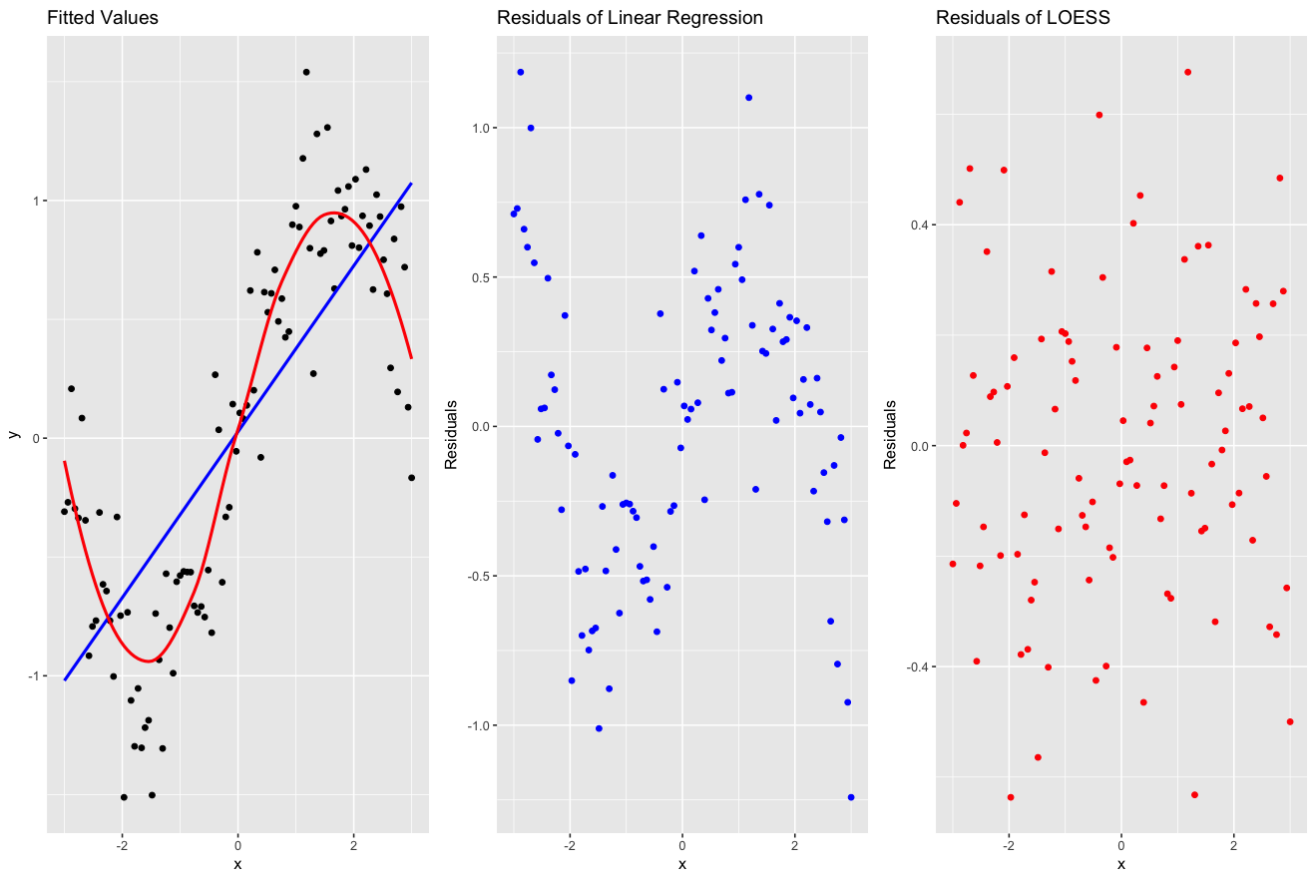


Figure 4: Fitted values and residuals for linear vs nonparametric regression

Linear vs Nonparametric Regression: Plots

Conclusion:

- Nonparametric regression adapts to local structure and captures nonlinear relationships.
- Parametric linear regression fails when the true relationship is nonlinear.

Nadaraya–Watson Estimator: Definition

Consider the regression model

$$Y_t = r(X_t) + \varepsilon_t, \quad \mathrm{E}(\varepsilon_t \mid X_t) = 0.$$

The Nadaraya–Watson estimator of $r(x)$ is

$$\widehat{r}(x) = \frac{\widehat{m}(x)}{\widehat{g}(x)},$$

where

$$\widehat{m}(x) = \frac{1}{T}\sum_{t=1}^{T} Y_t K_h(x - X_t), \qquad \widehat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t).$$

Nadaraya–Watson Estimator: Definition

Define weights

$$\widehat{W}_t(x) = \frac{K_h(x - X_t)}{\sum_{s=1}^{T} K_h(x - X_s)}, \quad \sum_{t=1}^{T} \widehat{W}_t(x) = 1.$$

Then

$$\widehat{r}(x) = \sum_{t=1}^{T} \widehat{W}_t(x)\, Y_t,$$

a locally weighted average of $\{Y_t\}$ near x.

Nadaraya–Watson Estimator and Regressograms

If the kernel is uniform,

$$K(u) = \tfrac{1}{2}\mathbf{1}(|u| \le 1),$$

then the estimator becomes

$$\widehat{r}(x) = \frac{\sum_{t=1}^{T} Y_t \mathbf{1}(|X_t - x| \le h)}{\sum_{t=1}^{T} \mathbf{1}(|X_t - x| \le h)},$$

the local sample mean in $[x-h, x+h]$.

This is called a **regressogram** (Tukey, 1961).

Interpretation:

- For each x, collect observations with X_t near x.
- Average the corresponding Y_t's with weights determined by the kernel.
- Provides a smooth estimate of $r(x)$ without specifying a parametric form.

Example: Horsepower and Fuel Efficiency

Data: `mtcars` dataset.

- x = horsepower (`hp`).
- y = miles per gallon (`mpg`).

We apply the Nadaraya–Watson estimator:

- Use a data-driven bandwidth selector (e.g. `dpill`).
- Use kernel regression (`locpoly` with degree 0).
- Plot fitted curve vs. scatter of (`hp`, `mpg`).

Figure 5 shows a smooth, nonlinear relationship:

- Fuel efficiency declines nonlinearly with horsepower.
- Nadaraya–Watson fit captures this pattern without specifying a functional form.

Nadaraya–Watson: Horsepower vs MPG

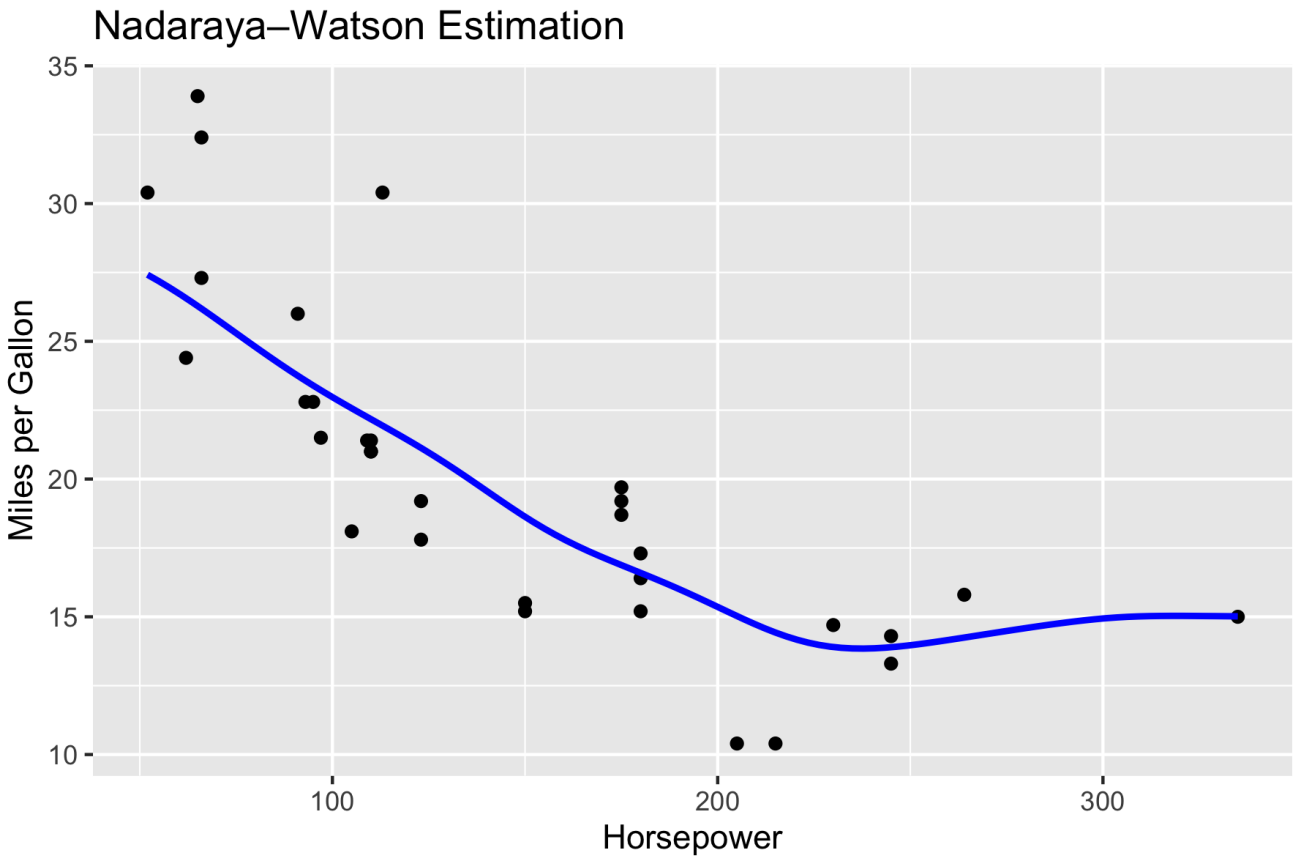


Figure 5: Effect of horsepower on MPG using the Nadaraya–Watson estimator

Nadaraya–Watson: Horsepower vs MPG

Key points:

- Black points: actual data.
- Blue line: nonparametric estimate of $r(x) = \mathrm{E}(\mathrm{mpg} \mid \mathrm{hp} = x)$.
- Relationship is clearly nonlinear and decreasing.

MSE and Optimal Bandwidth for Nadaraya–Watson

Recall the decomposition:

$$\widehat{r}(x) - r(x) = \frac{\widehat{m}(x) - r(x)\widehat{g}(x)}{\widehat{g}(x)}.$$

Approximate:

$$\widehat{r}(x) - r(x) \approx \frac{\widehat{m}(x) - r(x)\widehat{g}(x)}{\mathrm{E}\widehat{g}(x)},$$

since $\widehat{g}(x) \to \mathrm{E}\widehat{g}(x)$ in probability.
Write

$$\widehat{m}(x) - r(x)\widehat{g}(x) = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t K_h(x - X_t)}_{\widehat{V}(x)} + \underbrace{\frac{1}{T}\sum_{t=1}^{T} [r(X_t) - r(x)] K_h(x - X_t)}_{\widehat{B}(x)}.$$

MSE and Optimal Bandwidth for Nadaraya–Watson

Variance part: $\widehat{V}(x)$.
Bias part: $\widehat{B}(x)$.
Under i.i.d. assumptions:

$$\mathrm{E}[\widehat{V}(x)^2] = \frac{1}{Th}\,\sigma^2(x)g(x)\int K^2(u)\,du\,[1+o(1)],$$

where $\sigma^2(x) = \mathrm{Var}(\varepsilon_t \mid X_t = x)$.

Bias of Nadaraya–Watson and Bandwidth Choice

Define $m(x) = r(x)g(x)$ and consider

$$\mathrm{E}[\widehat{B}(x)] = \mathrm{E}\big[r(X_t)K_h(x - X_t)\big] - r(x)\mathrm{E}[K_h(x - X_t)].$$

Using Taylor expansions (and symmetry of K):

$$\mathrm{E}[\widehat{B}(x)] = \frac{1}{2}h^2\Big[r''(x)g(x) + 2r'(x)g'(x)\Big]C_K + o(h^2).$$

Normalizing by $\mathrm{E}[\widehat{g}(x)] \to g(x)$:

$$\mathrm{E}\left[\frac{\widehat{B}(x)}{\mathrm{E}\widehat{g}(x)}\right] = \frac{1}{2}h^2\left[r''(x) + \frac{2r'(x)g'(x)}{g(x)}\right]C_K + o(h^2).$$

Summary:

- As $h \to 0$, bias is $O(h^2)$ in the interior.
- Variance is $O\big((Th)^{-1}\big)$.
- Optimal bandwidth again balances bias and variance:

$$h_0 \propto T^{-1/5}, \quad \widehat{r}(x) - r(x) = O_P(T^{-2/5}).$$

Boundary Issues for Nadaraya–Watson

Near the boundary of support (e.g. $x \approx a$ or b):

- Kernel support extends beyond data range.
- Fewer observations available on one side of x.
- Local average is biased downward or upward depending on shape.

Consequences:

- Interior bias: $O(h^2)$.
- Boundary bias: typically $O(h)$ (larger).
- Estimates near boundaries are less reliable.

Possible remedies:

- Boundary-corrected kernels or reflection methods.
- Exclude boundary regions from analysis.
- Use local polynomial estimators (better boundary behavior).

Table of Contents

Local Polynomial Regression: Motivation

Local polynomial estimators generalize Nadaraya–Watson by fitting a local polynomial instead of a local constant.

Advantages:

- Better boundary behavior (reduces boundary bias).
- Can estimate derivatives of $r(x)$.
- Provides a flexible way to control bias–variance via polynomial degree p.

Key idea:

- Approximate $r(z)$ near x by a Taylor polynomial:

$$r(z) \approx \sum_{j=0}^{p} \alpha_j (z - x)^j,$$

and estimate $\alpha_j(x)$ by weighted least squares using observations with X_t near x.

Nadaraya–Watson as Local Constant LS

Recall the global constant fit:

$$\min_r \sum_{t=1}^{T} (Y_t - r)^2 \quad \Rightarrow \quad \widehat{r} = \bar{Y}.$$

Local constant fit at x:

$$\min_r \sum_{t=1}^{T} (Y_t - r)^2 K_h(x - X_t).$$

FOC:

$$\sum_{t=1}^{T} (Y_t - \widehat{r}(x)) K_h(x - X_t) = 0 \quad \Rightarrow \quad \widehat{r}(x) = \frac{\sum Y_t K_h(x - X_t)}{\sum K_h(x - X_t)},$$

the Nadaraya–Watson estimator.

Nadaraya–Watson as Local Constant LS

Interpretation:

- NW is local *constant* least squares with kernel weights.
- Local polynomial regression replaces constant r with a polynomial in $(X_t - x)$.

Local Polynomial Approximation

Suppose $r(z)$ has $(p+1)$ continuous derivatives near x:

$$r(z) = \sum_{j=0}^{p} \frac{1}{j!} r^{(j)}(x)(z-x)^j + \frac{1}{(p+1)!} r^{(p+1)}(\bar{x})(z-x)^{p+1}.$$

Define $\alpha_j(x) = \frac{1}{j!} r^{(j)}(x)$, so

$$r(z) \approx \sum_{j=0}^{p} \alpha_j(x)(z-x)^j.$$

Local Polynomial Approximation

Local polynomial fitting:

$$\min_{\alpha} \sum_{t=1}^{T} \left[Y_t - \sum_{j=0}^{p} \alpha_j (X_t - x)^j \right]^2 K_h(x - X_t).$$

Estimator:

- $\widehat{r}(x) = \widehat{\alpha}_0$.
- $\nu! \widehat{\alpha}_\nu$ estimates $r^{(\nu)}(x)$.

Local Polynomial vs Nadaraya–Watson

Figure 6 illustrates:

- Nadaraya–Watson estimator (local constant).
- Local linear estimator (local polynomial of order $p = 1$).

Observation:

- Local linear fitting yields smaller bias, particularly near boundaries.
- Nadaraya–Watson is a special case of local polynomial regression ($p = 0$).

Local Polynomial vs Nadaraya–Watson

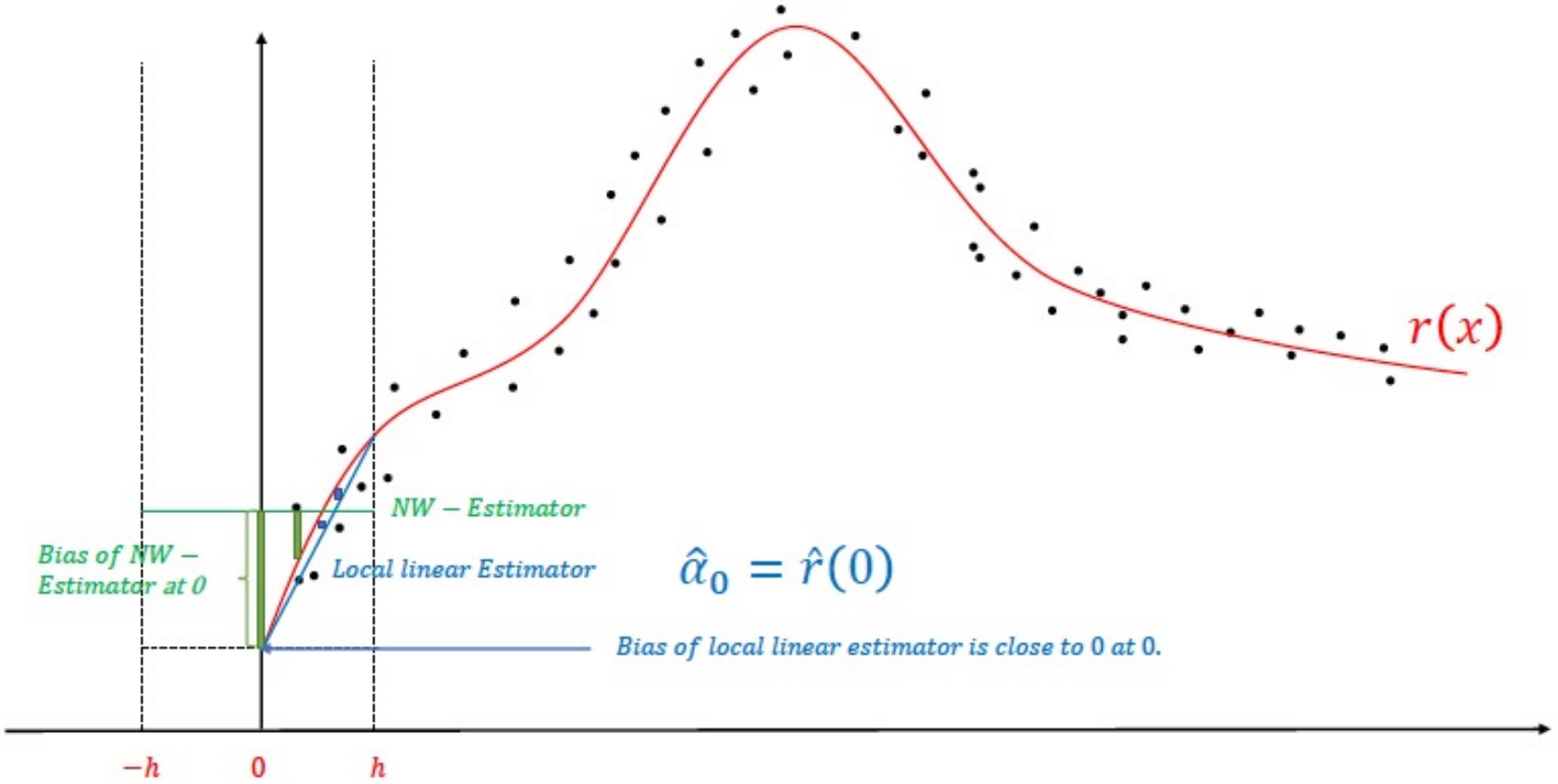


Figure 6: Comparison of Nadaraya–Watson and local linear estimators

Local Polynomial: Matrix Formulation

Define $(p+1) \times 1$ regressor:

$$Z_t(x) = \left[1, (X_t - x), (X_t - x)^2, \ldots, (X_t - x)^p\right]'.$$

Weights:

$$W_t(x) = K_h(x - X_t). \quad \Rightarrow \quad W(x) = \text{diag}(W_1, \ldots, W_T).$$

Stack:

$$Y = (Y_1, \ldots, Y_T)', \quad Z = [Z_1, \ldots, Z_T]'.$$

Weighted LS problem:

$$\widehat{\alpha}(x) = \left(Z'WZ\right)^{-1} Z'WY.$$

Then

$$\widehat{r}(x) = \widehat{\alpha}_0, \qquad \widehat{r}^{(\nu)}(x) = \nu!\widehat{\alpha}_\nu, \quad \nu \leq p.$$

Local Polynomial: Practical Aspects

Specification choices:

- Polynomial order p:
  - $p = 0$: local constant (Nadaraya–Watson).
  - $p = 1$: local linear.
  - $p = 2$: local quadratic, etc.
- Bandwidth h:
  - Selected by cross-validation, plug-in methods, etc.
- Kernel $K(\cdot)$:
  - Same families as in density estimation (Epanechnikov, Gaussian, etc.).

Implementation in R:

- `locpoly()` from `KernSmooth` package:
  - `degree = 0` $\Rightarrow$ Nadaraya–Watson,
  - `degree = 1` $\Rightarrow$ local linear,
  - `degree = 2` $\Rightarrow$ local quadratic.

Local Polynomial Regression in Practice

Compared to Nadaraya–Watson, local polynomial estimators:

- Provide better fit in many settings (lower bias).
- Are equivalent to NW when the density $g(x)$ is known.
- Handle boundary regions more effectively.

In asset pricing and risk management:

- Local polynomial regression can estimate:
  - time-varying risk premia,
  - nonlinear term structure relationships,
  - nonlinear factor effects in multi-factor models.
- Without imposing rigid parametric forms on these relationships.

Table of Contents

Nonparametric Tools in Finance

Nonparametric methods are valuable in finance because:

- Financial data often violate normality and linearity assumptions.
- Tail behavior, skewness, and volatility clustering are common.
- Robustness to misspecification is desired.

This section covers:

- Runs test for randomness,
- Nonparametric autoregression and nonlinear predictability,
- Semiparametric volatility models.

Kernel-based methods can also estimate time-varying parameters in consumption-based asset pricing (C-CAPM), relevant to the equity premium puzzle.

Runs Test

Efficient Market Hypothesis (EMH):

- Prices fully reflect available information.
- Past returns should not predict future returns.

Runs test:

- Nonparametric test of randomness.
- A *run*: sequence of consecutive identical outcomes (e.g. up vs down).
- Under randomness:
  - number and length of runs follow a known distribution,
  - large deviations indicate non-randomness.

Used in finance to:

- Test whether price changes are random,
- Detect serial dependence in sign or direction of returns.

Run Length Definition

Definition (Run length in a time series)

Let X_t be a series of returns and set $Z_0 = 0$. For $t \geq 1$:

1. If $\text{sign}(X_t) = \text{sign}(X_{t-1})$, then
$$Z_t = Z_{t-1} + \text{sign}(X_t).$$
2. If $\text{sign}(X_t) = -\text{sign}(X_{t-1})$, then
$$Z_t = 0.$$

Then Z_t is the signed length of the current run (e.g. a sequence of positives yields increasing positive Z_t). One can analyze the distribution of $\max_t Z_t$ under i.i.d. assumptions to form runs-based tests.

Nonlinear Predictability and Granger Causality

Traditional Granger causality tests:

- Based on VAR models,
- Capture only **linear** predictive relationships.

Definition (Complete (generalized) Granger causality)

Let $\{X_t, Y_t\}$ be strictly stationary. Define $\mathcal{F}_{t-1}^X = \{X_{t-1}, X_{t-2}, \dots\}$, $\mathcal{F}_{t-1}^Y = \{Y_{t-1}, Y_{t-2}, \dots\}$, and $\mathcal{F}_{t-1} = \{\mathcal{F}_{t-1}^X, \mathcal{F}_{t-1}^Y\}$. X_t Granger-causes Y_t in distribution if

$$P(Y_t \le y \mid \mathcal{F}_{t-1}) \neq P(Y_t \le y \mid \mathcal{F}_{t-1}^Y).$$

This notion allows for *nonlinear* predictive effects beyond linear correlation.

Nonlinear Predictability via Functionals g

To detect nonlinear predictability, consider functionals $g(Y_{t-1}, \ldots)$ and

$$\gamma(g) = \text{cov}(Y_t, g_{t-1}), \quad g_{t-1} = g(Y_{t-1}, \ldots).$$

Examples of g_{t-1}:

- Y_{t-1}^2,
- $\text{sign}(Y_{t-1})$,
- $\sum_{j=1}^{p} Y_{t-j}^2$.

Estimate:

$$\widehat{\gamma}(g) = \frac{1}{T} \sum_{t=2}^{T} (Y_t - \bar{Y})(g_{t-1} - \bar{g}).$$

Nonlinear Predictability via Functionals g

Under mild conditions and $\gamma(g) = 0$:

$$\sqrt{T}(\widehat{\gamma}(g) - \gamma(g)) \Longrightarrow N(0, V(g)).$$

Standardized statistic:

$$\widehat{S}(g) = \frac{\sqrt{T}(\widehat{\gamma}(g) - \gamma(g))}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \bar{Y})^2(g_{t-1} - \bar{g})^2}} \Longrightarrow N(0,1),$$

which allows testing for nonlinear predictability for a given g.

Nonparametric Autoregression

Consider the k-step-ahead conditional mean:

$$m_k(y_1, \ldots, y_p) = \mathrm{E}\big(Y_{t+k} \mid Y_t = y_1, \ldots, Y_{t+1-p} = y_p\big).$$

If m_k is smooth, estimate via nonparametric regression.

Univariate case ($p = 1$):

$$\widehat{m}_k(y) = \frac{\sum_{t=1}^{T-k} K\big((Y_t - y)/h\big)\, Y_{t+k}}{\sum_{t=1}^{T-k} K\big((Y_t - y)/h\big)}.$$

With appropriate kernel and bandwidth, under regularity:

- $\widehat{m}_k(y) \to m_k(y)$,
- CLT holds for $\widehat{m}_k(y)$.

This provides a fully nonparametric autoregressive model for Y_t.

Nonparametric AR and EMH

Consider the (relaxed) random walk condition:

Condition rw3': $\{\varepsilon_t\}$ is a martingale difference sequence (MDS):

$$\mathrm{E}(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) = 0.$$

If Y_t satisfies rw3', is stationary, and has finite variance:

- $Y_t = \mu + \varepsilon_t$,
- There should be no predictable component in Y_t given its past.

Nonparametric AR and EMH

Theorem

Assume rw3' and that Y_t has continuous density f_Y positive and differentiable at y. If $h \to 0$ and $Th \to \infty$, then for all k and y:

$$\sqrt{Th}\big(\widehat{m}_k(y) - \widehat{\mu}\big) \Longrightarrow N(0, \omega), \quad \omega = \int K^2(u)\, du \cdot \frac{\sigma_k^2(y)}{f_Y(y)},$$

and

$$\widehat{S}_k(y) = \frac{\widehat{m}_k(y) - \widehat{\mu}}{\sqrt{\frac{\sum K^2((Y_i - y)/h)(Y_{t+k} - \widehat{m}_k(y))^2}{(\sum K((Y_t - y)/h))^2}}} \Longrightarrow N(0, 1),$$

with $\widehat{\mu} = T^{-1} \sum Y_t$.

Testing Predictability with Nonparametric AR

Under rw3', $m_k(y) \equiv \mu$ for all k, y.

Test statistic:

- Evaluate $\widehat{S}_k(y)$ at a grid of points $\{y_1, \ldots, y_L\}$ and lags $k = 1, \ldots, K$.
- Consider the sum

$$\sum_{k=1}^{K} \sum_{l=1}^{L} \widehat{S}_k(y_l)^2.$$

- Under the null of no predictability, this approximately follows a χ^2_{KL} distribution.

Reject EMH (in this sense) if

$$\sum_{k=1}^{K} \sum_{l=1}^{L} \widehat{S}_k(y_l)^2 > \chi^2_{KL}(\alpha).$$

Sieve methods offer alternative approaches; see Chen (2007).

Semiparametric Volatility Models

Consider a GARCH model with unknown innovation distribution:

$$y_t = \nu_t \sigma_t, \quad \sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \alpha y_{t-1}^2,$$

where ν_t i.i.d. with unknown density f.

Issues with parametric f:

- Assuming normal or Student-t may be incorrect.
- Likelihood-based estimates for (ω, α, β) may lose robustness.

Semiparametric Volatility Models

Semiparametric approach:

- Estimate both f and GARCH parameters jointly.
- Use nonparametric density estimation for ν_t (see Section 1).
- Two-step ML (Engle and Gonzalez-Rivera 1991; Linton 1993; Drost, Klaassen, and Werker 1997):
  1. Get consistent $\widehat{\theta}$ via quasi-ML; compute standardized residuals.
  2. Estimate f nonparametrically from residuals.
  3. Re-estimate parameters using estimated f.

Nonparametric Volatility Function $g(\cdot)$

We can relax the GARCH form:

$$\sigma_t^2 = g(y_{t-1}, \ldots, y_{t-p}),$$

with unknown function g.

Estimate g nonparametrically (e.g. kernel or series estimator). For $p = 1$:

$$\widehat{g}(y) = \frac{\sum_{t=2}^{T} K\big((y - y_{t-1})/h\big) y_t^2}{\sum_{t=2}^{T} K\big((y - y_{t-1})/h\big)}.$$

Nonparametric Volatility Function $g(\cdot)$

To subtract conditional mean:

$$\widehat{g}_m(y) = \frac{\sum_{t=2}^{T} K((y - y_{t-1})/h) y_t^2}{\sum_{t=2}^{T} K((y - y_{t-1})/h)} - \left(\frac{\sum_{t=2}^{T} K((y - y_{t-1})/h) y_t}{\sum_{t=2}^{T} K((y - y_{t-1})/h)} \right)^2,$$

$$\widehat{g}_r(y) = \frac{\sum_{t=2}^{T} K((y - y_{t-1})/h) \widehat{u}_t^2}{\sum_{t=2}^{T} K((y - y_{t-1})/h)}, \quad \widehat{u}_t = y_t - \frac{\sum K((y - y_{t-1})/h) y_t}{\sum K((y - y_{t-1})/h)}.$$

This yields a fully nonparametric volatility function.

Table of Contents

Summary of Nonparametric Methods

This chapter developed nonparametric and semiparametric tools for econometrics, with emphasis on applications to financial data.

Key points:

- **Kernel density estimation**:
  - Flexible estimation of marginal and joint distributions.
  - Bias and variance depend on bandwidth h and kernel choice.
  - Optimal bandwidth balances bias–variance trade-off.
- **Nonparametric regression**:
  - Nadaraya–Watson estimator (local constant).
  - Local polynomial regression (local linear, quadratic, etc.) with better boundary behavior and derivative estimation.

Summary of Nonparametric Methods

- **Financial applications**:
  - Runs tests for randomness.
  - Nonparametric autoregression for nonlinear predictability.
  - Semiparametric volatility models and nonparametric volatility functions.

Nonparametric methods, by avoiding restrictive functional forms, provide a flexible, data-driven framework for modelling complex economic and financial relationships.

Table of Contents

References I

- Ait-Sahalia, Yacine (1996). "Testing continuous-time models of the spot interest rate". In: *The Review of Financial Studies* 9.2, pp. 385–426.
- Chen, Xiaohong (2007). "Large sample sieve estimation of semi-nonparametric models". In: *Handbook of econometrics* 6, pp. 5549–5632.
- Drost, Feike C, Chris AJ Klaassen, and Bas JM Werker (1997). "Adaptive estimation in time-series models". In: *The Annals of Statistics* 25.2, pp. 786–817.
- Engle, Robert F and Gloria Gonzalez-Rivera (1991). "Semiparametric ARCH models". In: *Journal of Business & Economic Statistics* 9.4, pp. 345–359.
- Hong, Yongmiao and Haitao Li (2005). "Nonparametric specification testing for continuous-time models with applications to term structure of interest rates". In: *The Review of Financial Studies* 18.1, pp. 37–84.

References II

 Linton, Oliver (1993). "Adaptive estimation in ARCH models". In: *Econometric Theory* 9.4, pp. 539–569.

Chapter 6 — HAR Inference

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

What is HAR Inference?

- **HAR** = heteroskedasticity- and autocorrelation-robust inference.
- Includes:
  - HAC (heteroskedasticity- and autocorrelation-consistent) methods.
  - "Inconsistent but robust" procedures:
    - Fixed-b asymptotics.
    - Self-normalization.
    - Bootstrap.
    - GMM-based robust inference.
- Goal: valid standard errors, tests, and confidence intervals when errors are heteroskedastic and/or serially correlated.

Sampling Distributions and Standard Errors

- Estimators are random variables with **sampling distributions**.
- Even if a regressor has no true effect, its OLS coefficient is almost never exactly zero.
- To test $H_0 : \beta_j = 0$ we need:

$$t = \frac{\widehat{\beta}_j - 0}{\widehat{\mathrm{se}}(\widehat{\beta}_j)}.$$

- Under heteroskedasticity/autocorrelation, conventional formulas for $\widehat{\mathrm{se}}$ fail $\Rightarrow$ misleading inference.
- HAR methods repair this by targeting the **long-run variance (LRV)**.

Long–Run Variance and HAC Estimators

- For a centered stationary process $\{X_t\}$,

$$\sigma^2_{\text{LRV}} = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h), \quad \gamma(h) = \text{Cov}(X_t, X_{t+h}).$$

- LRV captures cumulative effect of serial dependence and time-varying volatility.
- **HAC estimators** (e.g. Newey–West (Newey and West 1987)):
  - Plug-in estimators of autocovariances $\widehat{\gamma}(h)$.
  - Kernel weights and bandwidth b_T.
  - Bandwidth $b_T \to 0$, but $Tb_T \to \infty$ (classical asymptotics).
- Widely used in empirical economics for robust standard errors.

Whitney K. Newey

Whitney K. Newey is the Ford Professor of Economics at MIT and co-developer of the Newey–West estimator for heteroskedasticity- and autocorrelation-robust standard errors. He received his B.A. from Brigham Young University in 1978 and his Ph.D. from MIT in 1983 under Jerry A. Hausman, later teaching at Princeton before returning to MIT in 1990.

Newey's research has shaped modern econometrics, with major contributions to generalized method of moments (GMM), semiparametric estimation, and inference with many instruments. He is a Distinguished Fellow of the American Economic Association (2020) and a Fellow of the Econometric Society and the American Academy of Arts and Sciences. His recent work focuses on debiased machine learning methods for structural and causal parameters and on econometric techniques for models with general heterogeneity.

Fixed-b Asymptotics

- HAC: bandwidth fraction $b_T = \ell_T / T \to 0$ as $T \to \infty$.
- **Fixed-b** asymptotics (Kiefer and Vogelsang 2002a; Kiefer and Vogelsang 2002b; Kiefer and Vogelsang 2005):
  - Keep $b_T = b \in (0, 1]$ fixed in the limit.
  - Asymptotic distribution depends on b.
- Advantages (Kiefer and Vogelsang 2005):
  - More accurate first-order approximation to finite-sample distribution.
  - Informative local power analysis for HAC-robust tests.
- Trade-off:
  - Larger b: better size control, but lower power.
  - Limiting distribution depends on nuisance parameter b.

Self-Normalization: Basic Idea

- Originates with Student (1908):

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

  where S^2 is sample variance.
- S^2 is not consistent when n is small, but is *stochastically proportional* to the true variance.
- Key idea:
  - Use a random but well-behaved *self-normalizer* in the denominator.
  - Construct pivotal statistics with nuisance-parameter-free limits.
- Systematic theory in Shao (2010).

Self-Normalization for Time Series

- For a time series $\{X_t\}_{t=1}^n$, define

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t, \quad \bar{X}_t = \frac{1}{t}\sum_{i=1}^{t} X_i.$$

- Self-normalizer (Shao 2010):

$$W_n^2 = n^{-2}\sum_{t=1}^{n} t^2(\bar{X}_t - \bar{X}_n)^2.$$

- Under suitable ergodicity conditions:

$$W_n^2 \Longrightarrow \sigma^2 \int_0^1 [\mathbb{B}(s)]^2 ds,$$

  where $\mathbb{B}(s)$ is a Brownian bridge and σ^2 is the LRV.
- W_n^2 is proportional to σ^2 with a *known* limit distribution $\Rightarrow$ use as self-normalizer in tests and CIs.

Drawbacks and Adjusted Range

- W_n^2 is a variance of partial sums:
  - Sensitive to outliers, strong autocorrelation, heteroskedasticity, long memory.
- Hong et al. (2024) propose an **adjusted range** of the partial sum process:
  - Range-based statistics are more robust to:
    - Persistent autocorrelation.
    - Irregularities: outliers, skewness, heavy tails.
    - Some forms of infinite-variance behavior.
  - Asymptotically proportional to $\sqrt{\text{LRV}}$ up to a stochastic factor.
  - Limiting distribution is nuisance-parameter free.
- Provides an alternative self-normalizer with robustness advantages.

Bootstrap for HAR Inference

- Bootstrap: approximate sampling distribution by resampling.
- Advantages:
  - Less reliance on parametric assumptions.
  - Flexible for heteroskedasticity and dependence.
- Time series variants:
  - Residual/bootstrap of residuals.
  - Block bootstrap (moving, stationary blocks).
  - Wild bootstrap (for conditional heteroskedasticity).
  - Subsampling: no resampling, uses overlapping windows.
- Often improves finite-sample size and power of HAR tests.

GMM and HAR Inference

- GMM (Hansen 1982) uses population moment conditions:

$$\mathrm{E}[m(Z_t, \theta_0)] = 0.$$

- GMM estimator:

$$\widehat{\theta}_{\mathrm{GMM}} = \arg\min_{\theta} \widehat{g}(\theta)' \widehat{W} \widehat{g}(\theta),$$

  where $\widehat{g}(\theta)$ is sample average of moment conditions.
- Robustness:
  - Weighting matrix $\widehat{W}$ estimated via HAC (e.g. Newey–West).
  - Consistency and asymptotic normality under heteroskedasticity and autocorrelation.
- Issues:
  - Sensitive to instrument choice and weak instruments.
  - Small-sample distortions, numerical instability.

Extensions of GMM

- **Continuously updated GMM (CUE)**:
  - Updates weighting matrix at each θ.
- **Empirical likelihood-based GMM**:
  - Likelihood-style inference based on moment conditions.
  - Often better finite-sample properties.
- **Bootstrapped GMM**:
  - Use bootstrap to correct size and improve coverage.
- Widely used in dynamic panel models and asset pricing (e.g. C-CAPM).

HAR Toolbox: Summary

- **HAC**: consistent LRV estimation; ubiquitous in practice.
- **Fixed-b**: better finite-sample approximation; nuisance-dependent limiting distributions.
- **Self-normalization**: pivotal statistics without consistent LRV estimation.
- **Bootstrap**: data-driven approximation of sampling distributions.
- **GMM**: model-based, flexible, and robust via HAC weighting.

Table of Contents

Motivation for Spectral Analysis

- Frequency-domain view of time series:
  - Decompose a stationary process into cycles of different lengths.
- The **spectral density** (spectrum) describes how variance is distributed across frequencies.
- Long-run variance corresponds to behavior at *low frequencies*.
- Understanding spectral properties clarifies HAC estimators and LRV estimation.

Periodic Functions

Definition 1 (Periodic Function)

A real-valued function $g : \mathbb{R} \to \mathbb{R}$ is periodic with period (wavelength) $\alpha > 0$ if

$$g(t + k\alpha) = g(t) \quad \text{for all } t \in \mathbb{R},\ k \in \mathbb{Z}.$$

- $\sin(t)$ and $\cos(t)$ are standard examples with period 2π.
- Fourier series represents functions as sums of sines and cosines.

Sinusoidal Representation

Consider

$$g(t) = A\sin(\lambda t + \theta).$$

- $A > 0$: amplitude (peak height, overall scale).
- λ: frequency; period $2\pi/\lambda$.
- θ: phase shift; horizontal shift of the curve.

Equivalent form:

$$g(t) = a\sin(\lambda t) + b\cos(\lambda t),$$

where

$$A = \sqrt{a^2 + b^2}, \quad \theta = \arctan\left(\frac{b}{a}\right) \ (a \neq 0).$$

Autocovariance and Spectrum

Let $\{Y_t\}$ be covariance-stationary with mean μ and autocovariance

$$\gamma(h) = \mathrm{Cov}(Y_t, Y_{t+h}).$$

Define the spectral density $f_Y(\lambda)$:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{i\lambda h} f_Y(\lambda)\, d\lambda,$$

$$f_Y(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i\lambda h}.$$

- $f_Y(\lambda)$ describes how variance is allocated across frequencies.
- For real-valued $\{Y_t\}$ with $\gamma(h) = \gamma(-h)$:

$$f_Y(\lambda) = \frac{1}{2\pi}\Big\{\gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h)\cos(\lambda h)\Big\}.$$

Properties of the Spectrum

- If $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, then $f_Y(\lambda)$ is:
  - Continuous in λ.
  - Real-valued and nonnegative.
- Periodicity:

$$\cos[(\lambda + 2\pi k)h] = \cos(\lambda h) \Rightarrow f_Y(\lambda) = f_Y(\lambda + 2\pi k).$$

- Symmetry: $f_Y(\lambda) = f_Y(-\lambda)$.
- Total variance:

$$\int_{-\pi}^{\pi} f_Y(\lambda)\, d\lambda = \gamma(0).$$

- Normalized spectrum $f_Y(\lambda)/\gamma(0)$ integrates to 1 and acts like a probability density; cumulative spectral distribution:

$$F_Y(\lambda) = \int_{-\pi}^{\lambda} f_Y(s)\, ds.$$

Spectral Representation Theorem

Theorem 2 (Spectral Representation)

Let $\{Y_t\}$ be a real-valued stationary time series with mean 0. Then

$$Y_t = \int_{-\pi}^{\pi} e^{i\lambda t}\, dF_Y(\lambda),$$

where:

- $e^{i\lambda t}$ *are complex exponentials (equivalently, sines and cosines).*
- $\lambda \in [-\pi, \pi]$ *is frequency.*
- F_Y *is a (complex-valued) spectral distribution function linked to* f_Y*.*

Euler's Formula and De Moivre

- Euler's formula:
$$e^{i\theta} = \cos\theta + i\sin\theta.$$
- Derived from Taylor series expansions of e^z, $\cos z$, and $\sin z$.
- De Moivre's theorem:
$$(e^{i\theta})^n = e^{in\theta},$$
equivalently
$$(\cos\theta + i\sin\theta)^n = \cos(n\theta) + i\sin(n\theta).$$
- These results link complex exponentials to sines and cosines and underlie the spectral representation.

Example: White Noise

Example 3

Let $\{\varepsilon_t\}$ be white noise with variance σ^2.

- $\gamma(0) = \sigma^2$, $\gamma(h) = 0$ for $h \neq 0$.
- Spectral density:

$$f_\varepsilon(\lambda) = \frac{\sigma^2}{2\pi}, \quad \lambda \in [-\pi, \pi].$$

- Flat spectrum: all frequencies contribute equally.

Example: Dirac Delta and a Pure Sine Wave

Example 4

The Dirac delta $\delta_{\lambda_0}(\lambda)$ satisfies

$$\int g(s)\delta_{\lambda_0}(s)\,ds = g(\lambda_0).$$

Applying to $e^{i\lambda k}$:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda k}\delta_{\lambda_0}(\lambda)\,d\lambda = \frac{1}{2\pi}\cdot 2\pi\cos(\lambda_0 k) = \cos(\lambda_0 k).$$

- A single frequency λ_0 corresponds to a *line spectrum* at λ_0.
- Represents a purely periodic (deterministic) component.

MA(∞) and MA(1) Spectra

Consider

$$Y_t = \mu + \psi(L)\varepsilon_t, \quad \psi(L) = \sum_{j=0}^{\infty} \psi_j L^j,$$

with ε_t white noise, $\mathrm{Var}(\varepsilon_t) = \sigma^2$.

- Autocovariance generating function:

$$g_Y(z) = \sigma^2 \psi(z)\psi(z^{-1}).$$

- Spectrum:

$$s_Y(\lambda) = \frac{\sigma^2}{2\pi} \psi(e^{-i\lambda})\psi(e^{i\lambda}).$$

MA(∞) and MA(1) Spectra

Consider MA(1): $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$:

$$s_Y(\lambda) = \frac{\sigma^2}{2\pi}\left[1 + \theta^2 + 2\theta\cos\lambda\right].$$

- $\theta > 0$: decreasing spectrum in λ.
- $\theta < 0$: increasing spectrum in λ.

AR(1) and ARMA(p, q)

AR(1): $Y_t = c + \phi Y_{t-1} + \varepsilon_t$, $|\phi| < 1$.

$$s_Y(\lambda) = \frac{\sigma^2}{2\pi}\left[1 + \phi^2 - 2\phi\cos\lambda\right]^{-1}.$$

- $\phi > 0$: decreasing spectrum (low frequencies dominant).
- $\phi < 0$: increasing spectrum (higher frequencies more important).

AR(1) and ARMA(p, q)

General ARMA(p, q):

$$s_Y(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|1 + \sum_{k=1}^{q} \theta_k e^{-ik\lambda}\right|^2}{\left|1 - \sum_{k=1}^{p} \phi_k e^{-ik\lambda}\right|^2}.$$

- AR roots near the unit circle $\Rightarrow$ strong low-frequency components.
- MA roots shape short-run (higher-frequency) behavior.

Sample Autocovariances

Given observations $Y_1, \ldots, Y_T$, define sample mean

$$\bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y_t.$$

Sample autocovariances:

$$\widehat{\gamma}(j) = \begin{cases} T^{-1} \sum_{t=j+1}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}), & j = 0, 1, \ldots, T-1, \\ \widehat{\gamma}(-j), & j < 0. \end{cases}$$

- Up to $T - 1$ distinct lags.
- Symmetric: $\widehat{\gamma}(-j) = \widehat{\gamma}(j)$.

Sample Periodogram

Sample periodogram (analogue of spectrum):

$$\widehat{I}(\lambda) = \frac{1}{2\pi} \sum_{j=1-T}^{T-1} \widehat{\gamma}(j) e^{-i\lambda j} = \frac{1}{2\pi} \Big[\widehat{\gamma}(0) + 2 \sum_{j=1}^{T-1} \widehat{\gamma}(j) \cos(\lambda j) \Big].$$

- Symmetric around $\lambda = 0$.
- Total area:

$$\int_{-\pi}^{\pi} \widehat{I}(\lambda)\, d\lambda = \widehat{\gamma}(0),$$

so

$$2 \int_{0}^{\pi} \widehat{I}(\lambda)\, d\lambda = \widehat{\gamma}(0).$$

- Connects variance of Y_t with area under the periodogram.

Regression Interpretation of the Periodogram

- Think of Y_t regressed on sines and cosines at discrete frequencies.
- For T even, regress Y_t on

$$\cos\left(\frac{2\lambda jt}{T}\right), \ \sin\left(\frac{2\lambda jt}{T}\right), \quad j = 1, \ldots, T/2 - 1,$$

  plus $\cos(\lambda t)$ and an intercept.
- Regression:

$$Y_t = a_0 + \sum_{j=1}^{T/2-1} [a_j \cos(2\lambda jt/T) + b_j \sin(2\lambda jt/T)] + a_{T/2} \cos(\lambda t).$$

- Periodogram at frequency j:

$$I_j = a_j^2 + b_j^2.$$

- Interpreted as contribution of frequency j to variance of Y_t.

Limitations: Inconsistency of the Periodogram

Simplest case: IID $\{X_t\}$ with $\mathrm{E}(X_t) = 0$.

- True spectrum:
$$f_Y(\lambda) = \frac{1}{2\pi}\gamma(0).$$
- Periodogram:
$$\mathrm{E}[\widehat{I}(\lambda)] = f_Y(\lambda), \quad \text{unbiased.}$$
- But $\mathrm{Var}[\widehat{I}(\lambda)]$ does not vanish as $T \to \infty$:
$$\mathrm{Var}[\widehat{I}(\lambda)] = O(1).$$
- Hence $\widehat{I}(\lambda)$ is **not** consistent for $f_Y(\lambda)$.
- Intuition: we estimate too many autocovariances $\widehat{\gamma}(j)$ relative to sample size.

IMSE Perspective and Smoothing

Integrated MSE:

$$\mathrm{IMSE}(\widehat{I}) = \mathrm{E} \int_{-\pi}^{\pi} |\widehat{I}(\lambda) - f_Y(\lambda)|^2 \, d\lambda.$$

- Bias terms can vanish under summability of $\gamma(h)$.
- Variance term stays of order $O(1)$ as $T \to \infty$.

Remedy: smooth the periodogram:

$$\widehat{f}_Y(\lambda) = \frac{1}{T} \sum_{j=-n}^{n} w_b(\lambda - \lambda_j) \, \widehat{I}(\lambda_j),$$

where

- $w_b(u) = w(u/b)/b$ is a kernel with bandwidth b,
- $b \to 0$ and $Tb \to \infty$ as $T \to \infty$.

Under regularity conditions, $\widehat{f}_Y(\lambda)$ is **consistent**.

Practical Smoothing

- Smoothing reduces noise but may blur important peaks.
- Choice of bandwidth (or spans) is crucial:
  - Small bandwidth: noisy estimate.
  - Large bandwidth: oversmoothing, loss of detail.
- In R:
  - `spectrum()` and `spec.pgram()` implement nonparametric spectral estimation.
  - Modified Daniell smoothers used by default (moving averages).
- Experimentation is often needed to balance variance and bias.

Table of Contents

Heteroskedasticity- and Autocorrelation-Robust Inference

- We study HAC inference in the linear regression model.
- When regressors and errors may be serially correlated and heteroskedastic:
  - OLS is still consistent under exogeneity.
  - But usual OLS standard errors and t–statistics can be badly distorted.
- Classical assumptions (homoskedastic, uncorrelated errors) fail, so the usual variance formula no longer matches the true sampling variance.
- HAC standard errors correct for heteroskedasticity and autocorrelation and deliver robust inference when the classical model is misspecified.

Regression Model and OLS Estimator

- Model:

$$Y_t = X_t'\beta + u_t, \quad t = 1, 2, \ldots, T,$$

where

  - β is $q \times 1$,
  - X_t is $q \times 1$,
  - u_t satisfies $\mathrm{E}(u_t \mid X_t) = 0$ but may be autocorrelated and heteroskedastic.

- OLS estimator:

$$\widehat{\beta} = \Big(\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \sum_{t=1}^{T} X_t Y_t.$$

- Rewriting:

$$\sqrt{T}(\widehat{\beta} - \beta) = \Big(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t\Big).$$

Asymptotic Distribution and Long-Run Variance

- Assume WLLN and CLT:

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q, \quad \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t \Longrightarrow N(0, \mathbf{\Omega}_v),$$

  where $v_t = X_t u_t$.

- Then

$$\sqrt{T}(\widehat{\beta} - \beta) \Longrightarrow N\big(0,\ Q^{-1}\mathbf{\Omega}_v Q^{-1}\big).$$

- Q is easy to estimate; the main difficulty is estimating the **long-run variance** (LRV) matrix $\mathbf{\Omega}_v$.
- With $v_t = X_t u_t$ and $\mathrm{E}(v_t) = 0$:

$$\mathbf{\Omega}_v = \mathrm{Var}\Big(\tfrac{1}{\sqrt{T}}\sum_{t=1}^{T} v_t\Big) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{E}(v_t v_s').$$

LRV as Sum of Autocovariances and Spectrum at Zero

- If $\{v_t\}$ is second-order stationary, let

$$\gamma(j) = \mathrm{E}(v_t v_{t-j}'), \quad j \in \mathbb{Z}.$$

- Then

$$\mathbf{\Omega}_v = \frac{1}{T} \sum_{j=-(T-1)}^{T-1} (T - |j|)\gamma(j) \longrightarrow \sum_{j=-\infty}^{\infty} \gamma(j).$$

- In the spectral domain, if $S_v(\lambda)$ is the spectral density of v_t,

$$\mathbf{\Omega}_v = \sum_{j=-\infty}^{\infty} \gamma(j) = 2\pi S_v(0),$$

i.e. the LRV equals the spectrum at frequency zero up to the factor 2π.

Newey–West HAC Estimator

- HAC LRV estimators are widely used:
  - White (1980),
  - Newey and West (1987) and Newey and West (1994),
  - Andrews (1991) and Andrews and Monahan (1992), etc.
- Newey–West estimator:

$$\widehat{\mathbf{\Omega}}_v^{\mathrm{NW}} = \sum_{j=-m}^{m} \left(1 - \left|\frac{j}{m}\right|\right) \widehat{\gamma}(j),$$

where

$$\widehat{\gamma}(j) = \frac{1}{T} \sum_{t=|j|+1}^{T} \widehat{v}_t \widehat{v}_{t-j}', \quad \widehat{v}_t = X_t \widehat{u}_t.$$

- Here $\widehat{u}_t$ are OLS residuals, and m is the bandwidth (number of lags).
- Andrews (1991) recommend $m = 0.75\, T^{1/3}$.

Development of HAC LRV Estimators

- Two equivalent perspectives on HAC LRV estimation:
  1. **Sum of covariances** (time domain).
  2. **Weighted periodogram** (frequency domain).
- For a stationary $\{v_t\}$ with $\mathrm{E}\|v_t\|^2 < \infty$:

$$\mathbf{\Omega}_v = \sum_{j=-\infty}^{\infty} \gamma(j).$$

- Estimate via

$$\widehat{\mathbf{\Omega}}_v^{SC} = \sum_{j=-(T-1)}^{T-1} k(j)\,\widehat{\gamma}(j),$$

where $k(j)$ is a kernel (weight) function.

Sum-of-Covariances Estimator and Bartlett Kernel

- Sum-of-covariances (SC) estimator:

$$\widehat{\mathbf{\Omega}}_v^{SC} = \sum_{j=-(T-1)}^{T-1} k(j)\,\widehat{\gamma}(j),$$

with

$$\widehat{\gamma}(j) = \frac{1}{T}\sum_{t=1}^{T} v_t v_{t-j}'.$$

- $k(j)$ is chosen from a kernel family. Example: **Bartlett kernel**

$$k(j) = 1 - \left|\frac{j}{m}\right|, \quad |j| \le m,$$

and $k(j) = 0$ otherwise.

- m is a truncation parameter controlling how many autocovariances are used.

VAR Prewhitening

VAR Prewhitening (Andrews and Monahan 1992)

- If v_t exhibits strong autocorrelation/heteroskedasticity, prewhiten via a VAR(p):

$$v_t = \Phi_1 v_{t-1} + \cdots + \Phi_p v_{t-p} + \epsilon_t.$$

- Choose lag p by an information criterion (e.g. AIC).
- Step 1: estimate VAR, obtain residuals $\widehat{\epsilon}_t$.
- Step 2: compute HAC LRV estimator $\widehat{\Omega}_\epsilon$ for ϵ_t.
- Step 3: transform back to obtain LRV for v_t:

$$\widehat{\Omega}_v = \left(I_m - \sum_{i=1}^{p} \widehat{\Phi}_i\right)^{-1} \widehat{\Omega}_\epsilon \left(I_m - \sum_{i=1}^{p} \widehat{\Phi}_i\right)^{-1},$$

where I_m is an $m \times m$ identity matrix.

Cochrane–Orcutt Procedure

Cochrane–Orcutt (Cochrane and Orcutt 1949)

- Model:

$$Y_t = \alpha + \beta X_t + u_t, \quad u_t = \rho u_{t-1} + \epsilon_t.$$

- Steps:
  1. Run initial OLS, obtain residuals $\hat{u}_t$.
  2. Estimate ρ via regression:

$$\hat{u}_t = \hat{\rho}\, \hat{u}_{t-1} + \hat{\epsilon}_t.$$

  3. Transform variables:

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1}, \quad X_t^* = X_t - \hat{\rho} X_{t-1}.$$

  4. Re-estimate:

$$Y_t^* = \alpha^* + \beta^* X_t^* + u_t^*.$$

- Result: more efficient estimate of β under AR(1) errors.

Fourier Transform and Periodogram

- Fourier transform of v_t:

$$d_v(\lambda) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} v_t e^{-i\lambda t}.$$

- Decompose into real/imaginary parts:

$$d_v(\lambda) = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_t \cos(\lambda t) - i \frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_t \sin(\lambda t) \right) = v_1 - i v_2,$$

  where v_1, v_2 are (asymptotically) mean-zero normal vectors.

- Periodogram:

$$I_v(\lambda) = d_v(\lambda) \overline{d_v(\lambda)}',$$

  which in the scalar case is $v_1^2 + v_2^2$.

Distribution of the Periodogram

- Asymptotically,
$$I_v(\lambda) \Longrightarrow S_v(\lambda) \times (\chi^2_2/2),$$
so periodogram ordinates behave like scaled χ^2 variables.
- For Fourier frequencies
$$\lambda_j = \frac{2\pi j}{T}, \quad j = 0, 1, \ldots, T-1,$$
the $d_v(\lambda_j)$ are asymptotically independent for $j \neq k$.
- Orthogonality of sines and cosines $\Rightarrow$ asymptotic independence across frequencies.
- This simplifies frequency-domain inference and LRV estimation.

Weighted Periodogram Estimator of LRV

- Periodogram $I_v(\lambda)$ estimates the spectral density at λ.
- A **weighted periodogram** estimator of the LRV:

$$\widehat{\mathbf{\Omega}}^{WP} = 2\pi \sum_{l=-(T-1)}^{T-1} K(l)\, \widehat{I}_v\Big(\frac{2\pi l}{T}\Big),$$

where

  - $K(l)$ is a kernel (weight) function in the frequency domain,
  - typically gives more weight to low frequencies near zero.

- Averaging periodogram ordinates near zero frequency yields a consistent estimator of the LRV under suitable conditions.

Positive Semi-Definiteness Problem

- Not all covariance-matrix estimators $\widehat{\mathbf{\Omega}}$ are PSD.
- Example: m-period return, u_t is MA$(m-1)$, and we use

$$\tilde{\mathbf{\Omega}} = \sum_{j=-(m-1)}^{m-1} \widehat{\gamma}(j).$$

- For $m = 2$,

$$\tilde{\mathbf{\Omega}} = \sum_{j=-1}^{1} \widehat{\gamma}(j) = \widehat{\gamma}(0)\left(1 + 2\frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)}\right).$$

- This can be negative if $\widehat{\gamma}(1)/\widehat{\gamma}(0) < -0.5$.
- So $\tilde{\mathbf{\Omega}}$ is not guaranteed PSD w.p.1.

Ensuring PSD via Frequency-Domain Weights

- We want $\widehat{\Omega}^{WP}$ to be PSD:

$$\lambda' \widehat{\Omega}^{WP} \lambda \geq 0 \quad \text{for all vectors } \lambda.$$

- With

$$\widehat{\Omega}^{WP} = 2\pi \sum_{l} K(l)\, \widehat{I}_v(2\pi l / T),$$

we get

$$\begin{aligned} \lambda' \widehat{\Omega}^{WP} \lambda &= 2\pi \sum_{l} K(l) \left(\lambda' \widehat{I}_v(2\pi l / T) \lambda \right) \\ &= 2\pi \sum_{l} K(l) \left| \lambda' d_v(\lambda_l) \right|^2 \geq 0 \end{aligned}$$

if $K(l) \geq 0$ for all l.

- So a nonnegative frequency-domain weight function $K(l)$ guarantees PSD w.p.1.

Equivalence of WP and SC Estimators

- Starting from

$$\widehat{\Omega}^{WP} = 2\pi \sum_{l} K(l)\, \widehat{I}_v(2\pi l/T),$$

algebra shows

$$\widehat{\Omega}^{WP} = \sum_{j=-(T-1)}^{T-1} \widehat{\Gamma}_j\, k(j),$$

where

$$k(j) = \sum_{l=-(T-1)}^{T-1} K(l)\, e^{-i(2\pi j/T)l}$$

is the inverse discrete Fourier transform of K.

- Thus $\widehat{\Omega}^{SC}$ is PSD w.p.1 iff k arises from a nonnegative K.
- If K is symmetric, $k(j)$ is real and can be written as

$$k(j) = K(0) + 2\sum_{l=1}^{T-1} K(l)\cos\left((2\pi j/T)l\right).$$

Kernels in Time and Frequency Domains

- Time-domain kernel $k(x)$ and frequency-domain kernel $K(u)$ are related by Fourier transform:

$$K(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} k(x) e^{-iux}\, dx, \quad k(x) = \int_{-\infty}^{\infty} K(u) e^{iux}\, du.$$

- Common kernels have well-known transforms, linking:
  - time-domain weighting of autocovariances,
  - frequency-domain filtering of periodogram ordinates.
- Table 1 summarizes some widely used choices.

Common Kernels and Fourier Transforms

Table: Summary of kernels and their Fourier transforms.

Kernel	$k(x)$	**Fourier Transform** $K(u)$
Truncated	$\mathbf{1}(\lvert x\rvert \le 1)$	$\frac{1}{\pi}\frac{\sin u}{u}$
Bartlett	$(1-\lvert x\rvert)\mathbf{1}(\lvert x\rvert \le 1)$	$\frac{1}{2\pi}\left[\frac{\sin(u/2)}{u/2}\right]^2$
Daniell	$\frac{\sin(\pi x)}{\pi x}$	$\frac{1}{2\pi}\mathbf{1}(\lvert u\rvert \le \pi)$
Parzen	$\begin{cases} 1-6x^2+6\lvert x\rvert^3 & \lvert x\rvert \le \frac{1}{2} \\ 2(1-\lvert x\rvert)^3 & \frac{1}{2} < \lvert x\rvert \le 1 \\ 0 & \text{otherwise} \end{cases}$	$\frac{3}{8\pi}\left[\frac{\sin(u/4)}{u/4}\right]^4$
Quadratic-Spectral	$\frac{3}{(\pi x)^2}\left[\frac{\sin \pi x}{\pi x} - \cos(\pi x)\right]$	$\frac{3}{4\pi}[1-(u/\pi)^2]\mathbf{1}(\lvert u\rvert \le \pi)$

Common Kernels and Fourier Transforms

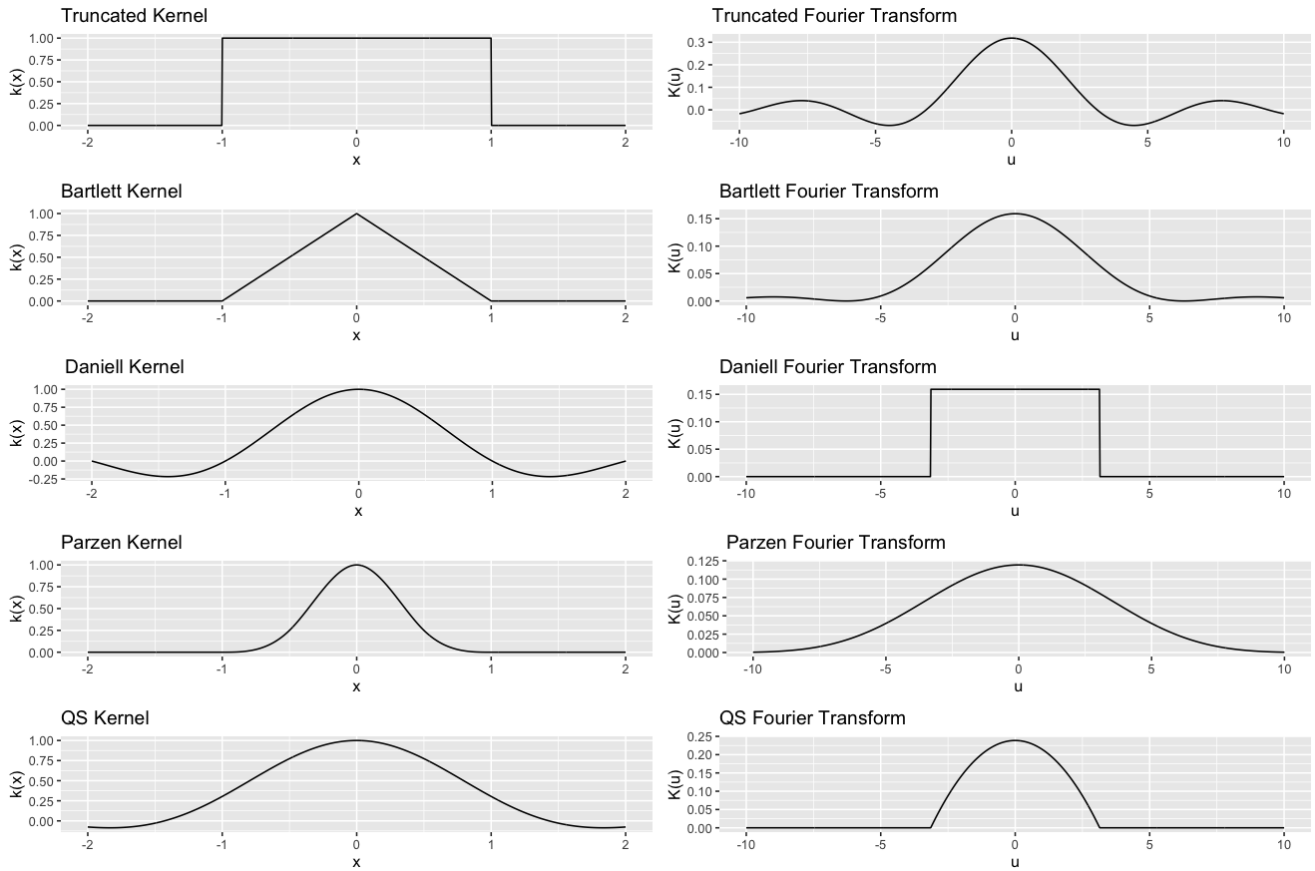


Figure: Common kernel functions $k(x)$ used in nonparametric spectral density estimation (left panels) and their corresponding Fourier transforms $K(u)$ (right panels).

Interpretation of Common Kernels (1)

1. **Truncated (Uniform) Kernel**
   - Time domain: $k(x) = 1$ on $[-1, 1]$, 0 otherwise (uniform weights).
   - Frequency domain: sinc-type transform $\propto \sin(u)/u$, slowly decaying with oscillations $\Rightarrow$ low bias, high variance.
2. **Bartlett Kernel**
   - Time domain: triangular, linearly decreasing from 1 to 0 on $[-1, 1]$.
   - Frequency domain: squared sinc; smoother response, lower variance, higher bias.
3. **Daniell Kernel**
   - Time domain: sinc function; oscillatory but decaying weights.
   - Frequency domain: rectangular; ideal low-pass filter shape.

Interpretation of Common Kernels (2)

4. **Parzen Kernel**
   - Time domain: smooth, compactly supported piecewise polynomial.
   - Frequency domain: quartic power of sinc; very smooth, low variance, but higher bias at high frequencies.
5. **Quadratic-Spectral (Priestley) Kernel**
   - Time domain: modified sinc with cosine correction.
   - Frequency domain: truncated parabola on $[-\pi, \pi]$; good compromise between bias and variance, emphasizing low frequencies.

Bias-Variance Trade-Off and Kernels

- Choice of kernel reflects a classic **Bias-Variance trade-off**.
- Kernels with broad, smooth Fourier transforms (e.g. Parzen):
  - More smoothing $\Rightarrow$ lower variance, higher bias.
- Kernels with sharper or oscillatory transforms (e.g. truncated, Daniell):
  - Less smoothing $\Rightarrow$ lower bias, higher variance.
- Suitable choice depends on:
  - Noise level,
  - Smoothness of the underlying spectral density,
  - Sample size.

Choices for Kernel and Bandwidth

- Two main decisions:
  1. Kernel shape $k(\cdot)$ or $K(\cdot)$.
  2. Bandwidth (or truncation parameter) controlling smoothness.
- In practice, **bandwidth choice** is usually more important than the specific kernel.
- Bandwidth determines the Bias-Variance trade-off and thus the effectiveness of the estimator.
- We illustrate this with a flat kernel in the frequency domain.

Flat Kernel in the Frequency Domain

- Consider

$$\widehat{\Omega}^{WP} = 2\pi \sum_{l=-(T-1)}^{T-1} K(l)\, \widehat{I}_v\Big(\frac{2\pi l}{T}\Big),$$

with a flat kernel

$$K(l) = \begin{cases} \frac{1}{2b+1}, & |l| \le b, \\ 0, & |l| > b. \end{cases}$$

- Then

$$\widetilde{\widehat{\Omega}}^{WP} = \frac{2\pi}{2b+1} \sum_{l=-b}^{b} \widehat{I}_v\Big(\frac{2\pi l}{T}\Big),$$

averaging $2b+1$ periodogram ordinates near zero.

- This estimator is PSD by construction.

Corresponding Time-Domain Kernel

- The time-domain kernel corresponding to the flat frequency-domain kernel is

$$k(j) = \sum_{l=-(T-1)}^{T-1} K(l) e^{-i(2\pi j/T)l}$$

$$= \frac{1}{2b+1} \sum_{l=-b}^{b} e^{-i(2\pi j/T)l}$$

$$\to \frac{\sin(2\pi j/m)}{2\pi j/m}, \quad m = T/b.$$

- Relationship $mb = T$ reflects a balance between:
  - number of periodogram ordinates,
  - number of autocovariances used.
- Note: A PSD frequency-domain kernel can yield negative weights in time domain.

Bias-Variance Trade-Off in Time and Frequency Domains

- Frequency domain:

$$\widehat{\Omega}^{WP} = 2\pi \sum_{l=-b}^{b} K(l)\, \widehat{I}_v\left(\frac{2\pi l}{T}\right).$$

  - Larger b: more smoothing, lower variance, but higher bias (broader window includes more noise).
- Time domain:

$$\widehat{\Omega}^{SC} = \sum_{j=-m}^{m} k(j)\, \widehat{\gamma}(j).$$

  - Larger m: use more autocovariances, lower bias, but higher variance.
- Choosing b (or m) is central to balancing bias and variance.

MSE Criterion and Andrews' Rule

- Ideal HAC estimator is:
  - PSD w.p.1,
  - consistent,
  - MSE-minimizing:

$$\text{MSE}(\widehat{\mathbf{\Omega}}) = \text{E}[(\widehat{\mathbf{\Omega}} - \mathbf{\Omega})^2] = \text{Bias}^2 + \text{Var}.$$

- This implies bandwidth m increasing with T but at a sufficiently slow rate.
- Andrews (1991) recommend

$$m = 0.75\, T^{1/3},$$

under AR(1) assumptions (coefficient 0.5) and χ^2 asymptotic critical values.

Size Distortions and Inference Focus

- Even with optimal MSE bandwidth, HAC-based tests can exhibit poor size when errors are serially correlated, especially in small/moderate samples (Den Haan and Levin 1996).
- For inference, we care about:
  - Test size / coverage of confidence intervals,
  - which depend more on *bias* than on bias^2.
- Sun, Phillips, and Jin (2008) emphasize that size control and coverage motivate different bandwidth choices than pure MSE minimization.
- In practice, we often:
  - Use fewer periodogram ordinates (smaller b),
  - Use more autocovariances (larger m),
  - which increases variance and can worsen size in finite samples.

Motivation for Inconsistent LRV Estimators

- One strategy: use an **inconsistent** LRV estimator that is stochastically proportional to the true LRV.
- This can yield **pivotal** statistics whose limiting distribution does not depend on nuisance parameters.
- Critical values can then be obtained via Monte Carlo.
- Fixed-b and self-normalization approaches fall into this class and provide alternative routes to robust HAR inference.
- We will discuss the fixed-b approach in detail.

Table of Contents

Motivation: Limitations of HAC LRV Estimators

- Andrews (1991): a "good" LRV estimator should
  1. be a.s. positive semi-definite (PSD);
  2. be HAC (consistent);
  3. minimize MSE:

$$\mathrm{MSE}(\widehat{\Sigma}) = \mathrm{E}[(\widehat{\Sigma} - \Sigma)^2] = \mathrm{bias}(\widehat{\Sigma})^2 + \mathrm{Var}(\widehat{\Sigma}).$$

- Implies bandwidth M increases with n but at a slower rate.
- Empirical studies: such estimators can yield **distorted test sizes** under moderate dependence.
- Reason: bandwidth rules are driven by MSE, not by size/coverage properties.

Two Approaches to Fix Size Distortions

- **Approach 1: keep LRV consistency, improve approximations**
  - Edgeworth expansions for smoothed spectral estimators (Velasco and Robinson 2001).
  - **Fixed-b** asymptotics (Kiefer and Vogelsang 2005):
    - Treat $b \in (0, 1]$ as *fixed*, not $b \to 0$.
    - More accurate first-order approximations.
    - Enables local power analysis for HAC-robust tests.
  - Evidence: improved finite-sample size (Kiefer, Vogelsang, and Bunzel 2000; Kiefer and Vogelsang 2002a; Kiefer and Vogelsang 2002b; Kiefer and Vogelsang 2005).
- **Approach 2: abandon consistency, use pivotal limits**
  - Use LRV estimators that are *inconsistent* but stochastically proportional.
  - Critical values adjusted for extra variance (often via simulation).
  - Fixed-b can also be interpreted this way.

Fixed-b as a Pivotal Approach

- Under fixed-b, limiting distributions depend on b but not on nuisance LRV parameters.
- Inference based on pivotal limits ((Kiefer and Vogelsang 2002b; Kiefer and Vogelsang 2005)):
  - Critical values obtained by Monte Carlo.
  - Often outperform χ^2 approximations.
- Lazarus et al. (2018) and Lazarus, Lewis, and Stock (2021):
  - Propose rules for choosing b via loss functions balancing size and power.
  - Characterize size–power trade-off frontiers.
- Like block bootstrap or subsampling (Kunsch 1989; Liu 1992; Politis, Romano, and Wolf 1999), fixed-b also needs tuning parameters.

Self-normalization and Adjusted Range

- Shao (2010) and Shao and Zhang (2010) self-normalization:
  - Often described as "better size but lower power" (Shao 2010; Zhang et al. 2011; Wang and Shao 2020).
  - Uses variance-of-partial-sums type normalizers.
- Fixed-b is a compromise: gives up some power to improve size when LRV is hard to estimate consistently.
- Hong et al. (2024) propose **adjusted-range-based** self-normalization:
  - Uses range of partial sums rather than variance.
  - Range has strong convergence properties even under infinite variance (Mandelbrot 1972; Mandelbrot 1975).
  - Achieves a better size–power trade-off.
- Extended to KS-type statistics, censored data, autocorrelation tests (Sun et al. 2022; Sun, Zhu, and Linton 2025).

Agenda for this Section

- **Fixed-b approach**
  - Regression setup and assumptions.
  - Fixed-b LRV estimator and its limit.
- **Self-normalization**
  - Kiefer (2000) setting.
  - Shao (2010) self-normalization:
    - Mean and approximately linear statistics.
  - Adjusted-range-based self-normalization (Hong et al. 2024).

Regression Setup and OLS

- Linear regression model:

$$Y_t = X_t'\beta + u_t, \quad t = 1, \ldots, T,$$

where

  - β is $q \times 1$,
  - X_t is $q \times 1$,
  - u_t may be heteroskedastic and autocorrelated, with $\mathrm{E}(u_t \mid X_t) = 0$.

- OLS estimator:

$$\widehat{\beta} = \Big(\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \sum_{t=1}^{T} X_t Y_t.$$

- Rewrite:

$$\sqrt{T}(\widehat{\beta} - \beta_0) = \Big(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t\Big),$$

for testing $\beta = \beta_0$ or building CIs.

Partial Sums and Long-run Variance

- Let $v_t = X_t u_t$ and $S_{\lfloor rT \rfloor} = \sum_{t=1}^{\lfloor rT \rfloor} v_t$.
- Then

$$T^{1/2}(\widehat{\beta} - \beta_0) = \left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)^{-1} T^{-1/2} S_T.$$

- Variance of $T^{-1/2}S_T$:

$$\mathrm{Var}(T^{-1/2}S_T) = \gamma(0) + \sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)\{\gamma(j) + \gamma(-j)\},$$

where $\gamma(j) = \mathrm{E}(v_t v_{t-j}')$.

- Long-run variance:

$$\mathbf{\Omega}_v = \gamma(0) + \sum_{j=1}^{\infty}\{\gamma(j) + \gamma(-j)\} = \Lambda\Lambda',$$

with Λ lower triangular (Cholesky factor).

Assumption 1: Functional CLT

Assumption 1 (Kiefer, Vogelsang, and Bunzel 2000)

$$T^{-1/2} S_{\lfloor sT \rfloor} \Longrightarrow \Lambda \mathbf{B}_q(s),$$

where $\mathbf{B}_q(\cdot)$ is q-dimensional Brownian motion.

- This is a **functional central limit theorem** (FCLT) for partial sums of v_t.
- Ensures that the process of partial sums converges to a continuous Gaussian process.
- See Kiefer, Vogelsang, and Bunzel (2000) for details.

Assumption 2: Regressor Design

Assumption 2 (Kiefer, Vogelsang, and Bunzel 2000)

$$T^{-1}\sum_{t=1}^{\lfloor sT \rfloor} X_t X_t' \xrightarrow{p} sQ$$

uniformly in $s \in [0,1]$ as $T \to \infty$, with Q positive definite.

- The normalized cumulative sum of $X_t X_t'$ behaves like sQ.
- Uniform convergence over s ensures good behavior for all subsamples.
- Q is the long-run second moment matrix of regressors.

Asymptotic Normality of OLS

Under Assumptions 1 and 2,

$$\sqrt{T}(\widehat{\beta}-\beta_0) \Longrightarrow Q^{-1}\Lambda \mathbf{B}_q(1) \sim N\big(0, Q^{-1}\mathbf{\Omega}_v Q^{-1}\big) = N(0, \mathbf{\Omega}_\beta).$$

- Q can be consistently estimated by

$$\widehat{Q} = T^{-1}\sum_{t=1}^{T} X_t X_t'.$$

- Difficulty: consistent estimation of $\mathbf{\Omega}_v$.
- HAC LRV estimators are typically plugged into

$$\widehat{\mathbf{\Omega}}_\beta = \widehat{Q}^{-1}\widehat{\mathbf{\Omega}}_v\widehat{Q}^{-1},$$

but may perform poorly in finite samples with dependence.

Kernel-based HAC LRV Estimator

- For heteroskedastic and autocorrelated errors, use:

$$\widehat{\Omega}_v^{(b)} = \sum_{j=-(T-1)}^{T-1} k(j/m)\,\widehat{\gamma}(j),$$

where

$$\widehat{\gamma}(j) = \frac{1}{T}\sum_{t=1}^{T} v_t v_{t-j}', \quad \widehat{\gamma}(j) = \widehat{\gamma}(-j)' \text{ for } j < 0.$$

- Kernel k satisfies:

$$k : \mathbb{R} \to \mathbb{R}, \quad k(x) = k(-x), \quad k(0) = 1, \quad |k(x)| \le 1,$$

with k continuous at 0 and $\int k^2(x)\,dx < \infty$.

- Bandwidth m, with $b = m/T$, is the key tuning parameter (fixed in fixed-b asymptotics).

Small-b vs Fixed-b Asymptotics

- Standard ("small-b") asymptotics:

$$m \to \infty, \quad b = m/T \to 0 \quad \text{as } T \to \infty.$$

- Kiefer and Vogelsang (2005) argue:
  - In practice, b is a fixed positive fraction of T.
  - The limit $b \to 0$ is unrealistic for applied HAC.
- **Fixed-b asymptotics**:
  - Treat $b \in (0, 1]$ as fixed in the asymptotic analysis.
  - Limiting distributions depend on b explicitly.
  - HAC estimators are *stochastically proportional* to the true LRV, not consistent.

Random Matrix $Q_i(b)$ under Fixed-b

Definition (Bartlett Kernel Case)

Let $K(x)$ be the Bartlett kernel. For $i \times i$ Brownian bridge $\mathbb{B}_i(s)$, define

$$\begin{aligned} \mathrm{Q}_i(b) = & \frac{2}{b} \int_0^1 \mathbb{B}_i(s) \mathbb{B}_i(s)' \, ds \\ & - \frac{1}{b} \int_0^{1-b} \left\{ \mathbb{B}_i(s+b) \mathbb{B}_i(s)' + \mathbb{B}_i(s) \mathbb{B}_i(s+b)' \right\} ds. \end{aligned}$$

- $\mathrm{Q}_i(b)$ summarizes the fixed-b limit of the HAC estimator for an i-dimensional process.
- General forms for other kernels given in Kiefer and Vogelsang (2005).

Limit of the Fixed-b HAC Estimator

- Under Assumptions 1 and 2, as $T \to \infty$,

$$\widehat{\mathbf{\Omega}}_V^{(b)} \Longrightarrow \Lambda \mathbf{Q}_q(b) \Lambda'.$$

- So $\widehat{\mathbf{\Omega}}_V^{(b)}$ is a random matrix stochastically proportional to $\mathbf{\Omega}_V$ via Λ and Λ'.
- Using fixed-b limits:
  - We do not get a consistent estimator of $\mathbf{\Omega}_V$.
  - But we get a pivotal limit for suitably self-normalized test statistics (with b-dependent critical values).

Kiefer (2000): Asymptotic Normality Revisited

- Same regression model and OLS estimator as before.
- Under Assumptions 1 and 2:

$$T^{1/2}(\widehat{\beta} - \beta) \Longrightarrow Q^{-1}\Lambda \mathbf{B}_q(1) \sim N\big(0, Q^{-1}\mathbf{\Omega}_v Q^{-1}\big) = N(0, V).$$

- Usual robust inference:
  - Estimate $V = Q^{-1}\mathbf{\Omega}_v Q^{-1}$ via HAC.
  - Construct t- or Wald-type tests using $\widehat{V}$.
- Problem: quality of inference hinges on HAC choice and finite-sample reliability of $\widehat{V}$.

Kiefer's Self-normalization Construction

- Start with a HAC estimator $\widehat{\Omega}_v$ and its Cholesky factor:

$$\widehat{\Omega}_v = \widehat{\Lambda}\widehat{\Lambda}'.$$

- Define

$$\widehat{V} = \left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \widehat{\Omega}_v \left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)^{-1}.$$

- Then

$$\widehat{V}^{-1/2} = \left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \widehat{\Lambda}.$$

- Transformation:

$$\widehat{V}^{-1/2} T^{1/2}(\widehat{\beta} - \beta) \Longrightarrow N(0, I_q),$$

under standard HAC-type assumptions.

Self-normalized Partial Sums

- Define

$$\widehat{S}_t = \sum_{j=1}^{t} X_j \widehat{u}_j,$$

and the matrix

$$\widehat{C} = T^{-2} \sum_{t=1}^{T} \widehat{S}_t \widehat{S}_t'.$$

- Under Assumptions 1 and 2:

$$T^{-1/2} \widehat{S}_{\lfloor sT \rfloor} \Longrightarrow \Lambda \mathbb{B}_q(s),$$

where $\mathbb{B}_q$ is Brownian bridge.

- Hence

$$\widehat{C} \Longrightarrow \Lambda \Big[\int_0^1 \mathbb{B}_q(s) \mathbb{B}_q(s)' ds \Big] \Lambda'.$$

- Denote $P_q = \int_0^1 \mathbb{B}_q(s) \mathbb{B}_q(s)' ds$ and write $P_q = Z_q Z_q'$ (Cholesky).

Limit Distribution of Kiefer's Self-normalized Statistic

- Define

$$\widehat{B} = \Big(T^{-1}\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \widehat{C} \Big(T^{-1}\sum_{t=1}^{T} X_t X_t'\Big)^{-1},$$

and

$$\widehat{M} = \Big(T^{-1}\sum_{t=1}^{T} X_t X_t'\Big)^{-1} \widehat{C}^{1/2}.$$

- Then

$$\widehat{M}^{-1} T^{1/2}(\widehat{\beta} - \beta) \Longrightarrow Z_q^{-1}\mathbf{B}_q(1).$$

- P_q and $\mathbf{B}_q(1)$ are independent, so conditional on P_q,

$$Z_q^{-1}\mathbf{B}_q(1) \sim N(0, P_q^{-1}).$$

- Unconditional distribution is a **mixture of normals** with heavier tails than $N(0, I_q)$.

Kiefer's t^* Statistic

- t^*: t-type statistic based on self-normalizer $\widehat{B}$.
- Limit distribution of t^*:

$$t^* \Longrightarrow \frac{B(1)}{\left[\int_0^1 \mathbb{B}(s)^2 ds\right]^{1/2}},$$

  where B and $\mathbb{B}$ are 1-dim Brownian motion/bridge.
- Critical values computed by Monte Carlo (Kiefer, Vogelsang, and Bunzel (2000); see their Table I).
- Symmetric, heavy-tailed distribution compared to $N(0,1)$.

Classical t Statistic as Self-normalization

- IID Gaussian case: $X_i \sim N(\mu, \sigma^2)$.
- CLT:

$$T^{1/2}(\overline{X} - \mu) \Longrightarrow N(0, \sigma^2).$$

- Exact finite-sample result:

$$T^{1/2} S^{-1}(\overline{X} - \mu) \sim t_{T-1},$$

where

$$S^2 = \frac{1}{T-1} \sum_{i=1}^{T} (X_i - \overline{X})^2.$$

- Here S^2 is not consistent for σ^2 when T is small, but is *stochastically proportional* to σ^2.
- The t-statistic is a classical self-normalized statistic.

Shao (2010): Self-normalization for Means

- Time series with mean μ and dependence/outliers:

$$X_t = \mu + \varepsilon_t, \quad t = 1, \ldots, T.$$

- Self-normalizer:

$$W_T^2 = T^{-2} \sum_{j=1}^{T} j^2 (\overline{X}_j - \overline{X}_T)^2,$$

where $\overline{X}_j$ is the partial sample mean.

- Under suitable mixing and moment conditions:

$$W_T^2 \Longrightarrow \sigma^2 \int_0^1 \mathbb{B}(s)^2 ds.$$

- W_T^2 is inconsistent for the LRV σ^2 but is stochastically proportional to it.

Regularity Conditions for Shao's Self-normalization

Assumptions (following Phillips (1987))

1. $\mathrm{E}(\varepsilon_t) = 0$.
2. Uniform (2β)-moment bound: $\sup_t \mathrm{E}|\varepsilon_t|^{2\beta} < \infty$ for some $\beta > 2$.
3. Long-run variance exists and is finite:

$$0 < \sigma^2 = \lim_{T\to\infty} \mathrm{E}\Big[T^{-1}\Big(\sum_{t=1}^{T} \varepsilon_t\Big)^2\Big] < \infty.$$

4. Strong-mixing with coefficients $\gamma(k)$ satisfying

$$\sum_{k=1}^{\infty} \gamma(k)^{1-2/\beta} < \infty.$$

- These conditions control outliers and dependence jointly.

Limit Distribution of Shao's Statistic

- Define
$$\mathrm{S} := T^{1/2} W_T^{-1} (\overline{X} - \mu).$$
- Under the previous assumptions,
$$\mathrm{S} \Longrightarrow \frac{B(1)}{\left[\int_0^1 \mathbb{B}(s)^2 ds\right]^{1/2}}.$$
- Same limiting distribution as Kiefer's t^*.
- Critical values obtained via Monte Carlo:
  - Simulate Brownian motions with discrete Gaussian increments.
  - Approximate integrals via Riemann sums.

Approximately Linear Statistics

- Let $\mathbf{F}^m$ be the m-dimensional marginal distribution of $\{X_t\}$; set

$$\mathbf{Y}_t = (X_t, \ldots, X_{t+m-1})', \quad t = 1, \ldots, N = T - m + 1.$$

- Quantity of interest:

$$\theta_t = \mathbf{T}(\mathbf{F}_t^m) \in \mathbb{R}^q,$$

  where $\mathbf{T}$ is a functional.
- Examples of θ:
  - Mean, variance, autocorrelation function,
  - Quantiles, regression coefficients, etc.
- For empirical distribution ρ_{t_1,t_2}, define

$$\widehat{\theta}_{t_1,t_2} = \mathbf{T}(\rho_{t_1,t_2}).$$

Influence Function Expansion

- Approximately linear statistic:

$$\mathbf{T}(\rho_{1,N}) = \mathbf{T}(\mathbf{F}^m) + N^{-1} \sum_{t=1}^{N} \mathbf{IF}(\mathbf{Y}_t; \mathbf{F}^m) + \mathbf{R}_{1,N}.$$

- Influence function:

$$\mathbf{IF}(\mathbf{y}; \mathbf{F}^m) = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}\{(1-\epsilon)\mathbf{F}^m + \epsilon\delta_{\mathbf{y}}\} - \mathbf{T}(\mathbf{F}^m)}{\epsilon}.$$

- For sub-blocks $[t_1, t_2]$:

$$\mathbf{T}(\rho_{t_1,t_2}) = \mathbf{T}(\mathbf{F}^m) + (t_2 - t_1 + 1)^{-1} \sum_{t=t_1}^{t_2} \mathbf{IF}(\mathbf{Y}_t; \mathbf{F}^m) + \mathbf{R}_{t_1,t_2}.$$

CLT for Approximately Linear Statistics

Assumption (Shao–Zhang 3.1)

$$\mathrm{E}\{\mathbf{IF}(\mathbf{Y}_t; \mathbf{F}^m)\} = 0$$

and

$$N^{-1/2} \sum_{t=1}^{\lfloor sN \rfloor} \mathbf{IF}(\mathbf{Y}_t; \mathbf{F}^m) \Longrightarrow \Delta \mathbf{B}_q(s),$$

with Δ lower triangular and $\Delta\Delta' = \mathbf{\Omega}(\mathbf{F}^m)$ positive definite.

Assumption (Shao–Zhang 3.2)

$$\sup_{1 \le k \le N} |k\mathbf{R}_{1,k}| = o_p(N^{1/2}), \qquad \sup_{1 \le k \le N} |k\mathbf{R}_{1,N-k+1}| = o_p(N^{1/2}).$$

- Then

$$N^{1/2}(\widehat{\theta}_{1,N} - \theta) \xrightarrow{d} N(0, \mathbf{\Omega}(\mathbf{F}^m)).$$

Self-normalizer for Approximately Linear Statistics

- Define

$$W_N^2 = N^{-2} \sum_{t=1}^{N} t^2 (\widehat{\theta}_t - \widehat{\theta}_N)(\widehat{\theta}_t - \widehat{\theta}_N)'.$$

- Shao and Zhang (2010) show that

$$N(\widehat{\theta}_N - \theta)'(W_N^2)^{-1}(\widehat{\theta}_N - \theta) \xrightarrow{d} U_q,$$

where

$$U_q = \mathbf{B}_q(1)' V_q^{-1} \mathbf{B}_q(1), \quad V_q = \int_0^1 \mathbb{B}_q(s)\mathbb{B}_q(s)' ds.$$

- Critical values for U_q obtained by Monte Carlo (depends only on q).

Upper Critical Values of U_q

Table: Upper critical values of U_q.

$\alpha \backslash q$	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
10.0%	28.2631	70.6532	127.7057	194.9004	280.4514
5.0%	46.1485	105.1023	177.9481	260.4148	358.8780
2.5%	66.8622	141.9688	229.5083	324.9254	446.9538
1.0%	99.6165	197.4371	319.7587	441.8954	555.0002
0.5%	126.9845	231.8552	389.9997	504.2213	658.3059
0.1%	247.4057	357.4201	527.8317	727.7721	844.1875
	$q=6$	$q=7$	$q=8$	$q=9$	$q=10$
10.0%	363.3813	471.2091	578.2518	704.1028	837.7239
5.0%	466.7869	591.6074	719.1749	863.9764	1022.0704
2.5%	588.5568	727.0452	864.1292	1032.3781	1208.3086
1.0%	737.5201	897.4758	1056.9578	1237.0748	1442.0418
0.5%	836.5370	1055.8835	1183.0328	1408.8165	1627.6028
0.1%	1200.8159	1397.8051	1631.4805	1836.0936	2078.1867

Limitations of Variance-based Self-normalization

- Empirically, self-normalized tests based on W_T^2 often show:
  - Good size (close to nominal),
  - But lower power (Shao 2010; Zhang et al. 2011; Wang and Shao 2020).
- Main reason: W_T^2 is a *variance* of a partial sum process, which:
  - reacts strongly to outliers and heavy dependence,
  - may over-stabilize the statistic and flatten its distribution.
- Hong et al. (2024) propose using a **range**-based self-normalizer instead.

Adjusted Range and M Statistic (Univariate)

- For $X_t = \mu + \varepsilon_t$, define the adjusted range

$$R_T = \max_{1\le k\le T} T^{-1/2}\sum_{t=1}^{k}(X_t - \overline{X}) - \min_{1\le k\le T} T^{-1/2}\sum_{t=1}^{k}(X_t - \overline{X}).$$

- Under similar assumptions as before,

$$R_T \Longrightarrow \sigma\Big[\sup_{s\in[0,1]} \mathbb{B}(s) - \inf_{s\in[0,1]} \mathbb{B}(s)\Big].$$

- Define

$$\mathrm{M} := T^{1/2} R_T^{-1}(\overline{X} - \mu) \Longrightarrow \frac{B(1)}{\sup_{s\in[0,1]} \mathbb{B}(s) - \inf_{s\in[0,1]} \mathbb{B}(s)}.$$

- Monte Carlo results: density of M is closer to $N(0,1)$ than that of S, with better size–power balance.

Critical Values for M and S

- Simulate $B(s)$ and $\mathbb{B}(s)$ via normalized sums of IID $N(0,1)$ increments.
- Use many replications (e.g. 10,000) and fine time grids (e.g. 200,000 steps).
- Obtain upper critical values for M and S:
  - For one-sided right-tail tests: levels 5%, 2.5%, 1%, etc.
  - S has much larger critical values (heavier tails).
  - M is more "normal-like", so tests have more power for a given size.

Densities of M and S

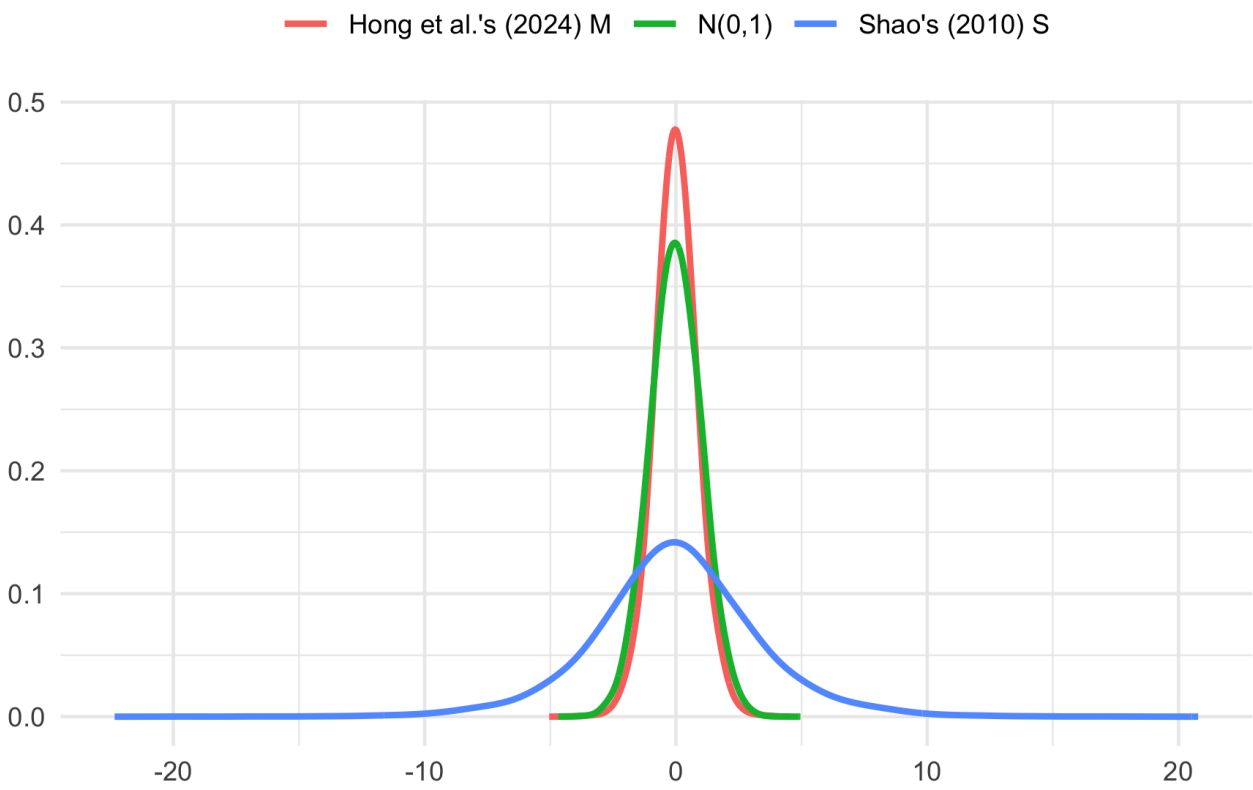


Figure: Densities of M according to Hong et al.'s (2024) adjusted-range-based self-normalization, the standard normal distribution, and S based on Shao's (2010) self-normalization.

Critical Values for M and S

Table: Upper critical values of M and S for one-sided tests (right tail).

Level (α)	Hong et al.'s (2024) M	Shao's (2010) S
5.0%	1.4100	5.3210
2.5%	1.7171	6.7805
1.0%	2.0682	8.5458
0.5%	2.3535	9.7029
0.1%	3.0706	13.0598

Adjusted Range for Approximately Linear Statistics (q = 1)

- When $q = 1$ (scalar parameter), $T = N$ and we can define

$$R_T = \max_{1 \le k \le T} T^{-1/2} \sum_{j=1}^{k} j(\widehat{\theta}_{1,j} - \widehat{\theta}_{1,T}) - \min_{1 \le k \le T} T^{-1/2} \sum_{j=1}^{k} j(\widehat{\theta}_{1,j} - \widehat{\theta}_{1,T}).$$

- This parallels the mean case but uses the approximately linear statistic $\widehat{\theta}_{1,j}$.
- Self-normalized statistic:

$$T^{1/2} R_T^{-1} (\widehat{\theta}_{1,T} - \theta_0)$$

has limiting distribution of the same type as M.

Multivariate Case and Partial Prewhitening

- When $q \geq 2$, cross-dependence complicates range-based normalization.
- Let

$$\mathbf{\Omega}(\mathbf{F}^m) = L_\theta D_\theta L_\theta',$$

be the LDL decomposition (square-root-free Cholesky).

- Use consistent estimator $\widehat{\mathbf{\Omega}}_\theta$ and decompose:

$$\widehat{\mathbf{\Omega}}_\theta = \widehat{L}_\theta \widehat{D}_\theta \widehat{L}_\theta'.$$

- Set $\widehat{C}_\theta = \widehat{L}_\theta$ and define the transformed series

$$\widehat{\theta}^*_{1,j} = \widehat{C}_\theta^{-1}(\widehat{\theta}_{1,j} - \widehat{\theta}_{1,N}).$$

- This "partial prewhitening" removes cross-dependence but preserves temporal dependence.

Range-based Normalizer for q-dimensional θ

- For each component $i = 1, \ldots, q$, define

$$R_N^{(i)} = \max_{1 \le k \le N} N^{-1/2} \sum_{j=1}^{k} j\widehat{\theta}_{1,j}^{*(i)} - \min_{1 \le k \le N} N^{-1/2} \sum_{j=1}^{k} j\widehat{\theta}_{1,j}^{*(i)}.$$

- Construct diagonal matrix

$$\widetilde{R}_N = \mathrm{diag}\big(R_N^{(1)}, \ldots, R_N^{(q)}\big).$$

- Adjusted-range self-normalizer:

$$V_N^R = \widehat{C}_\theta' \, \widetilde{R}_N^2 \, \widehat{C}_\theta.$$

- As $N \to \infty$,

$$(V_N^R)^{1/2} \Longrightarrow \mathbf{\Omega}(\mathbf{F}^m)^{1/2} \mathrm{R}_q,$$

where R_q collects componentwise ranges of Brownian bridges.

Limiting Distribution M_q^2

- Define

$$\mathrm{R}_q = \mathrm{diag}\Big[\sup_{r\in[0,1]} \mathbb{B}_q(r) - \inf_{r\in[0,1]} \mathbb{B}_q(r)\Big],$$

  where $\mathbb{B}_q$ is q-dimensional Brownian bridge.

- Set

$$\mathrm{M}_q^2 \stackrel{d}{=} \mathbf{B}_q(1)'\mathrm{R}_q^{-2}\mathbf{B}_q(1).$$

- Under Assumptions 3.1 and 3.2,

$$N(\widehat{\theta}_{1,N} - \theta_0)'(V_N^R)^{-1}(\widehat{\theta}_{1,N} - \theta_0) \Longrightarrow \mathrm{M}_q^2.$$

- Critical values for M_q^2 obtained by Monte Carlo; distribution becomes more dispersed as q increases.

Upper Critical Values of M_q^2

Table: Upper critical values of M_q^2.

$\alpha \backslash q$	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
10.0%	2.0779	3.4832	4.7167	5.8582	6.9952
5.0%	3.1079	4.6657	6.2193	7.3920	8.5455
2.5%	4.1448	5.8934	7.6381	8.8646	10.1399
1.0%	5.6191	7.7168	9.6347	10.9642	12.4322
0.5%	6.8637	9.4643	11.1350	12.5296	14.1282
0.1%	11.3016	13.2014	15.0835	16.2498	18.5210
	$q=6$	$q=7$	$q=8$	$q=9$	$q=10$
10.0%	8.0949	9.1488	10.2415	11.1577	12.3157
5.0%	9.8275	10.9115	12.1601	13.1386	14.3693
2.5%	11.7286	12.7780	14.4374	15.2984	16.5232
1.0%	14.4145	15.3810	17.1697	18.3134	19.5516
0.5%	16.1673	17.4250	19.3072	20.4066	21.7174
0.1%	20.8818	21.9212	23.6990	24.8022	25.4217

Applications and Advantages

- Hong et al. (2024) apply adjusted-range self-normalization to:
  - Structural breaks in means of approximately linear statistics.
  - Changes in correlations and correlation matrices.
  - Constancy of regression parameters over time.
- Simulation evidence:
  - Improves finite-sample performance of HAC-based tests (Müller 2007).
  - Helps fix non-monotonic power issues without forward/backward summations.
- Potential applications:
  - Confidence interval construction.
  - Sequential detection of structural changes.
  - Other forms of robust time-series inference.

Table of Contents

Bootstrap as HAR Inference

- Bootstrap is a HAR technique because it:
  - Avoids strong parametric assumptions on the error structure.
  - Can accommodate heteroskedasticity and serial correlation.
- Idea:
  - Resample data (residuals or blocks) in a way that mimics dependence.
  - Re-estimate the model repeatedly.
  - Use the empirical distribution of the statistic for inference.
- Particularly valuable for time series and econometrics where robust inference is essential.

Bootstrap: Basic Setup

- Data: i.i.d. sample $y_1, \ldots, y_T$ from unknown F.
- Statistic (or root): $R_T(\tau; y_1, \ldots, y_T)$.
- Target c.d.f.:
$$H_T(x, F) = \Pr(R_T \leq x).$$
- Standard asymptotics:
  - $H_T(x, F) \to H(x, F)$ as $T \to \infty$.
  - Replace F by estimate F_T:
$$\widehat{H}_A(x) = H(x, F_T).$$
- Example: $R_T = \sqrt{T}(\bar{y} - \mu)$ with asymptotic $N(0, \sigma^2)$, or pivotal version
$$R_T = \frac{\sqrt{T}(\bar{y} - \mu)}{\widehat{\sigma}} \xrightarrow{d} N(0, 1).$$

Bootstrap Approximation

- Empirical c.d.f. F_T puts mass $1/T$ on each observation.
- Bootstrap approximation:

$$\widehat{H}_B(x) = H_T(x, F_T).$$

- Implementation:
  1. Compute $\widehat{\tau}$ from original sample.
  2. For $s = 1, \ldots, S$:
     - Draw $y_1^{*(s)}, \ldots, y_T^{*(s)}$ with replacement from $\{y_1, \ldots, y_T\}$.
     - Compute $R_T^{*(s)} = R_T(\widehat{\tau}; y_1^{*(s)}, \ldots, y_T^{*(s)})$.
  3. Empirical bootstrap c.d.f.:

$$\widehat{H}_B(x) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\{R_T^{*(s)} \leq x\}.$$

Bootstrap Critical Values and Consistency

- Bootstrap critical value at level α:

$$\widehat{Q}_B(\alpha) = \widehat{H}_B^{-1}(\alpha).$$

- Classic result (e.g. Bickel and Freedman (1981)):
  - Under suitable conditions, for i.i.d. data with mean μ and variance σ^2:

$$\sup_x \left| \Pr\left(m^{1/2}(\bar{\mu}_m^* - \bar{\mu}) \le x \mid y_1, \ldots, y_T \right) - \Phi(x/\sigma) \right| \to 0,$$

    where $\bar{\mu}_m^*$ is the bootstrap mean.
  - And

$$m \operatorname{Var}(\bar{\mu}_m^* \mid y_1, \ldots, y_T) \to \sigma^2.$$

- Bootstrap consistently mimics finite-sample variability of many statistics.

Bootstrap for Dependent Data

- Standard i.i.d. bootstrap breaks down for time series:
  - Resampling individual observations destroys serial dependence.
- Remedies:
  - **Parametric bootstrap** (fit ARMA etc.).
  - **Wild bootstrap** (for conditional heteroskedasticity).
  - **Block bootstrap** (nonparametric dependence).
  - **Subsampling** (no resampling, rolling windows).

Parametric Bootstrap: ARMA Case

- Suppose
$$A(L)y_t = \mu + B(L)\varepsilon_t,$$
with ε_t i.i.d. mean 0, variance σ^2.
- Root:
$$R_T(\tau; y_1, \ldots, y_T) = \sqrt{T}\big(f(\widehat{\vartheta}, \widehat{M}) - f(\vartheta, M)\big),$$
where $\tau = (\vartheta', M')'$ includes ARMA parameters and shock moments.
- Algorithm (per bootstrap replication):
  1. Estimate $\widehat{\vartheta}$, compute residuals $\widehat{\varepsilon}_t$.
  2. Recenter residuals: $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t - T^{-1}\sum_s \widehat{\varepsilon}_s$.
  3. Sample ε_t^* with replacement from $\{\widetilde{\varepsilon}_t\}$.
  4. Generate y_t^* from fitted ARMA with shocks ε_t^*.
  5. Re-estimate $(\widehat{\vartheta}^*, \widehat{M}^*)$ and compute R_T^*.
- Repeat to build empirical distribution of R_T^*.

Parametric Bootstrap: AR(1) Example

- AR(1):

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad |\phi| < 1,$$

root:

$$R_T = \sqrt{T}(\widehat{\phi} - \phi).$$

- Per replication:
  1. Estimate $\widehat{\phi}$ by OLS on (y_t, y_{t-1}).
  2. Compute residuals $\widehat{\varepsilon}_t = y_t - \widehat{\phi} y_{t-1}$.
  3. Recenter: $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t - \bar{\widehat{\varepsilon}}$.
  4. Sample ε_t^* from $\{\widetilde{\varepsilon}_t\}$.
  5. Generate $y_t^* = \widehat{\phi} y_{t-1}^* + \varepsilon_t^*$.
  6. Estimate $\widehat{\phi}^*$ and compute $R_T^* = \sqrt{T}(\widehat{\phi}^* - \widehat{\phi})$.

Wild Bootstrap

- Purpose: allow conditional heteroskedasticity in ε_t.
- Start from residuals $\widehat{\varepsilon}_t$ of an ARMA (or other) model.
- Generate pseudo-residuals:

$$\varepsilon_t^* = z_t^* \widehat{\varepsilon}_t,$$

with z_t^* i.i.d. s.t. $\mathrm{E}z_t^* = 0$, $\mathrm{Var}z_t^* = 1$.

- Popular choices:
  - Rademacher: $z_t^* = \pm 1$ (probability 0.5 each).
  - Mammen two-point distribution (as in Mammen (1993)), which also controls third moments.
- Preserves

$$\mathrm{E}(\varepsilon_t^* \mid \text{data}) = 0, \quad \mathrm{E}[(\varepsilon_t^*)^2 \mid \text{data}] = \widehat{\varepsilon}_t^2,$$

thereby matching conditional heteroskedasticity.

Block Bootstrap

- Nonparametric method preserving dependence via resampling blocks.
- Partition $\{y_t\}_{t=1}^{T}$ into m blocks of length b:

$$Y_{(1)} = \{y_1, \ldots, y_b\}, \ldots, Y_{(m)} = \{y_{(m-1)b+1}, \ldots, y_T\}.$$

- For each replication:
  1. Sample m blocks with replacement from $\{Y_{(1)}, \ldots, Y_{(m)}\}$.
  2. Concatenate them to form bootstrap series $y_1^*, \ldots, y_T^*$.
  3. Compute statistic and root $R_T^* = \sqrt{T}(f(\widehat{M}^*) - f(\widehat{M}))$.
- Consistent if block length $b(T) \to \infty$ and $b(T)/T \to 0$.

Subsampling

- Alternative without resampling:
  - Slide a window of size b over the data.
  - Compute statistic on each window.
- For $j = 1, \ldots, T - b$:
  - Window: $Y^*_{(j)} = \{y_{j+1}, \ldots, y_{j+b}\}$.
  - Compute $\widehat{M}^*_{(j)}$ and root

$$R^*_{(j)} = \sqrt{b}\big(f(\widehat{M}^*_{(j)}) - f(\widehat{M})\big).$$

- Use empirical distribution of $\{R^*_{(j)}\}$ for inference.
- Again need $b(T) \to \infty$, $b(T)/T \to 0$ under weak dependence.

Case Study: S&P 500 Returns

- Data: S&P 500 daily returns (2000–2024).
- Illustrations:
  - Mean return:

$$\sqrt{T}\,\bar{r} \quad \text{vs.} \quad \sqrt{b}(\bar{r}^* - \bar{r})$$

    from subsampling windows of length $\approx$ 5 years.
  - Volatility persistence:

$$\sqrt{T}\,\widehat{\rho}(1) \quad \text{vs.} \quad \sqrt{b}\big(\widehat{\rho}^*(1) - \widehat{\rho}(1)\big),$$

    where $\widehat{\rho}(1)$ is lag-1 ACF of squared returns.
- Subsampling and block bootstrap provide robust inference under serial dependence and heteroskedasticity.

Bootstrap for Linear Regression

- Model:

$$y_t = \beta' x_t + \varepsilon_t,$$

  with stationary, mixing errors and $\mathrm{E}(\varepsilon_t \mid X) = 0$.
- Two bootstrap setups:
  1. Strongly exogenous regressors (only errors random).
  2. Stationary, mixing regressors (both x_t and ε_t dependent).

Bootstrap with Strongly Exogenous Regressors

- Assumption: x_t are fixed (or independent of error shocks).
- Algorithm:
  1. Compute OLS $\widehat{\beta}$ and residuals $\widehat{\varepsilon}_t = y_t - \widehat{\beta}' x_t$.
  2. If no intercept: recenter $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t - \overline{\widehat{\varepsilon}}$; otherwise set $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t$.
  3. Resample ε_t^* (e.g. i.i.d., blocks, wild).
  4. Generate $y_t^* = \widehat{\beta}' x_t + \varepsilon_t^*$.
  5. Compute $\widehat{\beta}^*$ from (x_t, y_t^*).
  6. Use conditional distribution of $\widehat{\beta}^* - \widehat{\beta}$ for inference.

Bootstrap with Stationary and Mixing Regressors

- Now both (x_t, ε_t) are stationary and mixing.
- Algorithm:
  1. Compute OLS $\widehat{\beta}$ and residuals $\widehat{\varepsilon}_t$.
  2. Recenter if needed: $\widetilde{\varepsilon}_t$.
  3. Resample pairs (ε_t^*, x_t^*) from $\{(\widetilde{\varepsilon}_t, x_t)\}$ (possibly via blocks).
  4. Generate $y_t^* = \widehat{\beta}' x_t^* + \varepsilon_t^*$.
  5. Compute $\widehat{\beta}^*$ from (x_t^*, y_t^*).
  6. Use conditional distribution of $\widehat{\beta}^* - \widehat{\beta}$ for inference on β.

Table of Contents

GMM as HAR Inference

- GMM uses **moment conditions** implied by economic models.
- Robust to heteroskedasticity and autocorrelation via appropriate weighting matrix.
- With optimal weighting matrix (based on HAC LRV of moments), GMM is asymptotically efficient within its class.
- Naturally fits into the HAR framework via HAC-based weighting.

GMM: Setup

- Population moment conditions:

$$G(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)] = 0,$$

  with $\boldsymbol{\theta} \in \mathbb{R}^a$, r moments.
- Sample analogue:

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{y}_T) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t),$$

  where $\mathbf{y}_T = (\mathbf{w}_1', \ldots, \mathbf{w}_T')'$.
- GMM objective:

$$Q(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta})' \mathbf{W}_T \mathbf{g}(\boldsymbol{\theta}),$$

  with positive definite weighting matrix $\mathbf{W}_T$.
- GMM estimator:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}).$$

GMM vs OLS

- OLS:
  - More observations than parameters; use *all* of them.
  - Minimize sum of squared residuals.
- GMM:
  - Often more moments than parameters.
  - Use all moments via quadratic criterion $Q(\theta)$.
  - Weight moments using $\mathbf{W}_T$:
    - Moments with larger variance get less weight.
    - Optimal: $\mathbf{W}_T \approx \Omega^{-1}$, inverse of long-run variance of $\sqrt{T}\,\mathbf{g}(\theta)$.

Asymptotic Distribution of the GMM Estimator

- Under regularity conditions:
  - Identification: $G(\theta) = 0$ uniquely at θ_0.
  - Uniform LLN for $\mathbf{g}(\theta)$.
  - Differentiability of G and $\mathbf{g}$ near θ_0.
  - $\mathbf{W}_T \xrightarrow{p} \mathbf{W}$ (positive definite).
- Then:

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0,$$

and

$$\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\big(0, V(\mathbf{W})\big),$$

where

$$V(\mathbf{W}) = (\Gamma'\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W}\Omega\mathbf{W}\Gamma(\Gamma'\mathbf{W}\Gamma)^{-1}.$$

- $\Gamma = \partial G(\theta)/\partial\theta'|_{\theta_0}$, $\Omega = \lim_{T\to\infty} \mathrm{Var}(\sqrt{T}\,\mathbf{g}(\theta_0))$.

Optimal Weighting and HAC Estimation

- **Optimal W**: $\mathbf{W}_{\text{opt}} = \Omega^{-1}$.
- In practice:
  - Estimate Ω via HAC estimator:

$$\widehat{\Omega} = \frac{1}{T} \sum_{|t-s| \le n(T)} K(|t-s|)\mathbf{h}(\widehat{\theta}, \mathbf{w}_t)\mathbf{h}(\widehat{\theta}, \mathbf{w}_s)'.$$

  - $K(\cdot)$ kernel, $n(T) \to \infty$ slowly.
- Standard errors:

$$\widehat{V} = V(\widehat{\Omega}^{-1}), \quad \text{CI for } \theta_i : \widehat{\theta}_i \pm z_{\alpha/2}\sqrt{\widehat{V}_{ii}/T}.$$

- Over-identifying restrictions test:

$$J_T = T \cdot Q(\widehat{\theta}) \xrightarrow{d} \chi^2(r-a).$$

C–CAPM: Setup

- Representative agent solves

$$\max_{C_t} \mathrm{E}_t\Big[\sum_{i=0}^{\infty} \beta^i u(C_{t+i})\Big],$$

with CRRA utility:

$$u(C_t) = \begin{cases} \dfrac{C_t^{1-\gamma}}{1-\gamma}, & \gamma > 0, \\ \log C_t, & \gamma = 1. \end{cases}$$

- Euler equation for asset i:

$$\mathrm{E}_t\Big[\beta(1+R_{i,t+1})\Big(\frac{C_{t+1}}{C_t}\Big)^{-\gamma} - 1\Big] = 0.$$

- Stochastic discount factor / IMRS:

$$M_{t+1} = \beta\Big(\frac{C_{t+1}}{C_t}\Big)^{-\gamma}.$$

C–CAPM GMM Moment Conditions

- Multiply Euler equation by instruments $\mathbf{x}_t \subseteq \mathcal{F}_t$:

$$\mathrm{E}\Big[\{1-\beta(1+R_{i,t+1})(C_{t+1}/C_t)^{-\gamma}\}\,\mathbf{x}_t\Big] = 0.$$

- Stack across N assets and M instruments:

$$\mathbf{h}(\theta, \mathbf{w}_{t+1}) = \begin{bmatrix} \{1-\beta(1+R_{1,t+1})(C_{t+1}/C_t)^{-\gamma}\}\mathbf{x}_t \\ \vdots \\ \{1-\beta(1+R_{N,t+1})(C_{t+1}/C_t)^{-\gamma}\}\mathbf{x}_t \end{bmatrix},$$

  with $\theta = (\beta, \gamma)'$.

- GMM estimator minimizes

$$\mathbf{g}(\theta)'\,\widehat{W}_T^{-1}\,\mathbf{g}(\theta),$$

  where $\widehat{W}_T$ estimates the optimal weighting matrix.

Two-step and Iterated GMM

- **Two-step GMM**:
  1. Step 1: use simple $\mathbf{W}_T$ (e.g. identity) to get $\widehat{\theta}^{(1)}$.
  2. Step 2: compute HAC estimate $\widehat{\Omega}$ of LRV of moments at $\widehat{\theta}^{(1)}$; set $\widehat{W}_T = \widehat{\Omega}^{-1}$ and re-estimate:

$$\widehat{\theta}^{(2)} = \arg\min_{\theta} \mathbf{g}(\theta)' \, \widehat{\Omega}^{-1} \mathbf{g}(\theta).$$

- **Iterated GMM**:
  - Re-estimate Ω and θ iteratively until convergence.
  - Gains: possible small-sample improvements.
  - Cost: higher computation; often modest gains in practice.

Empirical C–CAPM and GMM

- Hansen (1982) estimate C–CAPM via GMM:
  - Instruments: lagged consumption growth, lagged returns.
  - Returns: market indices, industry portfolios.
  - Consumption: nondurable + services.
- Typical findings:
  - $\widehat{\beta} \approx 0.99$,
  - $\widehat{\gamma}$ in a moderate range (e.g. ≈ 0.3–1).
- However, J-tests often reject over-identifying restrictions:
  - Suggests model misspecification or missing frictions.
  - Motivates extensions: habit formation, recursive preferences, time-varying risk aversion, etc.

Over-identifying Restrictions and J-test

- Over-identified GMM: $r > a$ (more moments than parameters).
- J-statistic:

$$J_T = T\,\mathbf{g}(\widehat{\theta})'\widehat{W}_T\mathbf{g}(\widehat{\theta}) \xrightarrow{d} \chi^2(r-a).$$

- Null: model and moment conditions are correct.
- Large $J_T \Rightarrow$ reject H_0:
  - Moments are not jointly consistent with data.
  - In C–CAPM, often indicates the simple model cannot match asset-return patterns.

Euler Equation Errors

- For asset i, pricing condition:

$$\mathrm{E}[M_{t+1}(\theta)(1+R_{i,t+1})] = 1.$$

- Sample Euler error:

$$\varepsilon_{i,t+1}(\theta) = M_{t+1}(\theta)R_{i,t+1} - 1.$$

- Even at GMM estimate $\widehat{\theta}$, residual pricing errors typically remain.
- Lettau and Ludvigson (2009):
  - C–CAPM exhibits sizable Euler errors, especially in downturns.
  - Indicates fundamental misspecification of the simple model.

Hansen–Jagannathan Distance

- Hansen and Jagannathan (1997) propose HJ distance as a robust model diagnostic.
- Define pricing error vector:

$$\mathbf{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}[M_t(\theta)\mathbf{R}_t - \mathbf{1}_N],$$

and return covariance:

$$G_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{R}_t\mathbf{R}_t'.$$

- HJ distance:

$$\text{Dist}_T(\theta) = \sqrt{\min_\theta \mathbf{g}_T(\theta)' G_T^{-1}\mathbf{g}_T(\theta)}.$$

Over-identifying Restrictions and J-test

- Geometric view:
  - Measures shortest distance from model SDF to set of SDFs that exactly price returns.
  - Focuses on pricing errors, not volatility of SDF.

Equity Premium Puzzle

- Empirical fact: equities have much higher average returns than risk-free assets.
- U.S. historical data: equity returns $\sim$ 7–9% vs. T-bills $\sim$ 1–3% $\Rightarrow$ equity premium $\sim$ 5–8%.
- Mehra and Prescott (1985):
  - Standard C–CAPM can match this only with $\gamma > 30$.
  - Implausibly high risk aversion relative to observed behavior.
- GMM estimates typically yield γ near 1, worsening the puzzle.

Time-varying Risk Aversion and Discounting

- Allow preference parameters to depend on state variables:

$$\beta_t = \beta(\mathbf{x}_t), \quad \gamma_t = \gamma(\mathbf{x}_t), \quad \mathbf{x}_t \in \mathcal{F}_{t-1}.$$

- Euler equation becomes nonlinear regression:

$$\beta(\mathbf{x}_t)\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma(\mathbf{x}_t)}(1 + R_{i,t+1}) = 1 + \varepsilon_{t+1},$$

  with $\mathrm{E}(\varepsilon_{t+1} \mid \mathcal{F}_t) = 0$.
- Estimate $\beta(\cdot)$, $\gamma(\cdot)$ locally (e.g. local polynomials) or via local GMM.
- This time-varying extension can better match observed equity premia and risk dynamics.

Case Study: GMM Estimation of C–CAPM

- Data sources:
  - FRED: real consumption (e.g. PCE), 3-month T-bill rate.
  - Yahoo Finance: S&P 500 index for market returns.
- Construction:
  - Log consumption growth: $\Delta c_t = \log C_t - \log C_{t-1}$.
  - Excess market return: $r_{m,t+1} - r_{f,t+1}$.
  - Instruments: constant, consumption growth, returns, etc.
- Moments:

$$m_t(\beta, \gamma) = \left[\beta(1 + r_{t+1}) \exp(-\gamma \Delta c_{t+1}) - 1\right] z_t,$$

  stacked across instruments.
- Use `gmm` package in R for two-step GMM estimation.

Empirical Results and Interpretation

- Example output (monthly data since 2000):
  - $\widehat{\beta} \approx 0.983$ (small s.e., highly significant).
  - $\widehat{\gamma} \approx 1.03$ (large s.e., not significant).
- J-test:
  - $J_T \approx 9.61$ with 1 d.f., $p \approx 0.0019$.
  - Reject over-identifying restrictions at 1% level.
- Interpretation:
  - Parameter estimates are plausible, but model is rejected overall.
  - Indicates misspecification: C–CAPM too simple to capture observed returns and consumption dynamics.
  - Motivates richer models (habit, rare disasters, heterogeneous agents, time-varying preferences, etc.).

Table of Contents

Summary of the Chapter

- **Spectral analysis**:
  - Population spectrum, periodogram, and smoothing.
  - Case study: GDP growth spectrum.
- **HAC estimation**:
  - PSD issues, kernels and bandwidth choice, LRV estimation.
- **Fixed-b and self-normalization**:
  - Fixed-b HAC limits, Kiefer–Shao self-normalization.
  - Adjusted-range-based self-normalization (Hong et al. (2024)) with improved size–power trade-off.
- **Bootstrap methods**:
  - Parametric, wild, block bootstrap, subsampling for time series.
  - Bootstrap for regression with different regressor assumptions.
- **GMM**:
  - HAR-consistent GMM with HAC weighting.
  - Application to C–CAPM, J-test, Euler errors, HJ distance, and equity premium puzzle.

Table of Contents

References I

- Andrews, Donald WK (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". In: *Econometrica* 59.3, pp. 817–858.
- Andrews, Donald WK and J Christopher Monahan (1992). "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator". In: *Econometrica* 60.4, pp. 953–966.
- Bickel, Peter J and David A Freedman (1981). "Some asymptotic theory for the bootstrap". In: *The Annals of Statistics* 9.6, pp. 1196–1217.
- Cochrane, D. and G. H. Orcutt (1949). "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms". In: *Journal of the American Statistical Association* 44.245, pp. 32–61. DOI: 10.1080/01621459.1949.10483290.
- Den Haan, Wouter J and Andrew T Levin (1996). *A practitioner's guide to robust covariance matrix estimation*.

References II

- Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: *Econometrica* 50.4, pp. 1029–1054.
- Hansen, Lars Peter and Ravi Jagannathan (1997). "Assessing specification errors in stochastic discount factor models". In: *The Journal of Finance* 52.2, pp. 557–590.
- Hong, Yongmiao et al. (2024). *Kolmogorov-Smirnov type testing for structural breaks: A new adjusted-range based self-normalization approach*.
- Kiefer, Nicholas M. and Timothy J. Vogelsang (2005). "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests". In: *Econometric Theory* 21.6, pp. 1130–1164.
- — (2002a). "Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation". In: *Econometrica* 70.5, pp. 2093–2095.

References III

- Kiefer, Nicholas M. and Timothy J. Vogelsang (2002b). "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size". In: *Econometric Theory* 18.6, pp. 1350–1366.
- Kiefer, Nicholas M, Timothy J Vogelsang, and Helle Bunzel (2000). "Simple robust testing of regression hypotheses". In: *Econometrica* 68.3, pp. 695–714.
- Kunsch, Hans R (1989). "The jackknife and the bootstrap for general stationary observations". In: *The Annals of Statistics* 17.3, pp. 1217–1241.
- Lazarus, Eben, Daniel J Lewis, and James H Stock (2021). "The Size-Power Tradeoff in HAR Inference". In: *Econometrica* 89.5, pp. 2497–2516.
- Lazarus, Eben et al. (2018). "HAR inference: Recommendations for practice". In: *Journal of Business & Economic Statistics* 36.4, pp. 541–559.

References IV

- Lettau, Martin and Sydney C Ludvigson (2009). "Euler equation errors". In: *Review of Economic Dynamics* 12.2, pp. 255–283.
- Liu, Regina Y (1992). "Moving blocks jackknife and bootstrap capture weak dependence". In: *Exploring the limits of bootstrap*.
- Mammen, Enno (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The Annals of Statistics* 21.1, pp. 255–285.
- Mandelbrot, Benoit (1972). "Statistical Methodology for Nonperiodic Cycles: From the Covariance to R/S Analysis". In: *Annals of Economic and Social Measurement* 1.3, pp. 259–290.
- Mandelbrot, Benoit B. (1975). "Limit Theorems on the Self-normalized Range for Weakly and Strongly Dependent Processes". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 33, pp. 271–285.
- Mehra, Rajnish and Edward C Prescott (1985). "The equity premium: A puzzle". In: *Journal of Monetary Economics* 15.2, pp. 145–161.

References V

- Müller, Ulrich K (2007). "A theory of robust long-run variance estimation". In: *Journal of Econometrics* 141.2, pp. 1331–1352.
- Newey, Whitney K and Kenneth D West (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". In: *Econometrica* 55.3, pp. 703–708.
- — (1994). "Automatic lag selection in covariance matrix estimation". In: *The Review of Economic Studies* 61.4, pp. 631–653.
- Phillips, Peter CB (1987). "Time series regression with a unit root". In: *Econometrica* 55.2, pp. 277–301.
- Politis, Dimitris N, Joseph P Romano, and Michael Wolf (1999). *Subsampling*. Springer Science & Business Media.
- Shao, Xiaofeng (2010). "A self-normalized approach to confidence interval construction in time series". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 343–366.

References VI

- Shao, Xiaofeng and Xianyang Zhang (2010). "Testing for Change Points in Time Series". In: *Journal of the American Statistical Association* 105.491, pp. 1228–1240.
- Student (1908). "The probable error of a mean". In: *Biometrika* 6.1, pp. 1–25.
- Sun, Jiajing, Meiting Zhu, and Oliver Linton (2025). "Adjusted-range-based self-normalized autocorrelation tests". In: *Economics Letters* 251, p. 112315. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2025.112315.
- Sun, Jiajing et al. (2022). "Adjusted-range self-normalized confidence interval construction for censored dependent data". In: *Economics Letters* 220, p. 110873.
- Sun, Yixiao, Peter CB Phillips, and Sainan Jin (2008). "Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing". In: *Econometrica* 76.1, pp. 175–194.

References VII

- Velasco, Carlos and Peter M Robinson (2001). "Edgeworth expansions for spectral density estimates and studentized sample mean". In: *Econometric Theory* 17.3, pp. 497–539.
- Wang, Runmin and Xiaofeng Shao (2020). "Hypothesis testing for high-dimensional time series via self-normalization". In: *Annals of Statistics* 48.5, pp. 2728–2758.
- White, Halbert (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *Econometrica* 48.4, pp. 817–838.
- Zhang, Xianyang et al. (2011). "Testing the structural stability of temporally dependent functional observations and application to climate projections". In: *Electronic Journal of Statistics* 5, pp. 1765–1796.

Chapter 7 — Filtering

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

What is Filtering?

- In econometrics, **filtering** = transforming raw data into a form more useful for:
  - signal extraction (trend, cycle, innovations),
  - forecasting,
  - model-based inference.
- Core idea: separate *signal* from *noise* in time series.
- Used heavily in:
  - Preprocessing (detrending, deseasonalization, noise reduction),
  - Modeling (e.g. ARMA residuals, SV models),
  - Economics, finance, engineering, signal processing.

Chapter Roadmap

- **Spectral-domain filters**:
  - Transfer functions, gain, phase; effect on spectra.
  - HP filter, moving averages, band-pass (Baxter–King).
- **Time-domain filters**:
  - Whitening, differencing, exponential smoothing.
- **State-space and Kalman filtering**:
  - Linear Gaussian models, Kalman filter/smoother.
  - Local level model, handling missing data, initialization.
- **Stochastic volatility (SV)**:
  - Log-squared transformation,
  - Gaussian-mixture approximation for non-Gaussian noise.

Whitening and Prewhitening

- ARMA(p, q) model residuals can be viewed as output of a **whitening filter**:

$$\text{(input series)} \xrightarrow{\text{filter}} \text{white noise.}$$

- Whitening removes serial dependence, isolating the *unpredictable* component.
- Important for:
  - Efficient inference (e.g. CLT on innovations),
  - HAC long-run variance (LRV) estimation.
- **VAR prewhitening** andrews1992improved:
  1. Fit VAR to multivariate series;
  2. Use residuals as approximately white;
  3. Apply kernel-based HAC estimator to residuals;
  4. Transform back to original scale via VAR coefficients.

Example: Hodrick–Prescott Filter

- HP filter decomposes Y_t into trend τ_t and cycle:

$$\min_{\{\tau_t\}} \left\{ \sum_{t=1}^{T} (Y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \right]^2 \right\}.$$

- First term: fit to data (small residuals).
- Second term: penalizes changes in trend growth $\Rightarrow$ smooth trend.
- Spectral connection:
  - HP filter acts roughly as a low-pass filter for the trend,
  - and a complementary high-pass filter for the cycle.

Linear Filters: Time-domain Definition

Definition 1 (Linear Filter)

A general linearly filtered time series is given by

$$X_t = \sum_{k=-\infty}^{\infty} \phi_{t,k} Y_{t-k} = \phi_t(L) Y_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\{Y_t\}$ is the input series, $\{X_t\}$ is the output series, and the filter coefficients satisfy

$$\sum_{k=-\infty}^{\infty} |\phi_{t,k}| \leq C < \infty$$

for all t. If the coefficients do not depend on t, i.e. $\phi_{t,k} = \phi_k$, the filter is called *time-invariant*.

- **One-sided (backward) filter**: $\phi_k = 0$ for $k < 0$.
- **Symmetric (two-sided) filter**: $\phi_k = \phi_{-k}$ for all k.
- We focus on time-invariant filters and drop t in $\phi_{t,k}$.

Transfer Function, Gain, and Phase

Apply a time-invariant filter to complex exponential input:

$$\phi(L)e^{i\lambda t} = \sum_{k=-\infty}^{\infty} \phi_k e^{i\lambda(t-k)} = e^{i\lambda t} \sum_{k=-\infty}^{\infty} \phi_k e^{-i\lambda k}.$$

Definition 2 (Transfer Function, Gain, and Phase)

$$B_\phi(\lambda) = \sum_{k=-\infty}^{\infty} \phi_k e^{-i\lambda k} = \phi(z), \quad z = e^{-i\lambda}.$$

$B_\phi(\lambda + 2\pi) = B_\phi(\lambda)$.
Gain and **phase**:

$$g(\lambda) = |B_\phi(\lambda)|^2, \qquad \theta(\lambda) = \tan^{-1} \frac{\Im B_\phi(\lambda)}{\Re B_\phi(\lambda)}.$$

Interpreting Gain and Phase

- Write $B_\phi(\lambda)$ in standard / polar form:

$$B_\phi(\lambda) = \sum_k \phi_k \cos(\lambda k) - i \sum_k \phi_k \sin(\lambda k) = g(\lambda)^{1/2} e^{i\theta(\lambda)}.$$

- Filter response:

$$\phi(L) e^{i\lambda t} = g(\lambda)^{1/2} e^{i(\lambda t + \theta(\lambda))}.$$

- Interpretation:
  - $g(\lambda)^{1/2}$ magnifies / attenuates amplitude at frequency λ.
  - $\theta(\lambda)$ phase-shifts that component.
- Some authors use $\partial\theta(\lambda)/\partial\lambda$ as lag displacement (group delay).

Effect on Power Spectrum

Theorem 3

If $\sum_k \phi_k^2 < \infty$ *and* Y_t *has spectrum* $f_Y(\lambda)$*, then filtered series* $X_t = \phi(L) Y_t$ *has spectrum*

$$f_X(\lambda) = |B_\phi(\lambda)|^2 f_Y(\lambda).$$

- Transfer function links input and output spectra:

$$f_X(\lambda) = g(\lambda)\, f_Y(\lambda).$$

- Inverse relation: for any integrable $B_\phi(\lambda)$,

$$\phi_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} B_\phi(\lambda) e^{-i\lambda k} d\lambda.$$

- Design filters either:
  - in time domain via $\{\phi_k\}$, or
  - in frequency domain via $B_\phi(\lambda)$.

Ideal Filter Types

Filters often aim to shape the spectrum $f_Y(\lambda)$ in specific ways.

- **Tweeter (ideal high-pass)**:

$$g(\lambda) = 0 \quad \text{for } |\lambda| \leq \lambda_0.$$

  Removes low frequencies; keeps high frequencies.

- **Woofer (ideal low-pass)**:

$$g(\lambda) = 0 \quad \text{for } |\lambda| > \lambda_1.$$

  Keeps low frequencies; removes high-frequency noise.

- **Mid-range (band-pass)**:

$$g(\lambda) = 0 \quad \text{for } |\lambda| \notin [\lambda_0, \lambda_1].$$

  Passes only a band of frequencies.

Ideal vs Practical Filters

- Ideal Tweeter/Woofer/Mid-range:
  - Discontinuous spectral windows,
  - Imply infinite-length, non-causal filters in time domain,
  - Depend on infinite past and future $\Rightarrow$ not implementable.
- Problems:
  - Boundary issues at sample start/end,
  - Need future data (non-causal),
  - Numerical instability for finite samples.
- Solution: use **approximate filters**:
  - Finite support, stable, computationally tractable,
  - Good approximation to ideal response in relevant band.

Simple Moving Average (SMA) Filter

Example 4 (Symmetric Moving Average (SMA))

Symmetric moving average:

$$X_t = \frac{1}{2r+1} \sum_{j=-r}^{r} Y_{t-j},$$

with $\phi_j = 1/(2r+1)$ for $|j| \leq r$.

Simple Moving Average (SMA) Filter

- Two-sided, symmetric, zero phase shift.
- Transfer function:

$$B_\phi(\lambda) = \begin{cases} 1, & \lambda = 0, \\ \dfrac{1}{2r+1} \dfrac{\sin((2r+1)\lambda/2)}{\sin(\lambda/2)}, & \lambda \neq 0. \end{cases}$$

- Acts as an **approximate low-pass filter**:
  - Retains low frequencies,
  - Attenuates high-frequency oscillations.

SMA on `AirPassengers` Data

- Data: monthly airline passenger counts (R's `AirPassengers`).
- Apply 12-month SMA (using `stats::filter()` with symmetric window):
  - Smooths within-year seasonality and short-run noise,
  - Highlights longer-run growth trend.
- Plot:
  - Blue: original series (strong seasonality + upward trend),
  - Red: 12-month SMA (smooth trend-cycle).

Differencing Filter as High-pass

Example 5 (Differencing Filter)

The differencing operator:

$$\Delta = 1 - L.$$

Transfer function:

$$B_\Delta(\lambda) = 1 - e^{-i\lambda},$$

gain:

$$g(\lambda) = (1 - \cos\lambda)^2 + \sin^2\lambda = 2(1 - \cos\lambda),$$

phase:

$$\theta(\lambda) = \tan^{-1}\left(\frac{\sin\lambda}{1 - \cos\lambda}\right).$$

Differencing Filter as High-pass

- At $\lambda \approx 0$: $g(\lambda) \approx 0 \Rightarrow$ low frequencies are *removed*.
- At $\lambda \approx \pi$: $g(\pi) = 4 \Rightarrow$ high frequencies retained.
- Hence differencing is a **high-pass filter**.
- Key use: remove trends / unit roots to induce stationarity.

Example: GDP per Capita Differencing

- Use `WDI` package to download US real GDP per capita (`NY.GDP.PCAP.KD`, 1960–2023).
- Construct:
  - Levels of GDP per capita,
  - First differences,
  - Log differences (approx. growth rates).
- Typical pattern:
  - Level: strong upward trend, clear nonstationarity.
  - First difference: large scale, nonstationary variance.
  - Log difference: bounded, often close to stationary.
- Differencing (especially log-difference) is key for many ARMA-type models.

Baxter–King Filter: Business Cycle Extraction

- Baxter–King (BK) filter is a symmetric, linear, time-invariant **band-pass** filter.
- Target: isolate business cycle frequencies (e.g. 2–8 years, or 12–32 quarters/years).
- Let P_l, P_u = lower/upper periodicity bounds (in periods).

$$\overline{\lambda} = \frac{2\pi}{P_l}, \quad \underline{\lambda} = \frac{2\pi}{P_u}.$$

- Ideal band-pass response:

$$B_{\text{ideal}}(\lambda) = \begin{cases} 1, & \underline{\lambda} \le |\lambda| \le \overline{\lambda}, \\ 0, & \text{otherwise.} \end{cases}$$

BK Filter Coefficients

Example 6 (Baxter–King Band-pass Filter)

Approximate ideal band-pass by symmetric finite-order MA with lags $k = \pm 1, \ldots, \pm K$:

$$\phi_k = \frac{\sin(k\overline{\lambda}) - \sin(k\underline{\lambda})}{k\pi}, \quad k = \pm 1, \ldots, \pm K; \quad \phi_0 = \frac{\overline{\lambda} - \underline{\lambda}}{\pi}.$$

- Two-sided: drops first and last K observations.
- Special case:

$$\phi_k = \frac{\sin(k\lambda)}{k\pi}, \quad \phi_0 = \frac{\lambda}{\pi}$$

  corresponds to ideal low-pass (Dirichlet kernel).
- Practical implementation in R: `mFilter::bkfilter()`.

BK Filter on US GDP per Capita (Conceptual)

- Apply BK filter to US GDP per capita (1960–2023):
  - Choose periodic band, e.g. 12–32 years.
  - Extract cyclical component (business-cycle frequencies).
- Plot:
  - Blue: original GDP per capita (trend + cycles).
  - Red: BK cycle (centered band-pass component).
- BK filter:
  - Removes high-frequency noise and low-frequency trend,
  - Retains medium-term cycles for business-cycle analysis.

Exponentially Weighted Moving Average (EWMA)

Example 7 (Exponentially Weighted Moving Average (EWMA))

$$X_t = \alpha X_{t-1} + (1-\alpha) Y_t, \quad 0 < \alpha < 1.$$

Equivalent infinite sum:

$$X_t = (1-\alpha) \sum_{j=0}^{\infty} \alpha^j Y_{t-j}.$$

Transfer function:

$$B_{\text{EWMA}}(\lambda) = \frac{1-\alpha}{1-\alpha e^{-i\lambda}}.$$

Exponentially Weighted Moving Average (EWMA)

- One-sided, exponentially decaying weights on the past.
- Acts as a **low-pass** filter:
  - Smooths high-frequency noise,
  - Retains low-frequency trends.
- Widely used for:
  - Price smoothing,
  - Volatility estimation,
  - Online signal tracking.

Gaussian Filter

Example 8 (Gaussian Filter)

Example (Gaussian Filter). Frequency-domain window:

$$B_G(\lambda) = \exp\left(-\frac{1}{2}\sigma^2\lambda^2\right), \quad \lambda \in [-\pi, \pi],\ \sigma^2 > 0.$$

- Time-domain kernel:
  - Symmetric, Gaussian-shaped, rapidly decaying,
  - Non-causal (uses past and future observations).
- Larger $\sigma^2 \Rightarrow$ stronger attenuation of high frequencies (more smoothing).
- Used for:
  - Trend estimation in macro series,
  - Noise reduction in financial or industrial signals.

Spectral Analysis: Suggested References

- Harvey (1990):
  - ARIMA models, seasonality, forecasting,
  - Integrates spectral methods into econometric modeling.
- Percival and Walden (1993):
  - General spectral analysis with applications,
  - Strong foundation independent of specific field.
- Chatfield (2013):
  - Comprehensive time series analysis overview,
  - Includes spectral techniques and practical considerations.

Table of Contents

Why State-Space Models?

- Many time series features are **unobserved**:
  - Trends, cycles, stochastic volatility, regimes.
- State-space models:
  - Provide a general framework to represent latent dynamics.
  - Naturally accommodate:
    - Missing data, irregular sampling,
    - Time-varying parameters,
    - Multivariate structures.
- Kalman filter:
  - Efficient recursive algorithm for linear Gaussian state-space models,
  - Core tool for estimation, prediction, and smoothing.

General (Possibly Nonlinear) State-Space Model

- **Observation equation**:

$$y_t = H(\alpha_t, \eta_t),$$

where $y_t \in \mathbb{R}^n$, $\alpha_t \in \mathbb{R}^m$, $\eta_t \in \mathbb{R}^n$.

- **State equation**:

$$\alpha_t = F(\alpha_{t-1}, \omega_t),$$

where $\omega_t \in \mathbb{R}^m$ is state noise.

- $H(\cdot)$ and $F(\cdot)$ may be nonlinear $\Rightarrow$ extended / particle filters.

Linear Gaussian Case

- Linear Gaussian state-space model:

$$\alpha_t = F_t \alpha_{t-1} + \omega_t, \quad \omega_t \sim N(0, Q_t),$$

$$y_t = H_t \alpha_t + \eta_t, \quad \eta_t \sim N(0, R_t).$$

- Dimensions:
  - $\alpha_t \in \mathbb{R}^m$, $y_t \in \mathbb{R}^n$,
  - $F_t \in \mathbb{R}^{m \times m}$, $H_t \in \mathbb{R}^{n \times m}$,
  - Q_t, R_t covariance matrices.
- **Process noise** ω_t:
  - Unmodeled dynamics, genuine shocks to the state.
- **Measurement noise** η_t:
  - Observation/measurement error.
- Markov assumption: α_t depends on past only via α_{t-1}.

AR(p) Model in State-Space Form

- AR(p) model:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$$

- Define state vector:

$$\alpha_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}.$$

AR(p) Model in State-Space Form

- State equation:

$$\alpha_t = F\alpha_{t-1} + \omega_t,$$

with

$$F = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \end{pmatrix}, \quad \omega_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

- Observation:

$$y_t = H\alpha_t, \quad H = (1, 0, \ldots, 0), \ \eta_t = 0.$$

MA(1) Model in State-Space Form

- MA(1): $y_t = \varepsilon_t + \theta \varepsilon_{t-1}$.
- State:

$$\alpha_t = \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{pmatrix}.$$

- State equation:

$$\alpha_t = F\alpha_{t-1} + \omega_t, \qquad F = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \omega_t = \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}.$$

- Observation:

$$y_t = H\alpha_t, \quad H = (1, \theta), \ \eta_t = 0.$$

Hamilton's Regime-Switching Model

- Model captures regime changes (e.g. expansions vs recessions):

$$y_t = \mu(s_t) + \sigma(s_t)\eta_t, \quad \eta_t \sim N(0,1).$$

- Regime indicator $s_t \in \{0,1\}$ follows a first-order Markov chain:

$$\Pr(s_t = i \mid s_{t-1} = j) = p_{ij}.$$

- Parameters (mean, variance) depend on regime:
  - Regime-specific means $\mu(s_t)$,
  - Regime-specific volatilities $\sigma(s_t)$.
- Can be embedded in a (nonlinear, non-Gaussian) state-space framework with hidden Markov states.

Inference Tasks in State-Space Models

- **Filtering**: estimate current state given data up to t:

$$f(\alpha_t \mid \mathcal{F}_t).$$

- **Prediction**: forecast future state/observation:

$$f(\alpha_{t+h} \mid \mathcal{F}_t), \quad f(y_{t+h} \mid \mathcal{F}_t).$$

- **Smoothing**: re-estimate past states using all data up to T:

$$f(\alpha_t \mid \mathcal{F}_T), \quad T > t.$$

- One-step-ahead predictive likelihood:

$$f(y_t \mid \mathcal{F}_{t-1}).$$

- Analogy (Tsay 2013):
  - Filtering: understand current word as you read.
  - Prediction: guess the next word.
  - Smoothing: reinterpret an earlier word after finishing the text.

Local Level Model

- Simple stochastic trend model:

$$\alpha_{t+1} = \alpha_t + \eta_t, \quad \eta_t \sim N(0, Q),$$

$$y_t = \alpha_t + \omega_t, \quad \omega_t \sim N(0, R).$$

- α_t = unobserved level; y_t = observed series.
- Q: variance of level innovations (**process noise**), R: variance of measurement noise.
- Matrix form:

$$\alpha_t = F\alpha_{t-1} + \eta_t, \ F = 1; \quad y_t = H\alpha_t + \omega_t, \ H = 1.$$

Conditional Moments in Multivariate Normal

Theorem 9 (Conditional Moments of a Multivariate Normal)

Let (α, y, z) be jointly normal, with nonsingular block covariances Σ_{ww} and $\Sigma_{yz} = 0$. Then

$$\mathrm{E}[\alpha \mid y] = \mu_\alpha + \Sigma_{\alpha y}\Sigma_{yy}^{-1}(y - \mu_y),$$

$$\mathrm{Var}(\alpha \mid y) = \Sigma_{\alpha\alpha} - \Sigma_{\alpha y}\Sigma_{yy}^{-1}\Sigma_{y\alpha},$$

$$\mathrm{E}[\alpha \mid y, z] = \mathrm{E}[\alpha \mid y] + \Sigma_{\alpha z}\Sigma_{zz}^{-1}(z - \mu_z),$$

$$\mathrm{Var}(\alpha \mid y, z) = \mathrm{Var}(\alpha \mid y) - \Sigma_{\alpha z}\Sigma_{zz}^{-1}\Sigma_{z\alpha}.$$

- Kalman filter is essentially repeated application of these formulas with appropriate partitioning of (α_t, y_t).

Local Level: Notation for Kalman Filter

- Information set: $\mathcal{F}_t = \{y_1, \ldots, y_t\}$.
- Conditional mean/variance:

$$\alpha_{t|j} = \mathrm{E}(\alpha_t \mid \mathcal{F}_j), \quad P_{t|j} = \mathrm{Var}(\alpha_t \mid \mathcal{F}_j).$$

- One-step-ahead forecast error (innovation):

$$v_t = y_t - \alpha_{t|t-1},$$

with variance

$$V_t = P_{t|t-1} + R.$$

Kalman Filter: Update and Prediction (Local Level)

Update step (given y_t):

$$K_t = \frac{P_{t|t-1}}{V_t}, \quad v_t = y_t - \alpha_{t|t-1}.$$

$$\alpha_{t|t} = \alpha_{t|t-1} + K_t v_t, \quad P_{t|t} = (1 - K_t) P_{t|t-1}.$$

Prediction step (for $t+1$):

$$\alpha_{t+1|t} = \alpha_{t|t}, \quad P_{t+1|t} = P_{t|t} + Q.$$

- Initial conditions: $\alpha_{1|0}$, $P_{1|0}$.

Deriving Update from Joint Normality

- Joint conditional distribution:

$$\begin{bmatrix} \alpha_t \\ v_t \end{bmatrix} \Big| \mathcal{F}_{t-1} \sim N\left(\begin{bmatrix} \alpha_{t|t-1} \\ 0 \end{bmatrix}, \begin{bmatrix} P_{t|t-1} & P_{t|t-1} \\ P_{t|t-1} & V_t \end{bmatrix} \right).$$

- Apply conditional normal formulas to get:

$$\alpha_{t|t} = \alpha_{t|t-1} + K_t v_t, \quad P_{t|t} = P_{t|t-1}(1 - K_t).$$

- Then prediction:

$$\alpha_{t+1|t} = \alpha_{t|t}, \quad P_{t+1|t} = P_{t|t} + Q.$$

Kalman Recursion (Local Level) Summary

$$
\begin{aligned}
v_t &= y_t - \alpha_{t|t-1}, \\
V_t &= P_{t|t-1} + R, \\
K_t &= \frac{P_{t|t-1}}{V_t}, \\
\alpha_{t+1|t} &= \alpha_{t|t-1} + K_t v_t, \\
P_{t+1|t} &= P_{t|t-1}(1 - K_t) + Q.
\end{aligned}
$$

- Repeat for $t = 1, \ldots, T$ to get filtered and one-step-ahead predicted states.
- Analogue of recursive least squares with time-varying gain.

Forecast Error Representation

- Innovations:

$$\nu_t = y_t - \alpha_{t|t-1}.$$

- First few:

$$\nu_1 = y_1 - \alpha_{1|0},$$

$$\nu_2 = y_2 - \alpha_{1|0} - K_1(y_1 - \alpha_{1|0}),$$

$$\nu_3 = y_3 - \alpha_{1|0} - K_2(y_2 - \alpha_{1|0}) - K_1(1 - K_2)(y_1 - \alpha_{1|0}), \ \ldots$$

- Matrix form:

$$\nu = K(y - \alpha_{1|0}\mathbf{1}_T),$$

with lower-triangular K whose entries depend on $\{K_t\}$.

Cholesky Decomposition via Kalman Filter

- Let $\Omega = \mathrm{Cov}(y)$.
- The transformation K produces uncorrelated innovations:

$$K\Omega K' = \mathrm{diag}(V_1, \ldots, V_T).$$

- Interpretation:
  - Kalman filter implicitly computes a Cholesky decomposition of Ω,
  - ν_t are independent Gaussian with variances V_t.

State Prediction Error Dynamics

- State prediction error:

$$x_t = \alpha_t - \alpha_{t|t-1}, \quad \mathrm{Var}(x_t) = P_{t|t-1}.$$

- Observation equation: $y_t = \alpha_t + \omega_t \Rightarrow$

$$\nu_t = y_t - \alpha_{t|t-1} = x_t + \omega_t.$$

- Updated state:

$$\alpha_{t|t} = \alpha_{t|t-1} + K_t \nu_t = \alpha_{t|t-1} + K_t(x_t + \omega_t).$$

- Prediction:

$$\alpha_{t+1} = \alpha_t + \eta_t, \quad \alpha_{t+1|t} = \alpha_{t|t}.$$

State Error Recursion (Continued)

- State error at $t+1$:

$$x_{t+1} = \alpha_{t+1} - \alpha_{t+1|t} = (\alpha_t + \eta_t) - \alpha_{t|t} = x_t + \eta_t - K_t(x_t + \omega_t).$$

- Hence:

$$x_{t+1} = (1 - K_t)x_t + \eta_t - K_t\omega_t.$$

- Denote $L_t = 1 - K_t = R/V_t$, then

$$x_{t+1} = L_t x_t + \eta_t - K_t\omega_t.$$

- Interpretation:
  - Old error persists scaled by L_t,
  - plus new process noise η_t,
  - minus correction from measurement noise $K_t\omega_t$.

State Smoothing: q_t and M_t

- Smoothing uses all data up to T to refine α_t.
- Introduce backward accumulators:

$$M_t = \frac{1}{V_{t+1}} + L_{t+1}^2 \frac{1}{V_{t+2}} + \cdots + \left(\prod_{j=t+1}^{T-1} L_j^2 \right) \frac{1}{V_T},$$

$$q_t = \frac{\nu_{t+1}}{V_{t+1}} + L_{t+1} \frac{\nu_{t+2}}{V_{t+2}} + \cdots + \left(\prod_{j=t+1}^{T-1} L_j \right) \frac{\nu_T}{V_T},$$

with $M_T = q_T = 0$.

Smoothed State and Variance

- Smoothed estimates:

$$\alpha_{t|T} = \alpha_{t|t-1} + P_{t|t-1} q_{t-1},$$

$$P_{t|T} = P_{t|t-1} - P_{t|t-1}^2 M_{t-1}.$$

- Backward recursions:

$$q_{t-1} = \frac{\nu_t}{V_t} + L_t q_t, \quad q_T = 0,$$

$$M_{t-1} = \frac{1}{V_t} + L_t^2 M_t, \quad M_T = 0.$$

- Start from $t = T$ and go backward to $t = 1$.

Handling Missing Observations

- Suppose $y_{\ell+1}, \ldots, y_{\ell+h}$ are missing.
- For missing y_t:

$$\nu_t = 0, \quad K_t = 0$$

$\Rightarrow$ skip update; use only prediction:

$$\alpha_{t|t-1} = \alpha_{t-1|t-2}, \quad P_{t|t-1} = P_{t-1|t-2} + Q.$$

- State evolves purely via state equation on missing segments.
- Smoothing step still uses future data to refine states at missing times.
- Kalman filter naturally handles irregular or unbalanced time series.

Kalman Initialization

- Initial state: $\alpha_1 \sim N(\alpha_{1|0}, P_{1|0})$.
- First-step forecast:

$$v_1 = y_1 - \alpha_{1|0}, \quad V_1 = P_{1|0} + R.$$

- Update:

$$\alpha_{2|1} = \alpha_{1|0} + \frac{P_{1|0}}{P_{1|0} + R}(y_1 - \alpha_{1|0}),$$

$$P_{2|1} = \frac{P_{1|0}}{P_{1|0} + R}R + Q.$$

- **Diffuse initialization**:
  - Let $P_{1|0} \to \infty$ to represent ignorance.
  - Then

$$\alpha_{2|1} = y_1, \quad P_{2|1} = R + Q.$$

  - First observation dominates initial state.

Initialization and Smoothing

- Diffuse prior affects mainly the earliest states (e.g. $\alpha_{1|T}$).
- Smoothing with diffuse prior yields:

$$\alpha_{1|T} = y_1 + Rq_1, \quad P_{1|T} = R - R^2 M_1,$$

  where q_1, M_1 use information from $y_2, \ldots, y_T$.
- Numerical issues:
  - Very large $P_{1|0}$ can cause instability.
- Alternatives:
  - Large but finite $P_{1|0}$,
  - Estimate $\alpha_{1|0}$ and $P_{1|0}$ jointly via ML.

Likelihood for Local Level Model

- Under Gaussian assumptions:

$$p(y_1, \ldots, y_T \mid R, Q) = p(y_1 \mid R, Q) \prod_{t=2}^{T} p(y_t \mid \mathcal{F}_{t-1}, R, Q).$$

- One-step-ahead residuals:

$$\nu_t = y_t - \alpha_{t|t-1} \sim N(0, V_t).$$

- Log-likelihood:

$$\log \mathcal{L}(R, Q) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \left[\log V_t + \frac{\nu_t^2}{V_t} \right].$$

- Estimate (R, Q) by:

$$(\widehat{R}, \widehat{Q}) = \arg \max_{R,Q} \log \mathcal{L}(R, Q),$$

using numerical optimization.

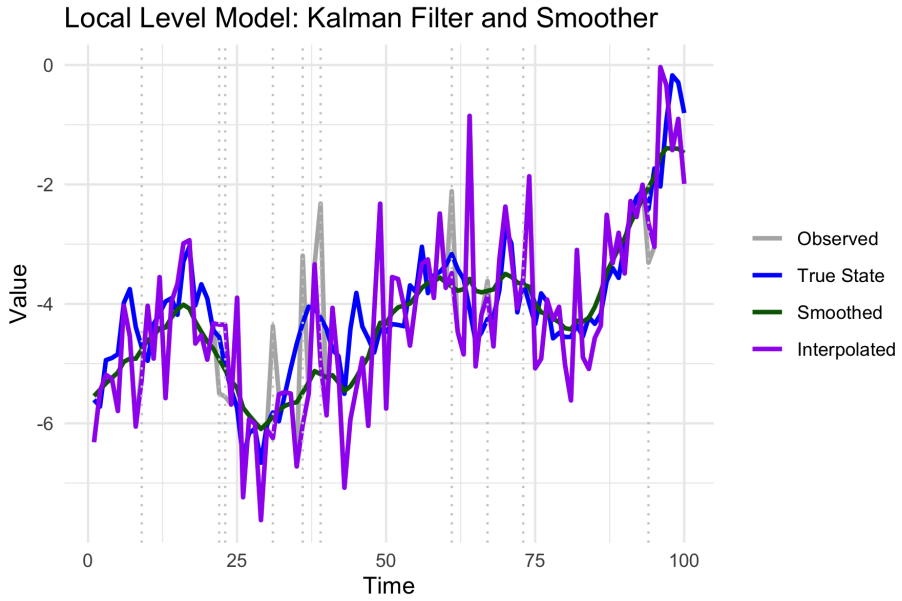
Local Level Model: Simulated Example

- Simulate local level model:

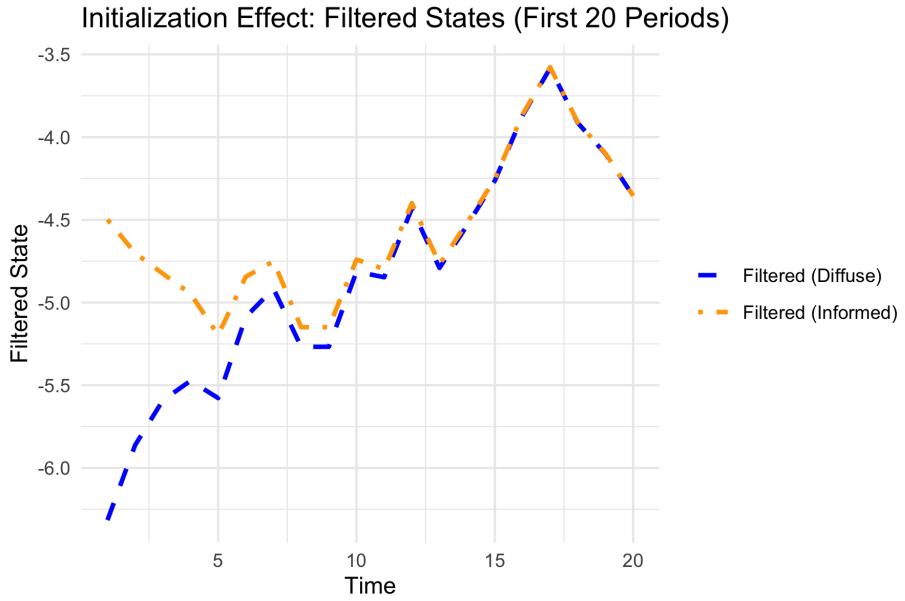
$$\alpha_t = \alpha_{t-1} + \eta_t, \quad y_t = \alpha_t + \omega_t.$$

- Introduce missing observations.
- Estimate model parameters via ML (e.g. R `dlm` package).
- Apply Kalman filter and smoother with:
  - Diffuse initialization,
  - Informed initialization (m_0 near mean, C_0 small).
- Key observations:
  - Missing data handled seamlessly (pure prediction steps).
  - Filtered paths differ early under different priors, but converge quickly.
  - Smoothed estimates track true state closely (in simulation).

Local Level Model: Simulated Example



(a) Observed series, true state, smoother, and interpolated missing values (vertical dotted lines mark missing data).



(b) Filtered states under diffuse and informed initialization (first 20 periods).

Figure: Kalman filtering and smoothing for simulated local-level data with missing observations.

Nile River Flow Example

- Data: annual Nile River flow (built-in R dataset).
- Fit local level state-space model via ML.
- Run Kalman filter and smoother under:
  - Diffuse initialization,
  - Informed initialization (mean of series as prior, small variance).
- Result:
  - Filtered paths differ at sample start,
  - After a few observations, prior influence fades,
  - Smoothed state shows gradual trend shifts in Nile flows.

Nile River Data: Kalman Filter and Smoother

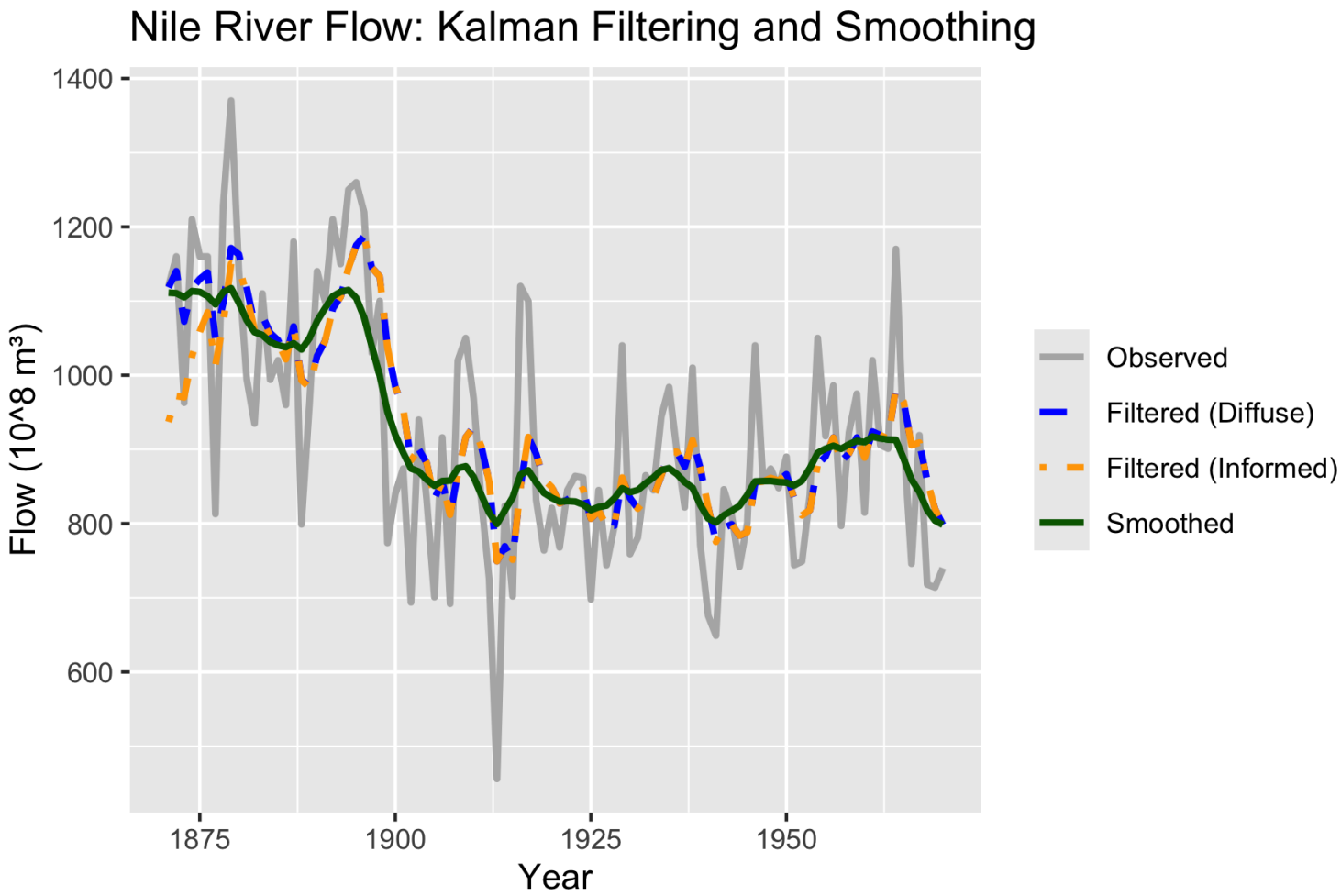


Figure: Nile River data: Kalman filter and smoother.

General Linear Gaussian State-Space Model

- Model:

$$\alpha_t = F_t \alpha_{t-1} + B_t x_t + \eta_t,$$
$$y_t = H_t \alpha_t + \omega_t,$$

  where
  - $\eta_t \sim N(0, Q_t)$, $\omega_t \sim N(0, R_t)$, independent,
  - $\alpha_t \in \mathbb{R}^n$ (state), $y_t \in \mathbb{R}^m$ (observations),
  - $x_t \in \mathbb{R}^p$ known regressors,
  - F_t, B_t, H_t, Q_t, R_t known up to parameters.
- Includes:
  - Time-varying parameter regressions,
  - VARMA/VARMAX with exogenous variables,
  - Many multivariate dynamic models.

Kalman Filter: General Case

Prediction:

$$\alpha_{t|t-1} = F_t \alpha_{t-1|t-1} + B_t x_t,$$
$$P_{t|t-1} = F_t P_{t-1|t-1} F_t' + Q_t.$$

Update:

$$v_t = y_t - H_t \alpha_{t|t-1},$$
$$S_t = H_t P_{t|t-1} H_t' + R_t,$$
$$K_t = P_{t|t-1} H_t' S_t^{-1},$$
$$\alpha_{t|t} = \alpha_{t|t-1} + K_t v_t,$$
$$P_{t|t} = P_{t|t-1} - K_t H_t P_{t|t-1}.$$

Interpreting the Kalman Gain

- Kalman gain:
$$K_t = P_{t|t-1} H_t' S_t^{-1}.$$
- If observation noise R_t is **large**:
  - S_t large, K_t small,
  - Filter trusts model more than data.
- If R_t is **small**:
  - S_t small, K_t large,
  - Filter adjusts strongly based on new observation.
- Kalman filter is an optimal *adaptive weighted average* of prediction and observation.

Kalman Smoother: General Case

- Smoothed mean: $\widehat{\alpha}_t = \mathrm{E}(\alpha_t \mid \mathcal{F}_T)$.
- Smoothed covariance: $\widehat{P}_t = \mathrm{Var}(\alpha_t \mid \mathcal{F}_T)$.
- Backward recursions:

$$\widehat{\alpha}_T = \alpha_{T|T}, \quad \widehat{P}_T = P_{T|T}.$$

For $t = T-1, \ldots, 1$:

$$\begin{aligned} G_t &= P_{t|t} F_t' P_{t+1|t}^{-1} \quad \text{(smoother gain)}, \\ \widehat{\alpha}_t &= \alpha_{t|t} + G_t(\widehat{\alpha}_{t+1} - \alpha_{t+1|t}), \\ \widehat{P}_t &= P_{t|t} + G_t(\widehat{P}_{t+1} - P_{t+1|t}) G_t'. \end{aligned}$$

- If α_t is itself random, $\widehat{\alpha}_t$ is the best linear predictor, not necessarily a consistent estimator of a fixed parameter.

Table of Contents

Basic Stochastic Volatility (SV) Model

- Basic SV model:

$$x_t = \sigma_t z_t,$$
$$\log \sigma_t = \log \sigma_{t-1} + \omega_t,$$

  where $z_t \sim N(0,1)$, $\omega_t \sim N(0, Q)$.
- Volatility follows a random walk in log-scale.
- Goal: estimate latent volatility σ_t and parameters.

Log-squared Transformation

- Transform:

$$y_t = \log x_t^2 = 2s_t + \varepsilon_t,$$
$$s_t = s_{t-1} + \omega_t, \quad s_t = \log \sigma_t,$$

  where $\varepsilon_t = \log z_t^2$.
- Now:
  - State equation is linear Gaussian: s_t random walk.
  - Observation equation is linear in s_t, but ε_t is non-Gaussian: $\varepsilon_t \sim \log \chi_1^2$.
- Non-Gaussian measurement noise violates Kalman assumptions.

Approaches to SV Estimation

- **1. QMLE (ignore non-Gaussianity)**:
  - Treat ε_t as Gaussian; apply Kalman filter.
  - Quasi-ML estimator is consistent under some conditions, but not fully efficient.
- **2. Conjugate Bayesian analysis**:
  - Carefully chosen priors keep posteriors tractable.
  - Model-specific and more complex.
- **3. Gaussian-mixture approximation**:
  - Approximate $\log \chi_1^2$ by finite mixture of normals (e.g. 7-component mixture of Kim, Shephard, and Chib 1998).
  - Conditional on mixture component, model becomes linear Gaussian.

Normal Mixture Approximation to $\log \chi_1^2$

- Approximate

$$\varepsilon_t \sim \log \chi_1^2 \approx \sum_{i=1}^{K} q_i N(m_i, \sigma_i^2).$$

- Parameters (q_i, m_i, σ_i^2) chosen to:
  - Match moments of $\log \chi_1^2$,
  - Minimize Kullback–Leibler divergence.
- Example (7-component mixture from Kim, Shephard, and Chib 1998): see Table of weights q_i, means m_i, variances σ_i^2.

Normal Mixture Approximation to $\log \chi_1^2$

Table: Parameters of the Normal Mixture Approximation to the $\log \chi_1^2$ Distribution

i	q_i (mixing weight)	m_i (mean)	σ_i^2 (variance)
1	0.00730	-10.12999	5.79596
2	0.10566	-3.97281	2.61369
3	0.00002	-8.56686	5.17950
4	0.04395	2.77786	0.16735
5	0.34001	0.61942	0.64009
6	0.24566	1.79518	0.34023
7	0.25750	-1.08819	1.26261

Note: This table reproduces Table 4 from Kim, Shephard, and Chib (1998).

Mixture-based SV Estimation

- Introduce latent discrete indicator $s_t \in \{1, \ldots, K\}$:

$$\varepsilon_t \mid (s_t = i) \sim N(m_i, \sigma_i^2).$$

- Conditional on $\{s_t\}$, the model is **Gaussian state-space**:

$$y_t = 2s_t + m_{s_t} + \text{Gaussian noise},$$

  with linear link to log-volatility.
- This allows:
  - Kalman filter and smoother to be used within each Gibbs iteration,
  - Efficient MCMC for SV models (Kim–Shephard–Chib scheme).
- Result: flexible, accurate SV estimation using state-space machinery.

Table of Contents

Chapter Summary

- **Spectral-domain filtering**:
  - Transfer functions, gain, phase,
  - Relationship: $f_X(\lambda) = |B_\phi(\lambda)|^2 f_Y(\lambda)$,
  - SMA, differencing, band-pass (Baxter–King), EWMA, Gaussian filters.
- **Time-domain filtering**:
  - Whitening/prewhitening (e.g. VAR prewhitening for HAC),
  - Differencing and smoothing,
  - HP filter as penalized-smooth trend.
- **State-space and Kalman filtering**:
  - General linear Gaussian state-space models,
  - Kalman filter (prediction/update) and smoother,
  - Local level example, missing data, initialization, ML estimation.
- **Stochastic volatility**:
  - Log-squared SV transformation,
  - Non-Gaussian noise handled via QMLE or Gaussian mixtures,
  - Mixture-based SV estimation using Kalman filter within MCMC.

Table of Contents

References I

- Chatfield, Christopher (2013). *The Analysis of Time Series: Theory and Practice*. Springer.
- Harvey, Andrew C. (1990). *The Econometric Analysis of Time Series*. MIT Press.
- Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). "Stochastic volatility: likelihood inference and comparison with ARCH models". In: *The review of economic studies* 65.3, pp. 361–393.
- Percival, Donald B. and Andrew T. Walden (1993). *Spectral Analysis for Physical Applications*. Cambridge University Press.
- Tsay, Ruey S. (2013). *Analysis of Financial Time Series*. 3rd. Hoboken, NJ: Wiley. ISBN: 978-1-118-47512-3.

Chapter 8 — Nonstationary Processes

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Table of Contents

Nonstationary Processes: Overview

- Nonstationarity is pervasive in economics and finance:
  - Macroeconomic aggregates often show long-run growth.
  - Asset prices often behave like random walks with drift.
- Violates weak stationarity assumptions of classical time series.
- Requires specialised tools for:
  - Modelling (deterministic vs stochastic trends),
  - Estimation and asymptotics,
  - Hypothesis testing (unit roots, stationarity).

Deterministic vs Stochastic Trends

When we say a process is nonstationary, we typically have in mind:

- **Deterministic trend**:
  - Time-varying mean driven by a deterministic function of t (e.g. linear or polynomial trend).
- **Stochastic trend (unit root)**:
  - Mean evolves stochastically, often via a unit root.
  - Conditional variance may also trend deterministically.

- First: model and remove deterministic component (trend),
- Later: unit root processes and integrated series.

Polynomial Trend Model

- Polynomial trend specification:

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + u_t = Q_p(t;\beta) + u_t,$$

where

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)', \quad p \geq 0.$$

- Disturbance u_t assumed stationary:

$$A(L)u_t = B(L)\varepsilon_t,$$

with $A(\cdot)$, $B(\cdot)$ satisfying standard stationarity and invertibility conditions, and ε_t an innovation sequence.

- Decomposition:
  - Deterministic trend $Q_p(t;\beta)$ ("strong" or "global" trend),
  - Stationary fluctuation u_t (transitory).

Interpretation of Polynomial Trend

- Because Q_p is a polynomial in t:

$$Q_p(t) \to \pm\infty \quad \text{as } t \to \infty$$

  (sign determined by β_p).
- Innovations $\{u_t\}$ are transitory:
  - Move series away from path temporarily,
  - Effects eventually dissipate; y_t returns to trend.
- Example: real GNP with steady 3% growth around a deterministic path.
- Common cases:
  - $p = 1$ (linear trend) most widely used,
  - Quadratic or higher for more curvature (with caution).

Mean and Variance under Additive Trend

- For $y_t = Q_p(t;\beta) + u_t$ with stationary u_t:

$$\mathrm{E}(y_t) = \beta_0 + \beta_1 t + \cdots + \beta_p t^p, \quad \mathrm{Var}(y_t) = \sigma_u^2 \quad \forall t.$$

- Nonstationarity enters only via the mean.

Trend in scale as well:

$$y_t = \alpha + \beta t + \sqrt{\omega + \gamma t}\,\varepsilon_t,$$

with ε_t i.i.d., $\mathrm{E}\varepsilon_t = 0$, $\mathrm{Var}\varepsilon_t = \sigma_\varepsilon^2$:

$$\mathrm{E}(y_t) = \alpha + \beta t, \quad \mathrm{Var}(y_t) = (\omega + \gamma t)\sigma_\varepsilon^2.$$

- Both mean and variance grow linearly in t,
- Same order as in unit-root processes (to be studied later).

Autocovariances and Spurious Persistence

- For additive model $y_t = Q_p(t;\beta) + u_t$ with stationary u_t:

$$(y_t, y_{t-s}) = (u_t, u_{t-s}).$$

- If u_t i.i.d.:

$$(y_t, y_{t-s}) = 0 \quad \forall s > 0.$$

- In practice: use *global* sample mean, not time-varying mean.

Autocovariances and Spurious Persistence

Example: $y_t = t + u_t$, $t = 1, \ldots, T$:

- Sample mean:

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} t = \frac{T+1}{2}.$$

- Sample autocovariance at lag k (using t/T normalization):

$$\widehat{\gamma}(k) \to \int_0^1 (u - 0.5)^2 \, du > 0.$$

- Sample autocorrelations near 1 for all fixed lags k as T grows, even though $\{u_t\}$ is uncorrelated.

OLS Estimation of Polynomial Trend

- Design matrix:

$$X = \begin{pmatrix} 1 & 1 & 1^2 & \cdots & 1^p \\ 1 & 2 & 2^2 & \cdots & 2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & T & T^2 & \cdots & T^p \end{pmatrix}.$$

- Then

$$X'X = \begin{pmatrix} T & \sum t & \cdots & \sum t^p \\ & \sum t^2 & & \\ & & \ddots & \\ & & & \sum t^{2p} \end{pmatrix},$$

with typical term $\sum_{t=1}^{T} t^{i+j} = O(T^{i+j+1})$.

- $X'X/T$ does *not* converge to a nondegenerate limit.

Normalization and Asymptotics

- Define diagonal matrix

$$\Delta = \text{diag}(T, T^3, \ldots, T^{p+1}).$$

- Normalised matrix:

$$\Delta^{-1/2} X'X \Delta^{-1/2} \to M,$$

where M is $(p+1) \times (p+1)$ positive definite.

- OLS estimator:

$$\widehat{\beta} = (X'X)^{-1} X'y,$$

fitted trend: $\widehat{T}_t = x_t' \widehat{\beta}$, residuals $\widehat{u}_t = y_t - \widehat{T}_t$.

i.i.d. disturbances:

- $u_t = \varepsilon_t$ i.i.d., $\text{E}\varepsilon_t = 0$, $\text{Var}\varepsilon_t = \sigma^2$.
- Grenander CLT:

$$\Delta^{1/2}(\widehat{\beta} - \beta) \Longrightarrow N(0, \sigma^2 M^{-1}).$$

Asymptotic Distribution of Trend Estimate

- For fixed t:

$$\sigma^{-1/2}\left(x_t'(X'X)^{-1}x_t\right)^{-1/2}(\widehat{T}_t - T_t) \Longrightarrow N(0,1),$$

  where $T_t = x_t'\beta$.
- Leading contribution to estimation error comes from intercept.
- In probability:

$$\widehat{T}_t - T_t = O_p(T^{-1/2}).$$

Wald Inference for Polynomial Trend

- Test linear hypotheses $R\beta = r$, R is $q \times (p+1)$, rank q.
- Wald statistic:

$$W = \widehat{\sigma}^{-2}(R\widehat{\beta} - r)' X'X (R\widehat{\beta} - r),$$

  with $\widehat{\sigma}^2 = \sum \widehat{u}_t^2 / T$.
- Asymptotically:

$$W \Longrightarrow \chi^2_q.$$

- E.g. test no trend: $H_0 : \beta_1 = \cdots = \beta_p = 0$.
- Pointwise CI for T_t:

$$\widehat{T}_t \pm z_{\alpha/2} \widehat{\sigma} \sqrt{x_t' (X'X)^{-1} x_t}.$$

- With serial correlation: replace $\widehat{\sigma}^2$ by HAC LRV estimate.

Example: Quadratic Trend ($p = 2$)

Example 1 (Quadratic deterministic trend)

- Regressors: $x_t' = (1, t, t^2)$.
- After normalisation:

$$\lim_{T\to\infty} \begin{pmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-3/2} & 0 \\ 0 & 0 & T^{-5/2} \end{pmatrix} X'X \begin{pmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-3/2} & 0 \\ 0 & 0 & T^{-5/2} \end{pmatrix} = M,$$

where

$$M = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{pmatrix}.$$

Quadratic Trend: Standard Errors and Test

- Asymptotic s.e.'s for $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$:

$$\widehat{\sigma}\left(\sqrt{\frac{9}{T}}, \sqrt{\frac{192}{T^3}}, \sqrt{\frac{180}{T^5}}\right).$$

- Test $H_0 : \beta_1 = \beta_2 = 0$:

$$\begin{aligned} W &= \widehat{\sigma}^{-2}\begin{pmatrix}\sqrt{T^3}\,\widehat{\beta}_1 & \sqrt{T^5}\,\widehat{\beta}_2\end{pmatrix}\begin{pmatrix}1/3 & 1/4\\ 1/4 & 1/5\end{pmatrix}\begin{pmatrix}\sqrt{T^3}\,\widehat{\beta}_1\\ \sqrt{T^5}\,\widehat{\beta}_2\end{pmatrix} \\ &= \widehat{\sigma}^{-2}\left(\frac{1}{3}T^3\widehat{\beta}_1^2 + \frac{1}{2}T^4\widehat{\beta}_1\widehat{\beta}_2 + \frac{1}{5}T^5\widehat{\beta}_2^2\right) \Longrightarrow \chi^2_2. \end{aligned}$$

Serial Correlation in Errors

- General disturbance: $A(L)u_t = B(L)\varepsilon_t$, $\{u_t\}$ stationary with Toeplitz covariance Γ_T.
- Asymptotic covariance of OLS:

$$\mathrm{Var}(\widehat{\beta}) = (X'X)^{-1}(X'\Gamma_T X)(X'X)^{-1}.$$

- By Amemiya (1985):

$$(X'X)^{-1}(X'\Gamma_T X)(X'X)^{-1} \simeq (X'\Gamma_T^{-1}X)^{-1}.$$

- So asymptotically, OLS and GLS share the same limit in this setup.
- If

$$\Delta^{-1/2}X'\Gamma_T^{-1}X\Delta^{-1/2} \to M_*$$

with M_* p.d., then

$$\Delta^{1/2}(\widehat{\beta} - \beta) \Longrightarrow N(0, M_*^{-1}).$$

Practical Variance Estimation

- To estimate $\mathrm{Var}(\widehat{\beta})$:
  - **Parametric**:
    - Specify ARMA model for u_t,
    - Estimate, recover $\widehat{\Gamma}_T^{-1}$,
    - Plug into $(X'\widehat{\Gamma}_T^{-1}X)^{-1}$.
  - **Nonparametric (HAC)**:
    - Kernel-based LRV estimates (e.g. Newey–West),
    - Approximate $X'\Gamma_T X$ without parametric model.
- Weak dependence (mixing) ensures standard limit theory.

Trending Heteroskedasticity

- Suppose $\mathrm{Var}(y_t)$ also trends:

$$\mathrm{Var}(y_t) = \gamma_0 + \gamma_1 t + \cdots + \gamma_p t^p.$$

- If variance trends at same order as mean:
  - Highest-order coefficient β_p still consistently estimable,
  - But converges at slower rate:

$$\mathrm{Var}(\widehat{\beta}_p) \sim \frac{T^{3p+1}}{T^{4p+2}} = T^{-(p+1)}.$$

  - For $p = 1$: rate T instead of $T^{3/2}$.
- Lower-order coefficients (including intercept) generally no longer consistently estimable: trending variance overwhelms information.

Linear Filters and Simple Moving Averages

- General linear filter (SMA):

$$\widehat{T}_t = \sum_{j=-n}^{n} w_j y_{t-j} = w(L) y_t,$$

  with $\sum_{j=-n}^{n} w_j = 1$.
- Examples:
  - Two-sided equal weights: $w_j = 1/(2n+1)$,
  - One-sided equal weights: $w_j = 1(j \leq 0)/(n+1)$.
- Polynomial trend fitting can be written as special case with data-dependent weights.

- Aim: smooth away short-run movements (noise).
- Examples:
  - 7-day moving average of COVID-19 cases,
  - Moving averages and Bollinger Bands in finance.

Nonparametric Trend Model

- Nonparametric regression model:

$$y_t = g\left(\frac{t}{T^{\varkappa}}\right) + \sigma\left(\frac{t}{T^{\varkappa}}\right)\varepsilon_t, \quad t = 1, \ldots, T,$$

  where:
  - $g(\cdot)$: smooth trend function,
  - $\sigma(\cdot)$: smooth scale (volatility) function,
  - ε_t: stationary, mixing "short-run noise".
- Scaling parameter $\varkappa \in (0, 1]$:
  - $\varkappa = 1$: g and σ defined on $[0, 1]$ (*local trend*),
  - $\varkappa < 1$: expanding domain $[0, T^{1-\varkappa}]$ (*global+local*).

Local vs Global Trend

- $\varkappa = 1$ (most common):
  - g bounded on $[0, 1]$,
  - *Local/weak trend*: suitable for sample at hand,
  - Avoids explosive extrapolation far from data.
- $\varkappa < 1$:
  - g can be defined on $\mathbb{R}_+$,
  - Asymptotics combine long-span (global) and infill (local) ideas,
  - Used e.g. in Bandi and Phillips (2003).
- We mainly focus on $\varkappa = 1$.

Kernel (Nadaraya–Watson) Trend Estimator

- Estimate $g(u)$ for $u \in [0,1]$ by:

$$\widehat{g}(u) = \frac{\sum_{t=1}^{T} K\big((u - t/T)/h\big) y_t}{\sum_{t=1}^{T} K\big((u - t/T)/h\big)},$$

  with bandwidth $h = h(T)$ and kernel K:
  - $\int K(s)\,ds = 1$, $\int sK(s)\,ds = 0$.
- Common kernels:
  1. Bartlett (triangular): $K(u) = 1 - |u|$ on $[-1,1]$,
  2. Epanechnikov: $K(u) = 0.75(1 - u^2)$ on $[-1,1]$,
  3. Gaussian: $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$,
  4. Double exponential: $K(u) = 0.5 e^{-|u|}$.

Kernel as Linear Smoother

- For interior $u \in (0,1)$, denominator is nearly constant $\Rightarrow$

$$\widehat{g}(u) = \frac{1}{Th} \sum_{t=1}^{T} K\Big(\frac{u - t/T}{h}\Big) y_t = \sum_{t=1}^{T} w_{Tt}(u) y_t,$$

  with deterministic weights $w_{Tt}(u)$, $\sum_{t=1}^{T} w_{Tt}(u) = 1$.
- *Linear smoother*: output is linear in data y.
- Goal: estimate $g(\cdot)$ under smoothness, not parametric form.

Asymptotic Distribution of Kernel Trend Estimator

Theorem 2

Assume:

- g *is twice continuously differentiable at* $u \in (0,1)$,
- $\sigma^2(\cdot)$ *continuous at* u,
- ε_t *stationary, mixing, Herrndorf conditions.*

If $h \to 0$ *and* $Th^5 \to c > 0$ $(h \propto T^{-1/5})$, *then*

$$\sqrt{Th}\big(\widehat{g}(u) - g(u)\big) \Longrightarrow N\big(b(u), v(u)\big),$$

where

$$b(u) = c^{1/2} g''(u)\mu_2(K), \quad v(u) = \sigma^2(u)\,\mathrm{Irvar}(\varepsilon)\|K\|_2^2,$$

with $\mu_2(K) = \int K(s)s^2\,ds$, $\|K\|_2^2 = \int K(s)^2\,ds$.

Rate and Bias–Variance Trade-off

- $\widehat{g}(u)$ is consistent, asymptotically normal, but:
  - Convergence rate: $T^{2/5}$ ($\sqrt{Th}$ with $h \propto T^{-1/5}$),
  - Slower than $T^{1/2}$ parametric rate.
- Reflects cost of estimating infinite-dimensional object $g(\cdot)$.
- Mean squared error (MSE):

$$\text{MSE}(\widehat{g}(u)) = O(T^{-4/5}) = \text{bias}^2 + \text{variance}.$$

- Serial correlation enters through $\text{lrvar}(\varepsilon)$, analogously to mean estimation of stationary processes.

Effect of Different Scaling $\varkappa$

- If $g(t/T^{\varkappa})$ with $0 < \varkappa < 1$:
  - Same estimator $\widehat{g}$ can be used.
  - Asymptotic normality still holds, but:

$$T^{\varkappa} h^5 \to c, \quad \sqrt{T^{\varkappa} h}(\widehat{g}(u) - g(u)).$$

  - Convergence rate becomes $T^{2\varkappa/5}$.
- Example: $\varkappa = 1/2 \Rightarrow$ MSE of order $T^{-2/5}$ (slower convergence).

Kernel Choice and Efficiency

- Optimal bandwidth depends on $g''(u)$, $\sigma^2(u)$, $\text{Irvar}(\varepsilon)$ and kernel constants.
- Example kernel constants:
  - Uniform: $\mu_2(K) = 1/3$, $\|K\|_2^2 = 1/2$,
  - Epanechnikov: $\mu_2(K) = 0.200$, $\|K\|_2^2 = 0.600$,
  - Gaussian: $\mu_2(K) = 1$, $\|K\|_2^2 = 0.282$,
  - Double exponential: $\mu_2(K) = 2$, $\|K\|_2^2 = 0.25$.
- Relative efficiency (Fan 1993):

$$\text{Eff}(K) = (\|K\|_2^2)^2 \mu_2(K).$$

- Epanechnikov optimal; Gaussian close; exponential much less efficient.

Confidence Intervals and LRV Estimation

- Pointwise CI:

$$I_\alpha(u) = \widehat{g}(u) \pm z_{\alpha/2}\sqrt{\frac{\widehat{v}(u)}{Th}},$$

  where $\widehat{v}(u)$ estimates $v(u)$.
- Practice:
  - Often ignore bias term in $b(u)$,
  - Estimate LRV using residuals $\widehat{u}_t = y_t - \widehat{g}(t/T)$,
  - Multiply LRV by $\int K^2$ and divide by Th,
  - Alternatively, apply LRV estimation directly to $w_{Tt}(u)\widehat{u}_t$.
- Self-normalization / bootstrap also possible.

Equal-weight Filters and Epanechnikov

- Two-sided equal-weight filter:

$$K(s) = \frac{1}{2} \quad \text{for } s \in [-1, 1].$$

- Minimises variance, but not MSE (ignores bias).
- Epanechnikov kernel:

$$K(s) = 0.75(1 - s^2) \quad s \in [-1, 1],$$

- Minimises asymptotic MSE $\Rightarrow$ better bias–variance compromise.
- Performance depends on smoothness of g.

Bandwidth Selection

- Bandwidth h controls window size / smoothing.
- Common methods:
  - Plug-in rules,
  - Cross-validation.
- Complications with time series:
  - Dependence, heteroskedasticity,
  - Must account for autocorrelation when tuning h.

One-sided Filters and EWMA

- One-sided equal-weight filter (e.g. Bollinger Bands):

$$w(s) = \frac{1}{1} \quad s \in [-1, 0] \text{ (idealised)}.$$

- EWMA smoother: one-sided exponential kernel:

$$w(u) = e^{-u}, \quad u > 0.$$

- For estimation:
  - Leading bias term of EWMA is $O(h)$ (worse than h^2).
- For prediction:
  - Cannot use future data; one-sided is unavoidable.
  - Bias can be reduced by allowing some negative weights with $\int_{-1}^{0} sw(s)\, ds = 0$.

Reducing Boundary Bias

- One-sided kernels with $\int_{-1}^{0} sw(s)\,ds = 0$:

$$w(s) = 4 + 6s, \quad s \in [-1, 0],$$

$$w(s) = 3 - 6s^2, \quad s \in [-1, 0],$$

  satisfy $\int w(s)\,ds = 1$, $\int sw(s)\,ds = 0$.
- Bias reduces to $O(h^2)$ as in two-sided case.
- Alternative: local linear regression (Fan–Gijbels) to reduce boundary bias.

Spline Smoothing and HP Filter

- Cubic smoothing spline $\widehat{g}_\lambda$ = minimiser of:

$$Q_\lambda(g) = \sum_{t=1}^{T}(y_t - g(t/T))^2 + \lambda \int [g''(u)]^2 du.$$

- Properties:
  - Cubic between observations,
  - Value and first two derivatives continuous,
  - Linear at boundaries.
- Equivalent discrete formulation:

$$Q_\lambda(g) = (y-g)'(y-g) + \lambda g' D g,$$

  where D is second-difference matrix.
- Solution:

$$\widehat{g}_\lambda = (I + \lambda D)^{-1} y.$$

HP Filter and Spline Connection

- HP filter is special case of cubic smoothing spline on equally spaced data.
- As $\lambda \to 0$:
  - Spline interpolates data.
- As $\lambda \to \infty$:
  - $\widehat{g}_\lambda$ tends to a straight line (LS trend).
- Widely used in macro to extract business cycle component.
- Caveats:
  - Detrending can distort dynamic relationships (Sims),
  - Detrended series are generated regressors.

Generated Regressors and Empirical Distribution

- Suppose $y_t = g(t/T) + \varepsilon_t$ with weakly dependent ε_t.
- Trend estimator $\widehat{g}(u)$ (e.g. kernel).
- Residuals: $\widehat{\varepsilon}_t = y_t - \widehat{g}(t/T)$.
- Empirical c.d.f. of residuals:

$$\widehat{F}_T(e) = \frac{1}{T}\sum_{t=1}^{T} 1\{\widehat{\varepsilon}_t \le e\}.$$

- Under regularity:

$$\sqrt{T}(\widehat{F}_T(e) - F(e)) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(1\{\varepsilon_t \le e\} - F(e)) + f(e)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t + o_p(1).$$

- Second term is the "price" of detrending; ignoring it leads to incorrect inference on F.

When Detrending Does Not Matter Asymptotically

- For some functionals, the effect of first-stage trend estimation vanishes asymptotically:
  - E.g. autocovariances of ε_t,
  - Their limiting distribution unaffected by preliminary deterministic trend estimation.
- But for distributions and tails, the additional term matters.
- Detrending should be treated as part of the model, not as harmless preprocessing.

Robust Trend Estimation: Median and Hampel Filter

- Local median estimator:

$$\widehat{\alpha} = \arg\min_{\alpha} \sum_{t=1}^{T} K\left(\frac{u - t/T}{h}\right) |y_t - \alpha|.$$

- Under conditions, $\widehat{g}(u) = \widehat{\alpha}$ is consistent and asymptotically normal.
- **Hampel filter**:
  - Use local median $\widehat{g}_k(t/T)$ in moving window,
  - Local scale estimate via MAD,
  - Replace outliers by local median if they exceed $cs_k(t)$.
- Provides compromise between smoothing and robustness.

SPY Example: Data and Setup

- Data: daily SPDR S&P 500 ETF (SPY) prices.
- Sample: 2015-01-01 to 2024-12-31 (Yahoo Finance).
- Steps:
  1. Extract adjusted closing prices.
  2. Construct time index $t = 1, \ldots, T$.
  3. Compute log prices $\log(\text{price}_t)$.
- Objective:
  - Compare **global** and **rolling local** quadratic trend fits on log-price and price scales.

Global Quadratic Trend

- Fit global quadratic regression:

$$\log(\text{price}_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + u_t.$$

- Use full sample to estimate $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$.
- Fitted global trend (log scale):

$$\widehat{\text{trend}}_{\text{global, log}}(t) = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2.$$

- Back-transform to price scale with bias correction:

$$\widehat{\text{trend}}_{\text{global}}(t) = \exp\left(\widehat{\text{trend}}_{\text{global, log}}(t)\right) \times \widehat{\text{bc}},$$

where $\widehat{\text{bc}} \approx \mathrm{E}(e^{u_t})$ estimated from residuals.

- Peak date of quadratic can be computed as $t_{peak} = -\widehat{\beta}_1/(2\widehat{\beta}_2)$ if $\widehat{\beta}_2 < 0$.

Rolling Local Quadratic Trend

- Rolling window of length n_{window} (e.g. 40 days):
  1. For each time $i \geq n_{\text{window}}$:
  2. Use most recent n_{window} observations $t \in \{i - n_{\text{window}} + 1, \ldots, i\}$.
  3. Fit local regression

$$\log(\text{price}_t) = \beta_0(i) + \beta_1(i)t + \beta_2(i)t^2 + u_t.$$

  4. Predict trend at time i:

$$\widehat{\text{trend}}_{\text{local, log}}(i).$$

  5. Optionally back-transform with local bias correction.
- Produces a *time-varying* quadratic trend that adapts to local dynamics.

SPY Trends on Log Scale

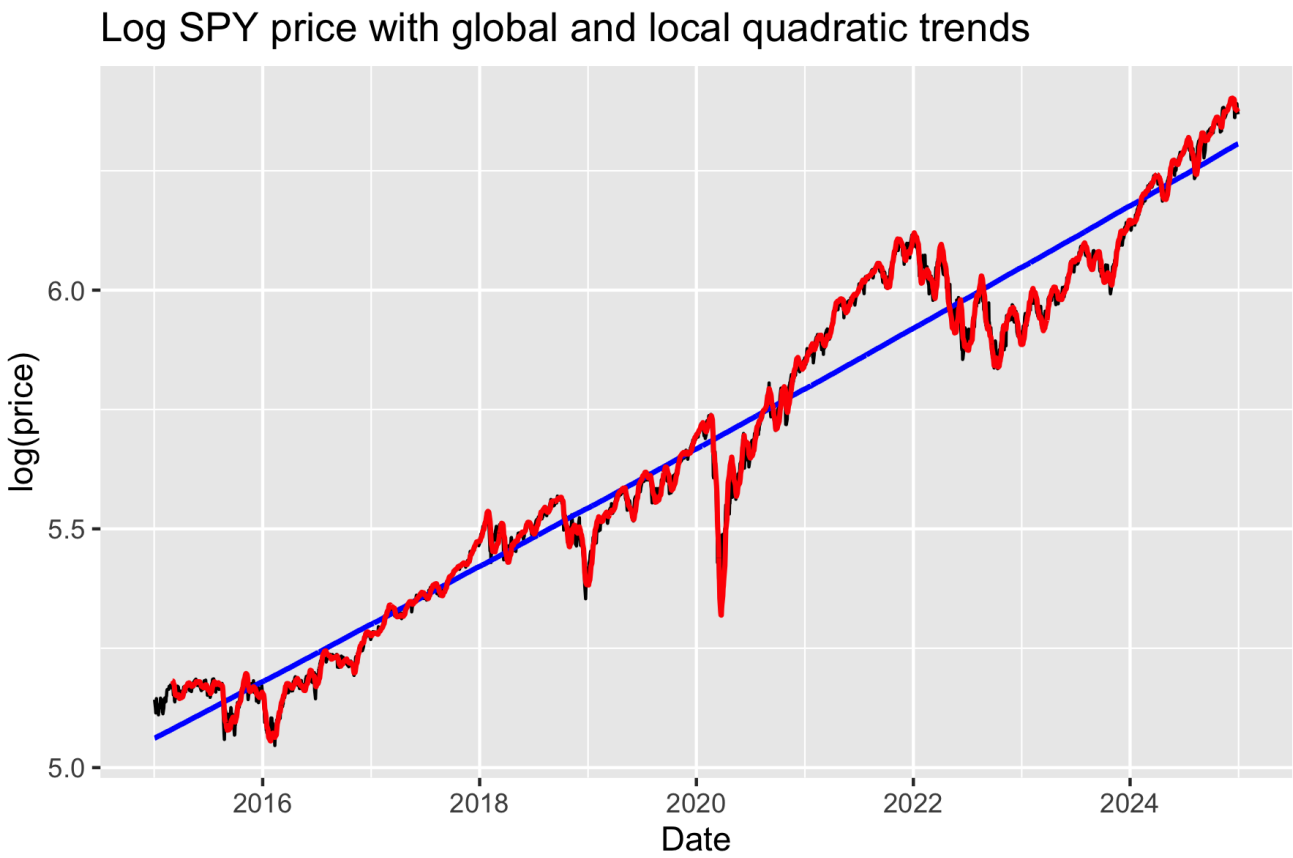


Figure: Log SPY price with global and rolling quadratic trends. Black: log price; blue: global quadratic log trend; red: rolling 40-day local quadratic log trend.

SPY Trends on Price Scale

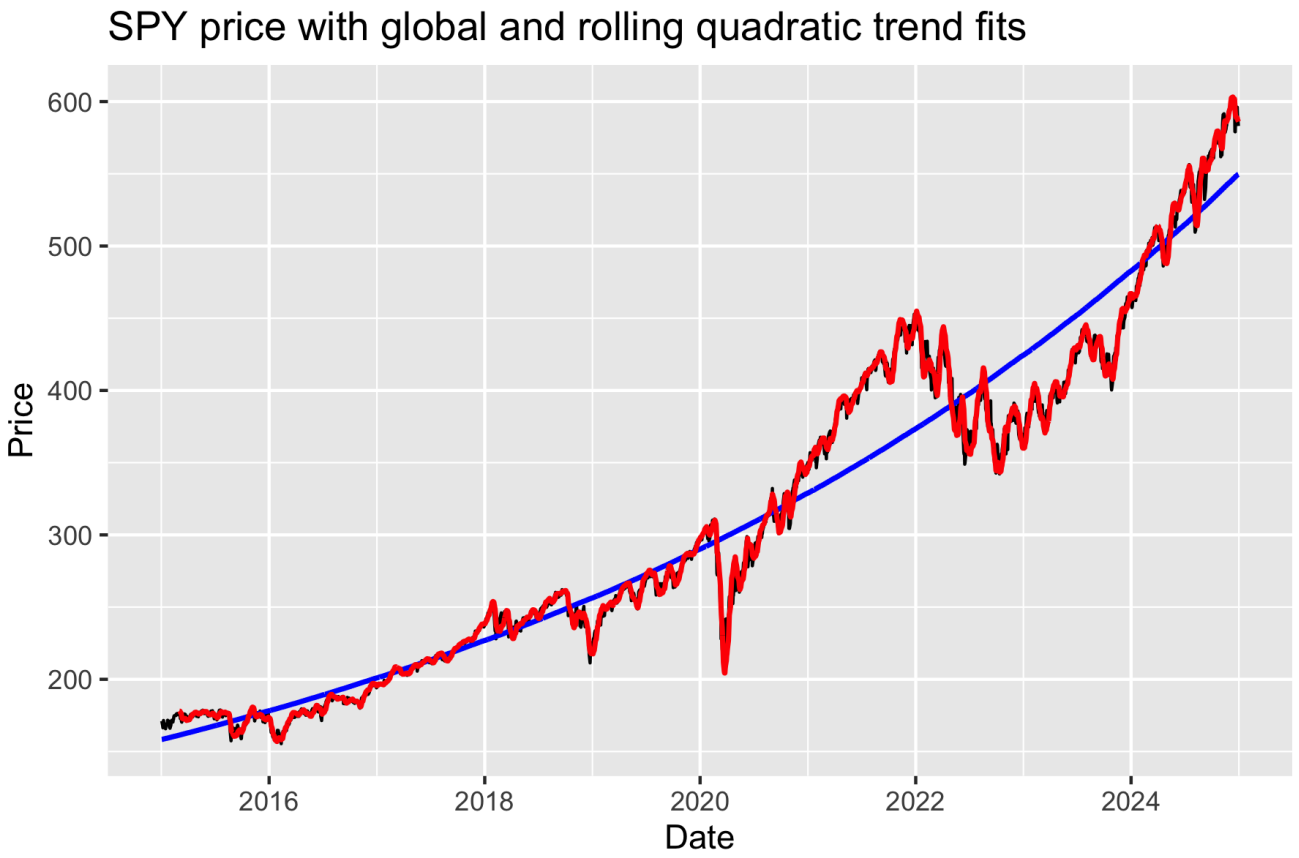


Figure: SPY price with global and rolling quadratic trend fits. Black: price; blue: global quadratic trend (back-transformed); red: rolling 40-day local quadratic trend (back-transformed).

Interpretation of SPY Trend Fits

- **Global Quadratic Trend**:
  - Provides a single smooth long-run path.
  - On log scale, appears almost linear (near-constant growth rate).
- **Rolling Local Quadratic Trend**:
  - More sensitive to local turning points and short-run dynamics.
  - Tracks local curvature and temporary accelerations/slowdowns.
- Comparison:
  - Global trend stable but can be misleading if structural breaks exist;
  - Local trend can overfit noise if window too short.
- Practical lesson: choice between global and local trends affects long-run vs short-run interpretation of price movements.

Table of Contents

Unit Root Processes: Motivation

- Many macro and financial series (output levels, prices, interest rates) exhibit:
  - High persistence,
  - Long memory of past shocks.
- A **unit root** means:
  - Characteristic polynomial has a root with modulus 1,
  - Shocks have permanent effects,
  - Series is nonstationary.
- Detecting unit roots is crucial for:
  - Model specification (differencing vs. levels),
  - Cointegration analysis,
  - Valid inference and forecasting.

AR(1) Example and Unit Root

Example 3

An AR(1) model:

$$Y_t = \phi Y_{t-1} + \varepsilon_t,$$

with lag polynomial $1 - \phi L$ and characteristic equation

$$1 - \phi z = 0.$$

- Root: $z = 1/\phi$.
- A **unit root** occurs when $\phi = 1$, so $z = 1$.

AR(1) Example and Unit Root

- Stationarity requires $|z| > 1 \Leftrightarrow |\phi| < 1$.
- If $\phi = 1$, the process is a random walk:

$$Y_t = Y_{t-1} + \varepsilon_t,$$

with variance exploding over time.

Variance and Stationarity in AR(1)

Consider

$$Y_t = \theta Y_{t-1} + \varepsilon_t.$$

- If $\theta = 1$:

  $\mathrm{Var}(Y_t) = \mathrm{Var}(Y_{t-1}) + \sigma^2 \Rightarrow \mathrm{Var}(Y_t) = t\sigma^2$ (no finite stationary variance).

- Only stationary solution when $\sigma^2 = 0$ (degenerate).
- For $|\theta| \geq 1$:
  - Nonstationary,
  - Shocks accumulate over time.
- AR(1) is stationary iff $|\theta| < 1$.

ARMA and Unit Roots

- General ARMA(p, q):

$$\theta(L)Y_t = \alpha(L)\varepsilon_t.$$

- Stationary iff all roots of AR polynomial $\theta(z) = 0$ satisfy $|z| > 1$.
- Important special case:
  - One root exactly 1, others outside unit circle,
  - Factor:

$$\theta(L) = (1 - L)\theta^*(L),$$

  - Then

$$\theta^*(L)\Delta Y_t = \alpha(L)\varepsilon_t,$$

  so ΔY_t is stationary ARMA.

Example: ARMA(2,1) with Unit Root

Example 4

$$Y_t = 1.2Y_{t-1} - 0.2Y_{t-2} + \varepsilon_t - 0.5\varepsilon_{t-1}.$$

Lag form:

$$(1 - 1.2L + 0.2L^2)Y_t = (1 - 0.5L)\varepsilon_t.$$

Factor:

$$(1 - 0.2L)(1 - L)Y_t = (1 - 0.5L)\varepsilon_t.$$

Characteristic equation:

$$(1 - 0.2z)(1 - z) = 0 \Rightarrow z = 1 \text{ is a root.}$$

Then

$$\Delta Y_t = 0.2\Delta Y_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}$$

is stationary ARMA(1,1).

Integrated Processes: $I(0)$, $I(1)$, $I(2)$

- Y_t is $I(1)$ if:
  - Y_t nonstationary,
  - ΔY_t stationary ARMA.
- Then Y_t follows an ARIMA$(p, 1, q)$.
- $I(2)$: requires differencing twice:

$$\Delta^2 Y_t = \Delta(\Delta Y_t)$$

  to obtain stationarity.
- General:
  - $I(0)$: stationary,
  - $I(1)$: one unit root,
  - $I(2)$: two unit roots, etc.

Comparing $I(0)$ and $I(1)$ Processes

- $I(0)$:
  - Fluctuates around fixed mean,
  - Finite, time-invariant variance,
  - Mean reversion: shocks have temporary effects,
  - ACF decays quickly.
- $I(1)$:
  - Broader fluctuations, no fixed mean,
  - Shocks have **permanent** effects,
  - "Infinite memory" of past shocks,
  - ACF decays very slowly (near unit root).

Simulated AR(1): Series and ACF/PACF

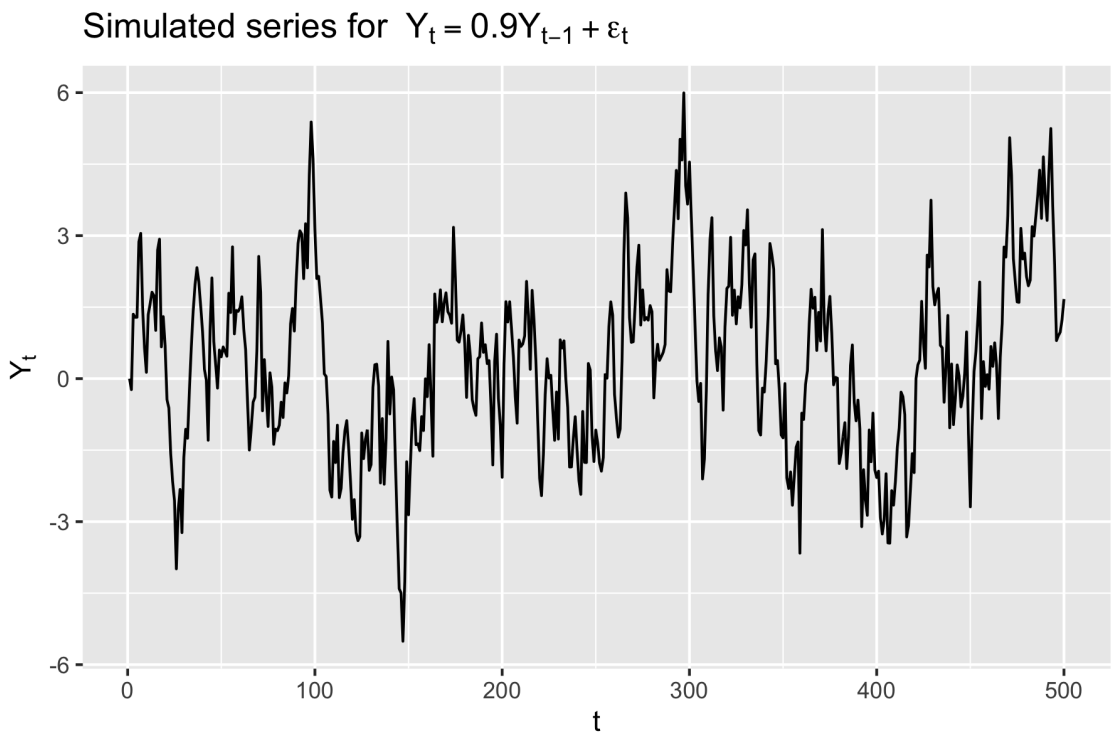


Figure: Simulated AR(1) time series: $Y_t = 0.9Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim i.i.d.N(0,1)$.

AR(1): ACF and PACF

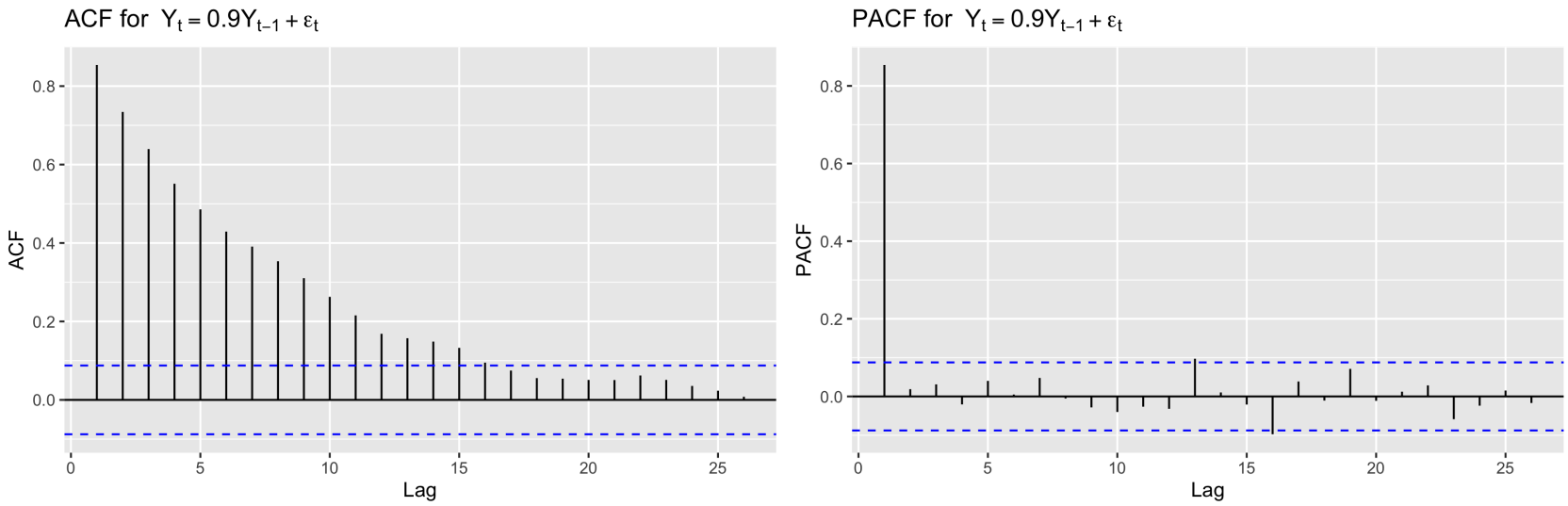


Figure: ACF and PACF of simulated AR(1) series.

Random Walk: Series and ACF/PACF

Random walk:

$$Y_t = Y_{t-1} + \varepsilon_t.$$

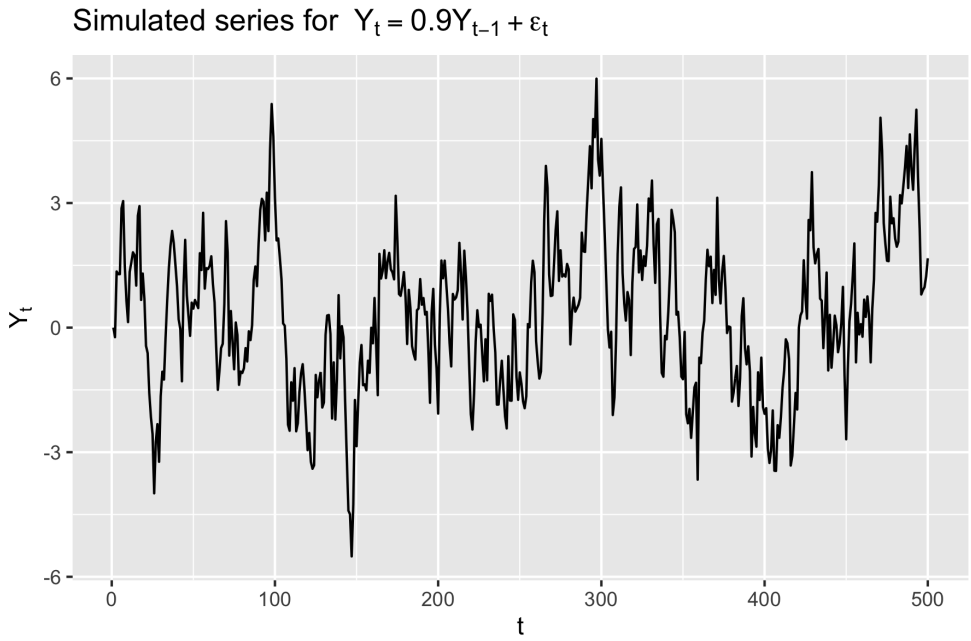


Figure: Time series plot of random walk process.

Random Walk: ACF and PACF

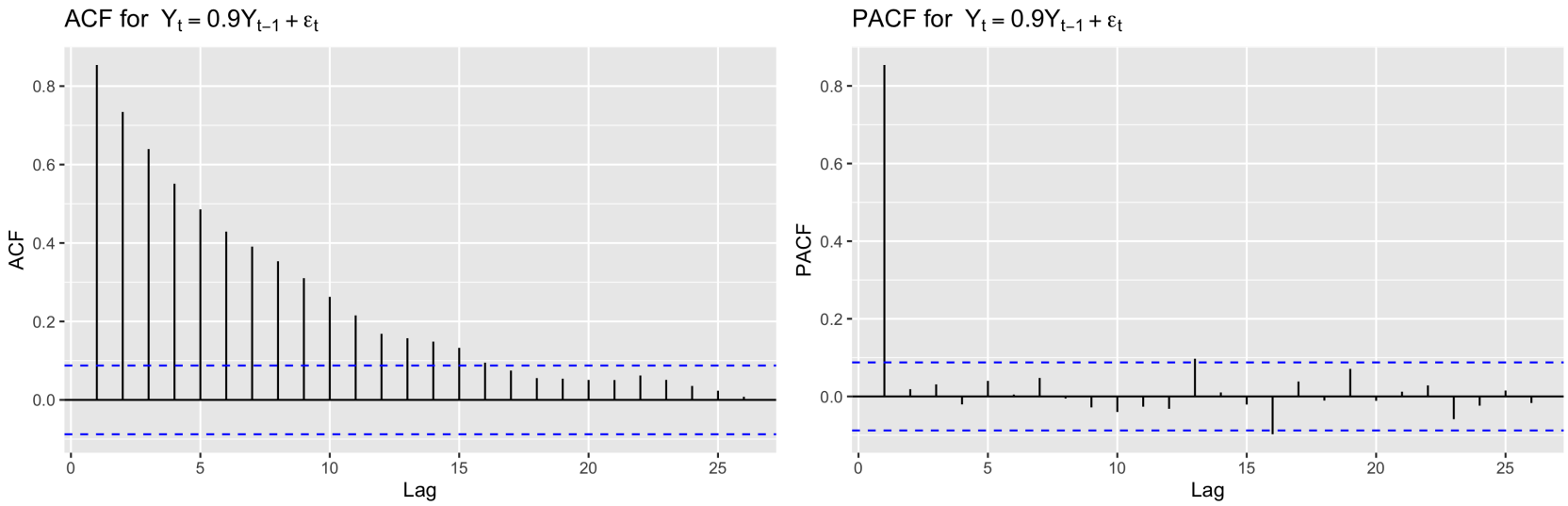
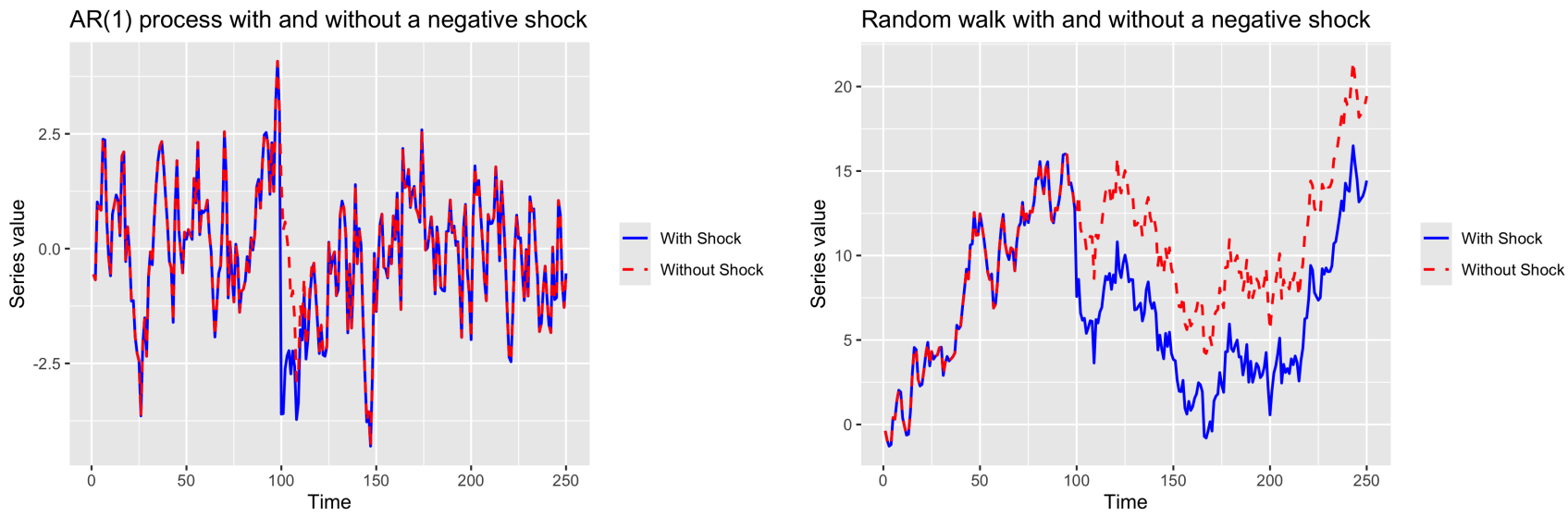


Figure: ACF and PACF for random walk process.

- ACF decays very slowly; first few lags near 1.
- Reflects nonstationarity and persistent effect of shocks.

Shock Response: AR(1) vs Random Walk



(a) AR(1) with and without a negative shock.

(b) Random walk with and without a negative shock.

Figure: Impact of a negative shock on AR(1) vs random walk.

Shock Persistence: Interpretation

- AR(1) with $|\phi| < 1$:
  - Shock at $t = 100$ causes drop,
  - Effect decays geometrically,
  - Process reverts to long-run mean/trend.
- Random walk:
  - Same shock shifts level permanently,
  - No tendency to return to previous path,
  - Cumulative effect of all past shocks.
- Practical implication:
  - $I(1)$ series may take much longer to "absorb" shocks,
  - Stationary series adjust more quickly.

Issues with Stochastic Trends

A series with a unit root exhibits a **stochastic trend**. Key issues:

1. In $Y_t = Y_{t-1} + \varepsilon_t$, the OLS estimator of θ is biased towards 0, even though $\theta = 1$.
2. With a stochastic-trend regressor, OLS coefficient no longer has a standard t-distribution (even in large samples).
3. Two independent $I(1)$ series can appear highly correlated: **spurious regression**.

Dickey–Fuller Test: Basic Idea

- AR(1):
$$Y_t = \delta + \theta Y_{t-1} + \varepsilon_t.$$
- Null: $\theta = 1$ (unit root, nonstationary).
- DF t-statistic:
$$DF = \frac{\widehat{\theta} - 1}{se(\widehat{\theta})}.$$
- Dickey and Fuller (1979):
  - Under $\theta = 1$, DF statistic has nonstandard distribution,
  - Critical values must be obtained from special tables (not t-dist).
- In practice: software provides DF test, critical values, and p-values.

DF Critical Values (Approximate)

Table: 1% and 5% critical values for DF tests (adapted from Fuller, 1976).

	Without trend		With trend	
Sample size	1%	5%	1%	5%
$T = 25$	-3.75	-3.00	-4.38	-3.60
$T = 50$	-3.58	-2.93	-4.15	-3.50
$T = 100$	-3.51	-2.89	-4.40	-3.45
$T = 250$	-3.46	-2.88	-3.99	-3.43
$T = 500$	-3.44	-2.87	-3.98	-3.42
$T = \infty$	-3.43	-2.86	-3.96	-3.41

Deterministic vs Stochastic Trends

Stochastic Trend vs Deterministic Trend

- Weak stationarity requires:

$$\mathrm{E}(Y_t) = \mu, \quad \mathrm{Var}(Y_t) = \gamma_0, \quad \mathrm{Cov}(Y_t, Y_{t-k}) = \gamma_k$$

  all finite and time-invariant.

- Deterministic trend: $\mathrm{E}(Y_t) = \mu(t)$ nonrandom:

$$Y_t = \delta + \gamma t + \alpha(L)\varepsilon_t \Rightarrow \mathrm{E}(Y_t) = \delta + \gamma t.$$

  Often called **trend stationary**.

- Stochastic trend: time-varying variance, e.g.

$$Y_t = Y_{t-1} + \varepsilon_t \Rightarrow \mathrm{Var}(Y_t) = \sigma^2 t.$$

Random Walk with Drift

- Random walk with drift:

$$Y_t = \delta + Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.(0, \sigma^2).$$

- Solution:

$$Y_t = Y_0 + \delta t + \sum_{i=1}^{t} \varepsilon_i.$$

- Moments:

$$\mathrm{E}(Y_t) = Y_0 + \delta t, \quad \mathrm{Var}(Y_t) = \sigma^2 t.$$

- Contains both:
  - Deterministic trend (linear drift),
  - Stochastic trend (cumulative random shocks).

Augmented Dickey–Fuller (ADF) Test

- DF test assumes AR(1) errors; ADF allows higher-order AR.
- Idea: add lagged differences to regression so errors are white noise.
- Example AR(2):

$$Y_t = \delta + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \varepsilon_t.$$

Rewrite:

$$\Delta Y_t = \delta + (\theta_1 + \theta_2 - 1) Y_{t-1} - \theta_2 \Delta Y_{t-1} + \varepsilon_t.$$

- Characteristic equation: $1 - \theta_1 z - \theta_2 z^2 = 0$.
- With unit root $z = 1$: $1 - \theta_1 - \theta_2 = 0$, so testing $\pi = \theta_1 + \theta_2 - 1 = 0$.

General ADF Regression

For AR(p):

$$Y_t = \delta + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + \varepsilon_t.$$

Equivalent ADF regression:

$$\Delta Y_t = \delta + \pi Y_{t-1} + c_1 \Delta Y_{t-1} + \cdots + c_{p-1} \Delta Y_{t-p+1} + \varepsilon_t,$$

with

$$\pi = \theta_1 + \cdots + \theta_p - 1.$$

- Null: $\pi = 0$ (unit root).
- Alternative: $\pi < 0$ (stationary).
- Trend version adds γt on RHS.
- Same nonstandard critical values as DF, depending on specification.

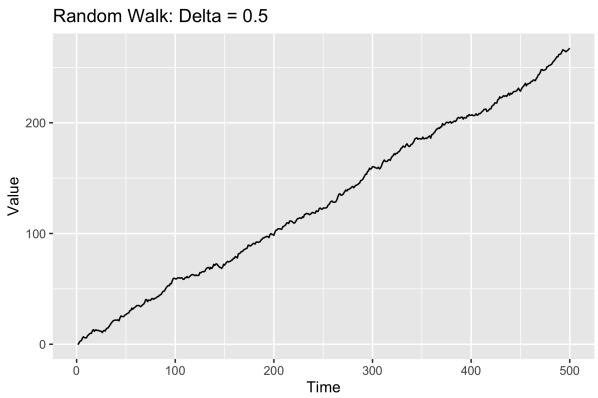
ADF on Simulated Random Walks with Drift

- Simulate random walks:

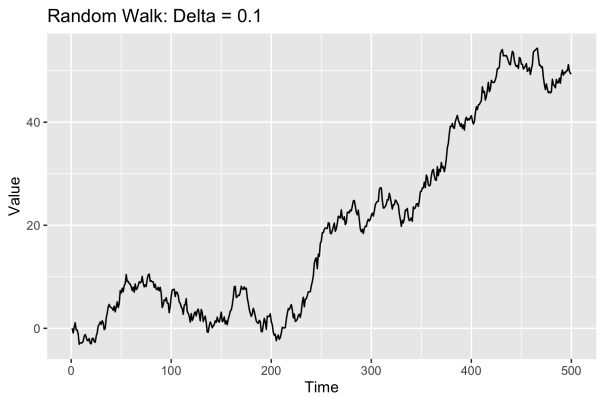
$$Y_t = Y_{t-1} + \delta + \varepsilon_t,$$

  with different drifts $\delta \in \{0.5, 0.1, 0.05\}$.
- Use `ur.df` with `type = "drift"` and 1 lag.
- Examine test statistics:
  - τ_2: DF-type statistic for unit root with drift,
  - ϕ_1: test for trend stationarity.

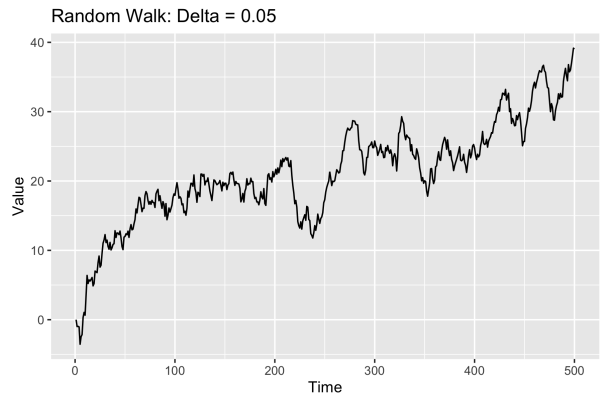
ADF Results for Different Drifts



(a) $\delta = 0.5$
$\tau_2 = 0.4965$, $\phi_1 = 64.55$
Critical values (1%, 5%, 10%):
τ_2: -3.44, -2.87, -2.57
ϕ_1: 6.47, 4.61, 3.79



(b) $\delta = 0.1$
$\tau_2 = -0.0598$, $\phi_1 = 2.46$
Critical values (1%, 5%, 10%):
τ_2: -3.44, -2.87, -2.57
ϕ_1: 6.47, 4.61, 3.79



(c) $\delta = 0.05$
$\tau_2 = -2.0785$, $\phi_1 = 3.75$
Critical values (1%, 5%, 10%):
τ_2: -3.44, -2.87, -2.57
ϕ_1: 6.47, 4.61, 3.79

Figure: ADF test statistics for random walks with different drifts.

Interpreting the ADF Results

- For all δ:
  - τ_2 is greater than 1%/5%/10% critical values $\Rightarrow$ fail to reject unit root.
- ϕ_1:
  - For $\delta = 0.5$, ϕ_1 far above critical values $\Rightarrow$ strong trend component (trend stationarity alternative).
  - For smaller δ, ϕ_1 weaker, trend part less dominant.
- The figure illustrates that a strong deterministic trend can overshadow the stochastic trend in large samples.

Beveridge–Nelson Decomposition (Concept)

Beveridge–Nelson Decomposition

- Decomposes nonstationary series into:
  - **Stochastic trend** (random walk component),
  - **Stationary component** (cyclical fluctuations).
- Helps:
  - Separate permanent from transitory movements,
  - Clarify long-run vs short-run behavior of unit root processes.

KPSS Test: Reverse Hypotheses

- DF/ADF:
  - H_0: unit root (nonstationary),
  - H_1: (trend-)stationary.
- KPSS reverses:
  - H_0: (trend-)stationary,
  - H_1: series has stochastic trend (unit root component).
- Useful complement to DF/ADF:
  - If DF/ADF and KPSS agree, conclusion robust,
  - If they conflict, borderline behaviour or model misspecification.

KPSS Model and Statistic

- Decomposition:

$$y_t = \alpha + \beta t + r_t + \varepsilon_t, \quad r_t = r_{t-1} + u_t,$$

  where $u_t \sim i.i.d.(0, \sigma_u^2)$, ε_t stationary.
- r_t: random walk stochastic trend.
- Hypotheses:

$$H_0 : \sigma_u^2 = 0 \ (\text{(trend-)stationary}), \quad H_1 : \sigma_u^2 > 0 \ (\text{stochastic trend}).$$

- Under H_0, regress y_t on intercept (and trend), get residuals:

$$\widehat{\varepsilon}_t = y_t - \widehat{\alpha} - \widehat{\beta} t.$$

- Partial sums:

$$S_t = \sum_{s=1}^{t} \widehat{\varepsilon}_s.$$

KPSS Test Statistic

- Let $\widehat{\omega}^2$ be HAC estimate of long-run variance of $\widehat{\varepsilon}_t$.
- KPSS statistic:

$$\text{KPSS} = \frac{1}{T^2} \sum_{t=1}^{T} \left(\frac{S_t}{\widehat{\omega}} \right)^2 .$$

- Under H_0, KPSS has nonstandard limit (Brownian functional).
- Critical values tabulated by Kwiatkowski et al. (1992).
- Variants:
  - Level-stationary null (no trend),
  - Trend-stationary null (intercept + trend).

Local-to-Unity Alternatives

- DF/ADF and KPSS are analysed under *local to unity* alternatives:

$$y_t = \phi_T y_{t-1} + u_t, \quad \phi_T = 1 - \frac{c}{T}, \ c > 0.$$

- $\phi_T \to 1$ at rate $1/T$:
  - Captures "near-unit-root" behaviour,
  - Appropriate asymptotic framework for power analysis.
- Contrasts with $I(0)$ tests (portmanteau, Ljung–Box) that target alternatives of order $c/\sqrt{T}$.

SPY: Log Price and Log Returns

- Data: daily SPY (S&P 500 ETF) from 2000-01-01 to 2024-12-31.
- Construct:
  - Log price: $\log P_t$ (candidate $I(1)$),
  - Log return: $\Delta \log P_t$ (candidate $I(0)$).
- Two windows:
  - Full sample (from 2000),
  - Recent subsample (from 2024-01-01).
- Tests:
  - ADF via `ur.df`,
  - KPSS via `ur.kpss` (level and trend).

ADF and KPSS for SPY Log Returns

```
Series: SPY log return (full sample)
Number of observations: 6287

ADF (none): tau1 = -60.04  (crit: -2.58, -1.95, -1.62)
=> Strong rejection of unit root.

KPSS (mu): statistic = 0.3967
Critical values: 0.347, 0.463, 0.574, 0.739
=> Reject level-stationarity at 10% but not at 5%.

Series: SPY log return (recent sample, from 2024)
Number of observations: 251

ADF (none): tau1 = -11.30
=> Strong rejection of unit root.

KPSS (mu): statistic = 0.0496
=> Far below critical values; retain level-stationary null.
```

Interpreting SPY Results

- ADF tests:
  - Strongly reject unit root for log returns in both full and recent samples.
- KPSS:
  - Full sample: weak evidence against pure level-stationarity (10% level), likely due to long-run heteroskedasticity.
  - Recent sample: strong support for level-stationary behaviour.
- Combined view:
  - SPY daily log returns behave approximately as $I(0)$ with time-varying volatility.
  - Consistent with standard financial theory.

Table of Contents

EMH and Unit Roots

- EMH (especially weak form) links naturally to unit-root behaviour:
  - Price process with a unit root behaves like a random walk,
  - Shocks have permanent effects,
  - Future price changes are unpredictable given past information.
- Econometric link:
  - Asset prices: often modelled as $I(1)$,
  - Returns: stationary $I(0)$ innovations.
- DF/ADF and stationarity tests give tools to:
  - Test whether prices are $I(1)$,
  - Check whether returns are approximately unpredictable.

Historical Development of EMH

- EMH originates in the 1960s:
  - Samuelson (1965): "properly anticipated prices fluctuate randomly".
  - Fama (1970): formal definition and taxonomy of market efficiency.
- Samuelson:
  - Focus on pricing of storable goods under uncertainty,
  - Efficient markets $\Rightarrow$ price changes are unpredictable.
- Fama:
  - Empirical work using early computers,
  - Defined EMH: prices *fully reflect all available information*.

Intuition Behind EMH

- More efficient $\Rightarrow$ more *random* price movements.
- Mechanism:
  - Many informed traders try to exploit information,
  - Their trades quickly push prices to incorporate that information,
  - Profit opportunities get competed away.
- In the limit:
  - Price changes look like **martingale differences** (unpredictable),
  - Hard to earn abnormal returns on publicly available information.

Neoclassical Extension of EMH

- 1970s: EMH extended to risk-averse investors.
- Neoclassical version:
  - Price changes must be unpredictable *given* risk preferences,
  - Prices consistent with marginal utilities and no-arbitrage.
- Extensions:
  - Non-traded assets (human capital),
  - State-dependent preferences, heterogeneity,
  - Information asymmetries, transaction costs.
- Core inference remains:
  - At any time, security prices fully reflect available information.

Three Forms of EMH

- **Weak-form efficiency**:
  - Prices incorporate all historical price and volume information.
  - No excess returns from trading rules based on past prices.
- **Semi-strong efficiency**:
  - Prices reflect all *public* information (fundamentals, news).
  - Event studies: prices adjust quickly to public news.
- **Strong-form efficiency**:
  - Prices reflect all information (public + private).
  - Even insiders cannot earn persistent abnormal returns.

Behavioural Finance and EMH

- Real markets often deviate from textbook EMH:
  - Noise traders, behavioural biases,
  - Limits to arbitrage.
- Behavioural finance:
  - Analyses impact of emotions, cognitive biases,
  - Explains anomalies and short-term deviations from efficiency.
- EMH remains:
  - A benchmark for modelling and testing,
  - A central reference in finance and econometrics.

Tests of EMH: Overview

Table: Tests of market efficiency by EMH form.

Hypothesis Type	Characteristics	Testing Method	Typical Conclusion
Weak form	Prices fully reflect all historical price/volume data.	Statistical tests for patterns in historical prices (ACF, AR tests, VR tests).	Historical patterns rarely yield persistent abnormal profits.
Semi-strong	Prices reflect all publicly available information.	Event studies around news releases, earnings, etc.	Prices usually adjust quickly to public news.
Strong form	Prices reflect all information (public + private).	Performance of insiders and professional managers.	Insiders sometimes earn abnormal returns $\Rightarrow$ strong form rarely holds.

Random Walk and Weak-form EMH

- Weak-form EMH: price changes are unpredictable from past prices.
- Random walk model:

$$P_t = P_{t-1} + \varepsilon_t,$$

  with ε_t innovation (mean zero).
- Random walk with drift:

$$P_t = \mu + P_{t-1} + \varepsilon_t = P_0 + t\mu + \sum_{s=1}^{t} \varepsilon_s.$$

- Interpretation:
  - P_t: price or log price,
  - μ: expected normal return,
  - ε_t: unexpected component (news).

Time-varying Expected Returns and Risk

- In general, expected return may be time-varying:

$$\mu_t = \Psi(P_{t-1}, P_{t-2}, \dots)$$

  or linked to risk:

$$\mu_t = h\Big(\mathrm{Var}(r_t \mid \mathcal{F}_{t-1})\Big),$$

  with h increasing.
- High-frequency data (intraday):
  - Often assume $\mu_t \approx 0$.
- Daily/weekly:
  - μ_t nearly constant, often approximated by constant μ.
- Quarterly/annual:
  - Need to model time-variation in expected returns explicitly.

Innovation Assumptions: rw1–rw3

To make random-walk model precise, we consider assumptions on ε_t:

- **rw1 (i.i.d.)**: ε_t i.i.d. with $\mathrm{E}[\varepsilon_t] = 0$.
  - Strongest assumption: independence and identical distribution.
- **rw2 (independent)**: ε_t independent over time with $\mathrm{E}[\varepsilon_t] = 0$.
  - Distribution allowed to change over time.
- **rw3 (uncorrelated)**: for all k, t, $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0$.
  - Uncorrelated, but other forms of dependence allowed.
  - Weakest of the three assumptions.

Hierarchy: rw1 $\Rightarrow$ rw2 $\Rightarrow$ rw3.

Trading Time vs Calendar Time

Trading Time vs Calendar Time

- **Trading time**: returns generated only on trading days.
  - Daily returns computed over fixed trading intervals,
  - Often simplest assumption.
- **Calendar time**: returns evolve continuously, including weekends/holidays.
  - Observed returns skip non-trading days,
  - Need adjustment for varying D_t (days since previous trade).

- Under rw1 with $\mathrm{E}r_t = \mu$, $\mathrm{Var}(r_t) = \sigma^2$,

$$\mathrm{E}(r_t^O) = \mu D_t, \quad \mathrm{Var}(r_t^O) = \sigma^2 D_t,$$

  where r_t^O observed return and D_t days between trades.
- Unless stated otherwise, we usually assume trading-time generation.

Autocorrelation Test for Weak-form EMH

- Efficient market (weak form) $\Rightarrow$ price changes (returns) behave like:
  - Serially uncorrelated (under rw1),
  - At least not predictably linear from past returns.
- If significant autocorrelation:
  - Past returns help predict future returns,
  - Potential violation of EMH.

- Autocovariance:

$$\gamma(j) = \mathrm{Cov}(Y_t, Y_{t-j}) = \mathrm{E}[(Y_t - \mu)(Y_{t-j} - \mu)].$$

- Autocorrelation:

$$\rho(j) = \frac{\gamma(j)}{\gamma(0)}.$$

Sample ACF and CLT

- Sample autocovariance and autocorrelation:

$$\widehat{\gamma}(j) = \frac{1}{T} \sum_{t=j+1}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}),$$

$$\widehat{\rho}(j) = \frac{\widehat{\gamma}(j)}{\widehat{\gamma}(0)} \quad \text{(or usual normalised form)}.$$

- Under rw1 (i.i.d. with finite variance):

Theorem 5

Assume Y_t i.i.d. with finite variance. Then for any fixed p,

$$\sqrt{T}\,(\widehat{\rho}(1), \ldots, \widehat{\rho}(p))' \Longrightarrow N(0, I_p),$$

i.e. each $\sqrt{T}\widehat{\rho}(j) \to_d N(0,1)$ and asymptotically independent.

Testing EMH via ACF

- Under rw1, null: $\rho(k) = 0$ (no predictability from lag k).
- Approximate 95% bounds:

$$\widehat{\rho}(k) \in \left[-\frac{z_{\alpha/2}}{\sqrt{T}}, \frac{z_{\alpha/2}}{\sqrt{T}} \right],$$

  with $z_{\alpha/2} \approx 1.96$.
- If $\widehat{\rho}(k)$ lies outside interval, evidence against EMH at lag k.
- Joint test: Box–Pierce statistic

$$Q = T \sum_{j=1}^{p} \widehat{\rho}(j)^2 \Longrightarrow \chi^2_p$$

  under rw1.

Small-sample Corrections

- In finite samples, $\widehat{\rho}(j)$ can be biased.
- Better autocovariance estimator:

$$\bar{\gamma}(j) = \frac{1}{T-j} \sum_{t=j+1}^{T} (Y_t - \bar{Y}_j)(Y_{t-j} - \bar{Y}_j),$$

with $\bar{Y}_j$ mean over $t = j+1, \ldots, T$.

- Bias-corrected autocorrelation:

$$\widehat{\rho}^{bc}(j) = \widehat{\rho}(j) + \frac{T-j}{(T-1)^2}(1 - \widehat{\rho}(j)^2).$$

- Box–Ljung statistic:

$$Q = T(T+2) \sum_{j=1}^{p} \frac{\widehat{\rho}(j)^2}{T-j},$$

improves small-sample properties.

AR(p) Model Test for EMH

- Consider AR(p):

$$Y_t = \mu + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \varepsilon_t,$$

  with $\mathrm{E}(\varepsilon_t \mid Y_{t-1}, \ldots, Y_{t-p}) = 0$.
- Under weak-form EMH:
  - Past returns should not predict current return,
  - Null: $H_0 : \beta_1 = \cdots = \beta_p = 0$.
- Alternative: at least one $\beta_j \neq 0$.

Wald Test for AR Coefficients

- Let X be $(T-p-1)\times(p+1)$ regressor matrix (constant + lags).
- OLS estimate $\widehat{\beta}=(\widehat{\beta}_1,\ldots,\widehat{\beta}_p)'$.
- Wald statistic:

$$W = T(\widehat{\beta}-\beta)'\widehat{V}^{-1}(\widehat{\beta}-\beta),$$

  with $\beta=0$ under EMH.
- Under rw1:

$$\widehat{V}=\widehat{\sigma}_e^2(X'X/T)^{-1}.$$

- Under heteroskedasticity (rw2, rw3), use White's robust covariance:

$$\widehat{V}_W=(X'X/T)^{-1}(X'DX/T)(X'X/T)^{-1},$$

  where $D=\mathrm{diag}(\widehat{\varepsilon}^2)$.
- $W\sim\chi_p^2$ under H_0.

Beyond rw1: rw2 and Variance Ratios

- Under rw2: $\bar{Y}_t = Y_t - \mathrm{E}Y_t$ independent but not identically distributed.
- Define:

$$\gamma_0 = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}(\bar{Y}_t^2),$$

$$\lambda_{ij} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=j+1}^{T} \mathrm{E}(\bar{Y}_{t-j}^2)\mathrm{E}(\bar{Y}_t^2),$$

and related variance ratio quantities.

Beyond rw1: rw2 and Variance Ratios

Theorem 6

Under rw2 with suitable moment conditions, for $j = 1, \ldots, p$,

$$\sqrt{T}(\widehat{\rho}(1), \ldots, \widehat{\rho}(p))' \Longrightarrow N(0, V_2(p)),$$

and

$$\sqrt{T}(\widetilde{VR}(p) - 1) \Longrightarrow N(0, \omega_2(p)),$$

where $V_2(p)$ *and* $\omega_2(p)$ *depend on* λ_{ij} *and* γ_0.

rw3 Assumptions and Robust VR Tests

- rw3 allows uncorrelated but possibly dependent $\widetilde{Y}_t = Y_t - \mathrm{E}Y_t$.
- Linton (2019) provide assumptions ensuring finite asymptotic variance:
  - Zero means and zero autocovariances,
  - Mixing conditions on products $\widetilde{Y}_t \widetilde{Y}_{t-j}$,
  - Fourth-moment restrictions and limits of averaged second moments.
- Under these conditions, the same CLTs as in rw2 hold.
- In practice:
  - Use sample estimators for λ_{ij}, γ_0,
  - Implement heteroskedasticity-robust VR tests (e.g. automatic VR).

SPY Daily Returns: Basic Features

- Data: SPY daily log returns, 2000–2024 and 2024 alone.
- Summary:
  - Full sample mean $\approx 3 \times 10^{-4}$ per day,
  - 2024 mean $\approx 9 \times 10^{-4}$ per day,
  - Min/max around $-11\%/+13\%$ (full), $\pm 3\%$ (2024).
- Returns are:
  - Highly volatile relative to their mean,
  - Roughly symmetric, thin-tailed relative to shocks.

Serial Correlation and AR Tests

- Full sample:
  - Ljung–Box (lag 20): $X^2 = 131.1$, $p < 10^{-16}$,
  - Robust `Auto.Q`: $p \approx 0.0018$,
  - AR(5) with robust Wald: joint lags significant ($p \approx 0.029$),
  - But $R^2 \approx 0.8\% \Rightarrow$ economically small predictability.
- 2024 subsample:
  - Ljung–Box: $p \approx 0.64$,
  - `Auto.Q`: $p \approx 0.56$,
  - AR(5) Wald: no evidence against H_0 ($p \approx 0.69$),
  - No linear predictability in recent data.

Variance Ratio Tests for SPY

- Full sample (classical VR):
  - $\mathrm{VR}(k) - 1 < 0$ for $k = 2, 5, 10, 20$, highly significant,
  - Indicates short-horizon mean reversion (variance lower than RW).
- Robust automatic VR (`Auto.VR`):
  - Strongly negative statistic, rejecting random-walk null even under heteroskedasticity.
- 2024 subsample:
  - Classical VR: insignificant at small k; some mean reversion at larger k.
  - Robust `Auto.VR`: small, not significant.

EMH Interpretation for SPY

- 2000–2024:
  - Statistically detectable deviations from i.i.d. random walk:
    - Some serial correlation,
    - Mean reversion at multi-day horizons.
  - But predictability is economically modest (very low R^2).
- 2024:
  - Daily returns look close to i.i.d. noise,
  - No robust evidence of linear predictability,
  - Weak-form EMH roughly consistent at daily horizon.

Table of Contents

Chapter Summary (1): Deterministic Trends

- Distinguished two types of nonstationarity:
  - Deterministic trends (time-varying mean),
  - Stochastic trends (unit roots).
- Polynomial trend models:
  - Additive decomposition $y_t = Q_p(t;\beta) + u_t$,
  - OLS estimation with nonstandard asymptotics,
  - Detrending crucial to avoid spurious persistence.
- Nonparametric trend fitting:
  - Kernel and spline smoothers (e.g. HP filter),
  - Bias–variance trade-off and bandwidth choice,
  - Generated regressor issues downstream.

Chapter Summary (2): Unit Roots and Tests

- Stochastic trends and integrated processes:
  - Stationarity of ARMA depends on AR roots,
  - Unit roots $\Rightarrow$ random walk behaviour and permanent shocks,
  - ARIMA(p, d, q) and $I(0)$ vs $I(1)$.
- Unit root tests:
  - DF/ADF: null is unit root; nonstandard critical values,
  - KPSS: null is (trend-)stationarity; alternative is stochastic trend,
  - Combining them gives richer diagnostic.

Chapter Summary (3): EMH and Weak-form Tests

- EMH provides economic interpretation of unit-root models:
  - Prices often $I(1)$; returns $I(0)$ innovations.
- Weak-form EMH testing:
  - Autocorrelation and portmanteau tests,
  - AR(p) regressions with robust Wald tests,
  - Variance ratio (VR) tests, including heteroskedasticity-robust versions.
- SPY illustration:
  - Long-horizon data: small but significant deviations from RW,
  - Recent data: returns close to serially uncorrelated,
  - Consistent with approximate weak-form efficiency at daily frequency.

Table of Contents

References I

- Bandi, Federico M. and Peter C. B. Phillips (2003). "Fully Nonparametric Estimation of Scalar Diffusion Models". In: *Econometrica* 71.1, pp. 241–283. DOI: 10.1111/1468-0262.00395.
- Dickey, David A and Wayne A Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American statistical association* 74.366a, pp. 427–431.
- Fama, Eugene (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *Journal of Finance* 25, pp. 383–417.
- Kwiatkowski, Denis et al. (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root". In: *Journal of Econometrics* 54.1-3, pp. 159–178.
- Linton, Oliver (2019). *Financial econometrics*. Cambridge University Press.
- Samuelson, Paul A. (1965). "Proof that properly anticipated prices fluctuate randomly". In: *Industrial Management Review* 6, pp. 41–50.

Chapter 9 — Continuous Time Finance

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Continuous-Time Finance: Overview

- Continuous-time models are central in modern financial econometrics:
  - Asset prices and returns,
  - Interest-rate dynamics,
  - High-frequency market data.
- Core building blocks:
  - Brownian motion and stochastic calculus,
  - Diffusion-based asset pricing models,
  - Realized volatility and covariance from high-frequency data.
- Practical focus:
  - Handling microstructure noise (bid–ask bounce, discreteness, asynchronicity),
  - Robust volatility/covariance estimation,
  - Applications to risk management and portfolio choice.

Table of Contents

From Random Walk to Brownian Motion

- Start with discrete-time i.i.d. innovations:

$$\varepsilon_t \sim N(0,1), \quad \text{i.i.d.}$$

- Define a random walk $\{z_t\}$:

$$\begin{aligned} z_1 - z_0 &= \varepsilon_0, \\ z_{t+1} - z_t &= \varepsilon_t, \quad t = 1, 2, \ldots \end{aligned}$$

- Then

$$z_t - z_0 = \sum_{j=1}^{t} \varepsilon_{t-j}.$$

From Random Walk to Brownian Motion

- Expectation:

$$\mathrm{E}(z_t - z_0) = \sum_{j=1}^{t} \mathrm{E}(\varepsilon_{t-j}) = 0.$$

- Variance:

$$\mathrm{Var}(z_t - z_0) = \mathrm{E}\Big(\sum_{j=1}^{t} \varepsilon_{t-j}\Big)^2 = \sum_{j=1}^{t} \mathrm{E}(\varepsilon_{t-j}^2) = t.$$

- Variance grows linearly with t, standard deviation $\propto \sqrt{t}$.

Toward Continuous-Time Brownian Motion

- Consider a process $\{z_t\}$ with increments:

$$z_{t+\Delta} - z_t \sim N(0, \Delta), \quad \forall\, \Delta > 0,$$

  and independent across disjoint intervals.
- In discrete time, Δ was an integer; in continuous time, we allow Δ to be arbitrarily small.
- A stochastic process with:
  - $z_0 = 0$,
  - Independent increments,
  - $z_{t+\Delta} - z_t \sim N(0, \Delta)$

  is called **Brownian motion** or a **Wiener process**.
- Brownian motion can be viewed as the continuous-time limit of a random walk.

Table of Contents

Stochastic Differentials: Intuition

- For small $\Delta t > 0$, define an increment:

$$dz_t \approx z_{t+\Delta t} - z_t.$$

- In discrete time:

$$\varepsilon_{t+1} = z_{t+1} - z_t.$$

- For Brownian motion B_t:

$$B_{t+\Delta t} - B_t \sim N(0, \Delta t).$$

- Hence:

$$\mathrm{Var}(dB_t) = \Delta t, \quad \mathrm{sd}(dB_t) \propto \sqrt{\Delta t}.$$

Stochastic Differentials: Intuition

- As $\Delta t \to 0$:
  1. **Non-differentiability**: $\frac{\sqrt{\Delta t}}{\Delta t} \to \infty$, so B_t is continuous but nowhere differentiable.
  2. **Random fluctuations**: Brownian paths are highly irregular at all scales.

Properties of Brownian Increment dB_t

- In differential notation, Brownian motion B_t satisfies:

$$\mathrm{E}_t[dB_t] = 0,$$

$$\mathrm{Var}_t(dB_t) = \mathrm{E}_t[dB_t^2] = dt.$$

- These can be heuristically viewed as:

$$dB_t \sim N(0, dt).$$

General Brownian-Driven SDE

A simple stochastic differential equation for an asset price process:

$$dX_t = \mu\, dt + \sigma\, dB_t,$$

where μ is the drift and σ is the volatility.

Properties of Brownian Increment dB_t

- Conditional mean:

$$\mathrm{E}_t[dX_t] = \mu dt.$$

- Conditional variance:

$$\mathrm{Var}_t(dX_t) = \sigma^2 dt.$$

Table of Contents

Hölder Regularity and Modulus of Continuity

- Brownian motion is continuous but highly irregular.
- Hölder continuity:
  - For any $\gamma < \frac{1}{2}$, with probability 1:

$$|B_t - B_s| \leq C|t - s|^{\gamma}$$

    for some random constant C.
  - For any $\gamma > \frac{1}{2}$, paths are almost surely *not* Hölder continuous.

Hölder Regularity and Modulus of Continuity

- Modulus of continuity:

$$g(\delta) = \left(2\delta \log(1/\delta)\right)^{1/2}.$$

- Result:

$$\Pr\left(\limsup_{\delta \to 0} \frac{1}{g(\delta)} \max_{\substack{0 \le s < t \le 1 \\ t-s \le \delta}} |B_s - B_t| = 1\right) = 1.$$

- Interpretation: the maximum increment over intervals of length δ behaves asymptotically like $g(\delta)$.

Crossing Times and Stopping Times

- **Crossing time** (first exit time) for Brownian motion:

$$\tau_a = \inf\{t \geq 0 : |B_t| > a\},$$

  is the first time B_t exits the interval $[-a, a]$.
- Such crossing times are examples of **stopping times**.

Crossing Times and Stopping Times

Stopping Time

A random time τ is a stopping time w.r.t. filtration $\{\mathcal{F}_s\}$ if

$$\{\tau \leq s\} \in \mathcal{F}_s \quad \forall\, s.$$

That is, whether τ has occurred by time s can be determined using information up to s.

- Stopping times play a key role in:
  - Optimal stopping problems (e.g. American options),
  - Martingale theory,
  - Sequential testing.

Table of Contents

First Passage Time of Brownian Motion

- First passage time (or hitting time) for Brownian motion:

$$\tau_a = \inf\{t \geq 0 : |B_t| > a\}.$$

- Bachelier (1900) derived the distribution:

$$\Pr(\tau_a \leq t \mid x, a) = 1 - \frac{2}{\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{j+\frac{1}{2}} \cos\Big(\big(j+\tfrac{1}{2}\big)\frac{\pi x}{a}\Big) \exp\Big(-\big(j+\tfrac{1}{2}\big)^2 \frac{\pi^2 t}{2a^2}\Big).$$

- Corresponding density:

$$f_0(t \mid x, a) = \frac{\pi}{a^2} \sum_{j=0}^{\infty} (-1)^j \Big(j+\tfrac{1}{2}\Big) \cos\Big(\big(j+\tfrac{1}{2}\big)\frac{\pi x}{a}\Big) \exp\Big(-\big(j+\tfrac{1}{2}\big)^2 \frac{\pi^2 t}{2a^2}\Big).$$

Expected First Passage Time

- In the special case $x = 0$ (starting at the origin), the expected first passage time has a simple form:

$$\mathrm{E}(\tau_a) = \frac{a^2}{\sigma^2}.$$

- This shows:
  - Hitting times grow quadratically with the barrier a,
  - Inversely proportional to variance parameter σ^2.
- First passage times appear in:
  - Barrier options,
  - Default models,
  - Risk limits and safety thresholds.

One-Sided Crossing Probability: Definition

One-Sided Crossing Time

Let $X(t)$ be a stochastic process and a a threshold. The one-sided crossing time is

$$\tau_a^+ = \inf\{t \geq 0 : X(t) > a\}.$$

The one-sided crossing probability up to time t is

$$\Pr(\tau_a^+ \leq t) = \Pr\big(\min\{s : X(s) > a\} \leq t\big).$$

- Compared to general stopping times, one-sided crossings of Brownian motion admit simple closed-form expressions.
- We now specialise to standard Brownian motion B_t with $B_0 = 0$.

One-Sided Crossing for Brownian Motion

- Define
$$\tau_a^+ = \inf\{t \geq 0 : B_t > a\}.$$
- For $B_0 = 0$, Feller (1991, f p. 171):
$$\Pr(\tau_a^+ \leq t) = 2\Big(1 - \Phi\Big(\frac{a}{\sqrt{t}}\Big)\Big),$$
where Φ is the standard normal c.d.f.
- Density:
$$f_0(t \mid 0, a) = \frac{a}{\sqrt{2\pi t^3}} \exp\Big(-\frac{a^2}{2t}\Big), \quad t > 0.$$
- This density:
  - Is bounded on $(0, \infty)$,
  - Its shape depends strongly on the level a.

Moments and Interpretation

- For one-sided hitting time τ_a^+:

$$\mathrm{E}(\tau_a^+) = \infty,$$

but

$$\mathrm{E}\left(\frac{1}{\tau_a^+}\right) = \frac{\sigma^2}{a^2}.$$

- Intuition:
  - There is a non-negligible probability of very late crossings, which pushes $\mathrm{E}(\tau_a^+)$ to infinity,
  - But inverse moments remain finite.
- For large a:
  - Crossing events are rare,
  - Short-horizon crossing probability is small.

Applications: Circuit Breakers and Limits

- One-sided crossings are central to market mechanisms such as **circuit breakers**.
- Examples:
  - London Stock Exchange (LSE):
    - Price monitoring halts if prices move beyond static/dynamic thresholds (e.g. $\pm 8\%$ static for FTSE 100).[1]
  - U.S. markets:
    - Market-wide breakers at S&P 500 drops of 7%, 13%, 20%;
    - Single-stock "Limit Up–Limit Down" bands (e.g. $\pm 5\%$ or $\pm 10\%$ over 5 minutes).[2]
  - Mainland China (Jan 2016):
    - CSI 300 halt at -5%, full-day halt at -7%; mechanism later suspended.[3]
- Empirically, thresholds are more often hit early and late in the trading day.

[1] LSE price monitoring thresholds: see FCA circuit-breaker analysis factsheet.

[2] See SEC/Investor.gov description of stock market circuit breakers.

[3] See Reuters coverage of Chinese circuit-breaker suspension January 2016.

Role in Surveillance and Risk Management

- Stopping times and one-sided crossing probabilities are key tools for:
  - Market surveillance and prudential oversight,
  - Risk management (VaR, stress testing),
  - Design of robust trading and hedging strategies.
- Modelling threshold events helps:
  - Anticipate extreme moves,
  - Quantify risk of hitting loss or margin limits,
  - Engineer products with barrier features (e.g. knock-in/out options).

Table of Contents

Markov Processes

Markov Process

A stochastic process $\{X_t,\ t \in [0,\ T]\}$ is a **Markov process** if, for all $t > s$ and all x,

$$\Pr(X_t \leq x \mid \mathcal{F}_s) = \Pr(X_t \leq x \mid X_s),$$

where $\mathcal{F}_s$ is the σ-algebra generated by the process up to time s.

- Future depends on the present state only, not on the full past path.
- Brownian motion and many diffusion models used in finance are Markov.

Diffusion Processes

Diffusion Process

A **diffusion process** is a continuous-time Markov process with continuous sample paths. It is used to model systems whose state evolves randomly but continuously over time.

- The **strong Markov property** extends Markov behaviour to random times (stopping times) τ:
  - For any stopping time τ and $t \geq 0$, the conditional distribution of $X_{\tau+t}$ given X_τ is independent of the path before τ.
- In finance:
  - Diffusions are the backbone of continuous-time interest rate and asset price models.

SDEs: Drift and Diffusion

- Diffusions are typically specified via stochastic differential equations (SDEs):

$$dX_t = \mu(X_t, t)\, dt + \sigma(X_t, t)\, dB_t,$$

  where:
  - $\mu(X_t, t)$ is the **drift** (deterministic trend component),
  - $\sigma(X_t, t)$ is the **diffusion** (volatility / noise component),
  - B_t is standard Brownian motion.
- SDEs provide a rigorous way to combine:
  - Deterministic dynamics,
  - Random shocks evolving continuously in time.

Integral Form of a Diffusion Process

Integral Representation

Let X_0 be a given random variable and B_t a standard Brownian motion. Consider

$$dX_t = \mu(X_t, t)\, dt + \sigma(X_t, t)\, dB_t.$$

Then X_t can be written as:

$$X_t = X_0 + \int_0^t \mu(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dB_s,$$

where:

- The first integral is a (deterministic) Riemann integral,
- The second is an Itô stochastic integral.

Existence and Uniqueness of SDE Solutions

Existence and Uniqueness Theorem

The SDE

$$dX_t = \mu(X_t, t)\, dt + \sigma(X_t, t)\, dB_t$$

admits a unique solution $\{X_t, t \in [0, T]\}$ with continuous paths if:

1. **Moment condition**: $\mathrm{E}(X_0^2) < \infty$.
2. **Lipschitz continuity**: there exists K such that for all x, y,

$$|\mu(x, t) - \mu(y, t)| \leq K|x - y|, \quad |\sigma(x, t) - \sigma(y, t)| \leq K|x - y|.$$

3. **Linear growth**: there exists K such that for all x,

$$|\mu(x, t)| \leq K(1 + x^2)^{1/2}, \quad |\sigma(x, t)| \leq K(1 + x^2)^{1/2}.$$

Examples of Diffusion Processes: I

- **Black–Scholes Model** (geometric Brownian motion):

$$dX_t = \beta X_t\, dt + \sigma X_t\, dB_t,$$

used to model asset prices in option pricing.

- **Ornstein–Uhlenbeck Process** (Vasicek interest rate model) vasicek1977equilibrium:

$$dX_t = \beta(\alpha - X_t)\, dt + \sigma\, dB_t,$$

where α is the long-term mean, β the speed of mean reversion.

- **Feller Square-Root Process** (CIR model) cox1985intertemporal:

$$dX_t = \beta(\alpha - X_t)\, dt + \sigma\sqrt{X_t}\, dB_t,$$

ensuring X_t stays non-negative (often used for interest rates).

Examples of Diffusion Processes: II

- **Courtadon Model** Courtadon1982:

$$dX_t = \beta(\alpha - X_t)\,dt + \sigma X_t\,dB_t.$$

- **Marsh Model** Marsh1983:

$$dX_t = \left(\alpha X_t^{-(1-\delta)} + \beta\right)dt + \sigma X_t^{\delta/2}\,dB_t.$$

- **Cox Process** Cox1975:

$$dX_t = \beta(\alpha - X_t)\,dt + \sigma X_t^{\gamma}\,dB_t,$$

a generalization of OU with state-dependent volatility.

Examples of Diffusion Processes: III

- **Constantinides Model** constantinides1992theory:

$$dX_t = (\alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2)\, dt + (\sigma_0 + \sigma_1 X_t)\, dB_t.$$

- **Affine Models** duffie1996yield,dai2000specification:

$$dX_t = \beta(\alpha - X_t)\, dt + \sqrt{\sigma_0 + \sigma_1 X_t}\, dB_t,$$

  widely used in term-structure modelling.

- **Nonlinear Mean-Reversion Models** ait1996testing:

$$dX_t = \left(\alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2 + \frac{\alpha_{-1}}{X_t}\right) dt + \left(\beta_0 + \beta_1 X_t + \beta_2 X_t^{\beta_3}\right) dB_t.$$

Diffusions in Finance: Summary

- Diffusion processes provide flexible continuous-time models for:
  - Asset prices and returns,
  - Short rates and term structures,
  - Volatility and stochastic factors.
- Key features:
  - Markov and strong Markov properties,
  - Continuous paths driven by Brownian motion,
  - Rich behaviour via state-dependent drift and diffusion.
- These SDE-based models underlie:
  - Option pricing and hedging,
  - Term-structure models,
  - Continuous-time portfolio choice and risk management.

Table of Contents

MLE for Diffusion Processes: Setup

- We consider continuous-time Markov diffusions observed discretely.
- Observations at times $t_0 < t_1 < \cdots < t_n$:

$$\{X_0, X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}\}, \quad \Delta = t_i - t_{i-1}.$$

- By Markov property, joint likelihood factorises:

$$\begin{aligned}&\Pr(X_{n\Delta}, \ldots, X_0; \theta) \\ &= \Pr(X_{n\Delta} \mid X_{(n-1)\Delta}; \theta) \cdots \Pr(X_\Delta \mid X_0; \theta) \Pr(X_0; \theta),\end{aligned}$$

  where $p_X(\Delta; X_{i\Delta} \mid X_{(i-1)\Delta}; \theta)$ is the **transition density**.
- Idea: plug transition density into log-likelihood, maximise over θ.

General Diffusion and Log-Likelihood

- Diffusion SDE:

$$dX_t = \mu(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dB_t.$$

- Transition density:

$$p_X(\Delta, x \mid x_0;\theta) = \text{density of } X_{t+\Delta} = x \mid X_t = x_0.$$

- Log-likelihood for discrete data:

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log p_X\left(\Delta; X_{i\Delta} \mid X_{(i-1)\Delta};\theta\right).$$

- MLE:

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

General Diffusion and Log-Likelihood

- Under standard regularity:

$$\sqrt{n}(\widehat{\theta} - \theta) \Longrightarrow N\big(0, \mathcal{I}^{-1}(\theta)\big),$$

where

$$\mathcal{I}(\theta) = \lim_{n\to\infty} -\mathrm{E}\Big[\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial\theta\partial\theta'}\Big].$$

Fisher Information and Variance

- Asymptotic variance:

$$\mathrm{Var}(\widehat{\theta}) \approx \frac{1}{n}\mathcal{I}^{-1}(\theta).$$

- Sample estimate:

$$\widehat{\mathcal{I}} = \frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\widehat{\theta})}{\partial\theta\partial\theta'}.$$

- Standard errors:

$$\widehat{\mathrm{se}}(\widehat{\theta}) = \sqrt{\mathrm{diag}\left(\widehat{\mathcal{I}}^{-1}\right)}.$$

- In practice:
  - Use numerical optimisation to find $\widehat{\theta}$,
  - Use numerical derivatives or automatic differentiation for $\nabla\mathcal{L}$ and Hessian.

Black–Scholes Diffusion and Log Returns

- Black–Scholes SDE:

$$dX_t = \beta X_t\, dt + \sigma X_t\, dB_t.$$

- Apply Itô to $\log X_t$:

$$d\log X_t = \left(\beta - \tfrac{1}{2}\sigma^2\right) dt + \sigma\, dB_t = \alpha\, dt + \sigma\, dB_t,$$

where $\alpha = \beta - \frac{1}{2}\sigma^2$.

- Continuously compounded return over Δ:

$$r_t(\Delta) = \log \frac{X_t}{X_{t-\Delta}}.$$

- Then

$$r_t(\Delta) \sim N(\alpha\Delta, \sigma^2\Delta),$$

i.i.d. across t.

BS Log-Likelihood and MLEs

Log-likelihood for $\{r_1(\Delta), \ldots, r_n(\Delta)\}$:

$$\mathcal{L}_n(\alpha, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2\Delta) - \frac{1}{2\sigma^2\Delta}\sum_{t=1}^{n}\big(r_t(\Delta) - \alpha\Delta\big)^2.$$

MLEs:

$$\widehat{\alpha} = \frac{1}{n\Delta}\sum_{t=1}^{n} r_t(\Delta),$$

$$\widehat{\sigma}^2 = \frac{1}{n\Delta}\sum_{t=1}^{n}\big(r_t(\Delta) - \widehat{\alpha}\Delta\big)^2.$$

Because returns are i.i.d. Gaussian:

- Regularity conditions hold,
- MLEs are consistent, asymptotically normal,
- They are also asymptotically efficient among CUAN estimators.

Other Models with Closed-Form Transition Densities

- **Ornstein–Uhlenbeck** (Vasicek) process:

$$dX_t = \beta(\alpha - X_t)\,dt + \sigma\,dB_t.$$

  - Gaussian transition density,
  - MLE straight from normal likelihood.

- **Feller Square-Root** (CIR) process:

$$dX_t = \beta(\alpha - X_t)\,dt + \sigma\sqrt{X_t}\,dW_t.$$

  - Noncentral chi-squared transition density,
  - MLE based on noncentral χ^2 density.

- In these cases, exact transition densities $\Rightarrow$ exact MLE feasible.

Why Approximate the Transition Density?

- In many diffusion models, $p_X(\Delta, x \mid x_0; \theta)$ has no closed form.
- Approximate methods are needed:
  - Simple normal approximation often poor for finite Δ,
  - Direct Edgeworth or similar expansions unstable when tails are heavy (e.g. geometric Brownian motion).
- Ait-Sahalia (1996), Aït-Sahalia (2002), and Ait-Sahalia, Fan, and Peng (2009) propose:
  - Transform diffusion to be "more Gaussian",
  - Use Hermite expansions in transformed space,
  - Map back to original state space.

Aït-Sahalia's 3-Step Approximation

1. **Transform** $X \mapsto Y$ to remove state-dependent volatility:

$$dX_t = \mu(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dB_t$$

becomes

$$Y_t = \gamma(X_t;\theta) = \int^{X_t} \frac{du}{\sigma(u;\theta)},$$

and

$$dY_t = \mu_Y(Y_t;\theta)\,dt + dB_t.$$

2. **Standardize** $Y \mapsto Z$:

$$Z_t = \Delta^{-1/2}(Y_t - y_0).$$

Then approximate the density of Z via Hermite expansion.

3. **Map back** $Z \mapsto Y \mapsto X$:
   - Use change-of-variable formulas to obtain approximate p_X.

First Transformation: $X \to Y$

- Original SDE:
$$dX_t = \mu(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dB_t.$$
- Define:
$$Y_t = \gamma(X_t;\theta) = \int^{X_t} \frac{du}{\sigma(u;\theta)}.$$
- Apply Itô's lemma:
$$dY_t = \mu_Y(Y_t;\theta)\,dt + dB_t,$$
where
$$\mu_Y(y;\theta) = \frac{\mu(\gamma^{-1}(y;\theta);\theta)}{\sigma(\gamma^{-1}(y;\theta);\theta)} - \frac{1}{2}\frac{\partial\sigma}{\partial x}(\gamma^{-1}(y;\theta);\theta).$$
- Result: unit diffusion (constant variance), but more complex drift.

Second Transformation: $Y \to Z$ and Hermite Expansion

- Standardize:
$$Z_t = \Delta^{-1/2}(Y_t - y_0).$$
- For Z close to standard normal, approximate its density via Hermite polynomials:
$$H_j(z) = e^{z^2/2}\frac{d^j}{dz^j}e^{-z^2/2}, \quad \phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}.$$
- Hermite expansion (order J):
$$p_Z^{(J)}(\Delta, z \mid y_0; \theta) = \phi(z)\sum_{j=0}^{J}\eta_j(\Delta, y_0; \theta)H_j(z),$$

  where
$$\eta_j(\Delta, y_0; \theta) = \frac{1}{j!}\int_{-\infty}^{\infty} H_j(z)\, p_Z(\Delta, z \mid y_0; \theta)\, dz.$$

Back-Transformation to X

- Once $p_Z^{(J)}$ is known, approximate density of Y:

$$p_Y^{(J)}(\Delta, y \mid y_0; \theta) = \Delta^{-1/2} p_Z^{(J)}\big(\Delta, \Delta^{-1/2}(y - y_0) \mid y_0; \theta\big).$$

- Approximate density of X via change-of-variable:

$$p_X^{(J)}(\Delta, x \mid x_0; \theta) = \sigma(x; \theta)^{-1} p_Y^{(J)}\big(\Delta, \gamma(x; \theta) \mid \gamma(x_0; \theta); \theta\big).$$

Convergence Theorem

There exists $\overline{\Delta} > 0$ such that for all $\Delta \in (0, \overline{\Delta})$, $\theta \in \Theta$, and (x, x_0) in the state space,

$$p_X^{(J)}(\Delta, x \mid x_0; \theta) \to p_X(\Delta, x \mid x_0; \theta) \quad \text{as } J \to \infty,$$

uniformly in θ, x and x_0 (on compacts).

Approximate MLE via Likelihood Expansion

- Approximate log-likelihood:

$$\mathcal{L}_n^{(J)}(\theta) = \sum_{i=1}^{n} \log p_X^{(J)}\left(\Delta, X_{i\Delta} \mid X_{(i-1)\Delta}; \theta\right).$$

- Approximate MLE:

$$\widehat{\theta}_n^{(J)} = \arg\max_{\theta \in \Theta} \mathcal{L}_n^{(J)}(\theta).$$

- As $J \to \infty$, $\widehat{\theta}_n^{(J)} \to \widehat{\theta}_n$:
  - Approximate MLE inherits asymptotic properties of true MLE,
  - In practice, low J (e.g. $J = 3$ or 4) often suffices.

Closed-Form Density Expansion

Closed-Form Expansion (Aït-Sahalia)

For Z and Y as defined earlier:

$$\widetilde{p}_Z^{(K)}(\Delta, z \mid y_0; \theta) = \Delta^{-1/2}\phi\left(\frac{y - y_0}{\Delta^{1/2}}\right)\exp\left(\int_{y_0}^{y} \mu_Y(\omega; \theta)\, d\omega\right) \times \sum_{k=0}^{K} c_k(y \mid y_0; \theta)\frac{\Delta^k}{k!},$$

with $c_0(y \mid y_0; \theta) = 1$ and

$$\lambda_Y(y; \theta) = -\tfrac{1}{2}\left(\mu_Y^2(y; \theta) + \partial_y \mu_Y(y; \theta)\right).$$

Higher c_k given by recursive integral expressions.

- Coefficients c_k are computable in closed form for many models.
- Provides highly accurate and fast approximate densities.

Specification Testing via Transition Densities

- Continuous-time SDE is specified by (μ, σ), but we observe only discrete-time:
  - Marginal density π_X,
  - Transition density p_X.
- Ait-Sahalia (1996) and Ait-Sahalia, Fan, and Peng (2009) propose tests based on:
  - Comparing parametric $p_X(y \mid x, \Delta; \theta)$ to empirical or nonparametric counterparts.
- E.g. under H_0 (correct model),

$$U_i = P_X(X_{i\Delta} \mid X_{(i-1)\Delta}, \Delta, \theta)$$

should be i.i.d. $U(0, 1)$.

Nonparametric vs Parametric Transition Density

- Hypothesis:

$$H_0 : p_X(y \mid x, \Delta) = p_X(y \mid x, \Delta, \theta),$$
$$H_1 : p_X(y \mid x, \Delta) \neq p_X(y \mid x, \Delta, \theta).$$

- Strategy (Ait-Sahalia, Fan, and Peng (2009)):
  - Estimate $p_X(y \mid x, \Delta)$ nonparametrically (kernel),
  - Compare to parametric $p_X(y \mid x, \Delta, \widehat{\theta})$.
- Related approaches:
  - PIT-based tests,
  - Other kernel-based tests (e.g. Chen, Gao, and Tang 2008).

MLEMVD: Practical MLE for Diffusions

- `MLEMVD` (GitHub: `mfrdixon/MLEMVD`) implements:
  - Maximum likelihood estimation of multivariate diffusion models,
  - Model specification tests and diagnostics.
- Supports:
  - Geometric Brownian motion (GBM),
  - CIR, Heston, and other interest rate/volatility models.
- Features:
  - Approximate and exact likelihoods,
  - Score, Hessian, information matrices,
  - Robust (sandwich) standard errors.

Table of Contents

Quadratic Variation and Volatility

Quadratic Variation

For a square-integrable process X_t, the quadratic variation over $[0, t]$ is

$$\langle X, X \rangle_{0:t} = \text{plim}_{\max(t_{k+1}-t_k)\to 0} \sum_{t_k \le t} |X_{t_{k+1}} - X_{t_k}|^2,$$

where $0 = t_1 < \cdots < t_n = t$.

- For smooth (bounded variation) functions, $QV = 0$.
- For diffusions $dX_t = \mu(X_t)dt + \sigma(X_t)dB_t$:
  - Quadratic variation typically random,
  - If $\sigma(X_t) \equiv \sigma$ constant, $QV_t = \sigma^2 t$.

QV and Conditional Variance

- Under general conditions andersen2003modeling:

$$\mathrm{E}\big[(X(t+h)-X(t))^2 \mid \mathcal{F}_t\big] = \mathrm{E}\big(\langle X, X\rangle_{t:t+h} \mid \mathcal{F}_t\big).$$

- Interpretation:
  - Conditional variance of returns over $[t, t+h]$ equals conditional expectation of QV over the same interval,
  - QV is the natural measure of return volatility.

Local Martingales and Semimartingales

Local Martingale

A process M is a **local martingale** if there exists an increasing sequence of stopping times τ_k with $\tau_k \to \infty$ a.s., such that the stopped process $M_{t \wedge \tau_k}$ is a martingale for each k.

Local Martingales and Semimartingales

Semimartingale

A real-valued process X is a **semimartingale** if it can be decomposed as

$$X(t) = M(t) + A(t),$$

where M is a local martingale and A is an adapted process of locally bounded variation with càdlàg paths.

- Semimartingales are the most general class for which stochastic integration is well-defined.
- Includes:
  - All martingales,
  - Diffusions and jump-diffusions,
  - Many asset price models.

Semimartingales and Finance

- Key properties:
  - Girsanov's theorem holds for semimartingales,
  - Fundamental theorem of asset pricing:
    - No-arbitrage $\Leftrightarrow$ existence of equivalent martingale measure,
    - Asset prices modelled as semimartingales under P and martingales under Q.
- Quadratic variation exists for any continuous square-integrable semimartingale.
- High-frequency volatility estimation is often built on this semimartingale framework.

Realized Volatility as QV Estimator

- Observe n equally spaced log-prices X_t on $[0, 1]$:

$$0, \frac{1}{n}, \ldots, \frac{n}{n} = 1.$$

- **Realized volatility** (RV):

$$RV_X^n = \sum_{l=1}^{n-1} \left(X_{\frac{l+1}{n}} - X_{\frac{l}{n}} \right)^2.$$

- As $n \to \infty$, $RV_X^n \to \langle X, X \rangle_{0:1}$ in probability (under mild conditions).
- RV is thus a consistent estimator of QV.

Asymptotic Distribution of RV

- For Itô semimartingales, Jacod and Protter (1998) derived CLTs for RV_X^n.
- Barndorff-Nielsen and Shephard (2002) specialise to Brownian semimartingales.

Brownian Semimartingale

A process $(X_t)_{t\geq 0}$ is a Brownian semimartingale if

$$X_t = \int_0^t \mu_v \, dv + \int_0^t \sigma_v \, dB_v,$$

with predictable processes (μ_v), (σ_v) and càdlàg (σ_v).

Asymptotic Distribution of RV

- In **no-leverage** case (drift/vol independent of B):

$$n^{1/2}(RV_X^n - QV) \Longrightarrow \sqrt{2}\int_0^1 \sigma_u^2\, dB_u,$$

  a mixed normal limit with variance $2\int_0^1 \sigma_u^4\, du$.

Integrated Quarticity and CLT

- Integrated quarticity:

$$IQ = \int_0^1 \sigma_u^4 \, du.$$

- Estimator:

$$\widehat{IQ} = \frac{n}{3} \sum_{i=1}^{n-1} \left(X_{\frac{i+1}{n}} - X_{\frac{i}{n}} \right)^4.$$

- CLT for RV (Barndorff-Nielsen and Shephard (2002)):

$$\frac{n^{1/2}(RV_X^n - QV)}{\sqrt{2}\,\widehat{IQ}^{1/2}} \Longrightarrow N(0,1).$$

- Uses:
  - Construct confidence intervals for integrated volatility,
  - Formal hypothesis tests on volatility dynamics.

Table of Contents

Market Microstructure Noise and Measurement Error

- High-frequency data are contaminated by **market microstructure noise**:
  - Discreteness of trades and quotes,
  - Order-processing lags, bid–ask bounce,
  - Tick size constraints, trading rules, fragmentation across venues.
- This noise complicates:
  - Volatility estimation,
  - Cross-asset correlation and covariance analysis,
  - Market efficiency tests.
- We model observed log prices as:

$$Y_{t_j} = X_{t_j} + \varepsilon_{t_j},$$

where X is efficient price, ε microstructure noise.

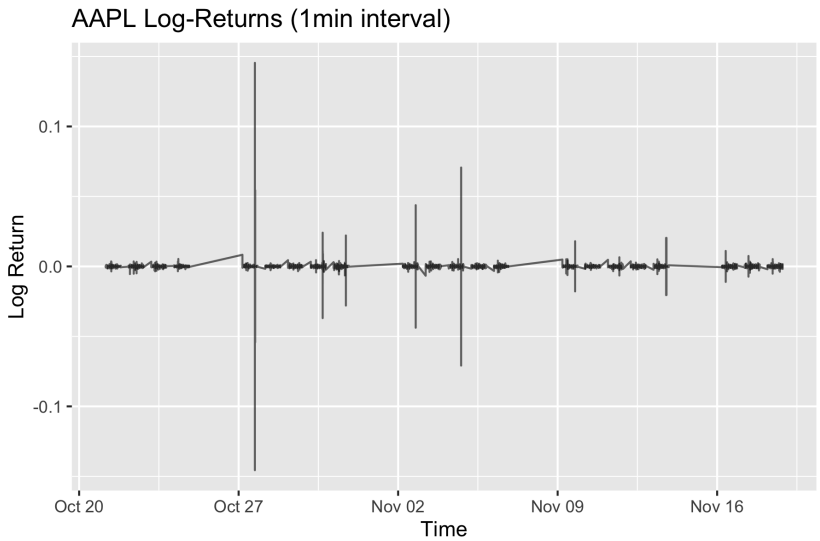
Alpha Vantage and High-Frequency AAPL Data

Alpha Vantage

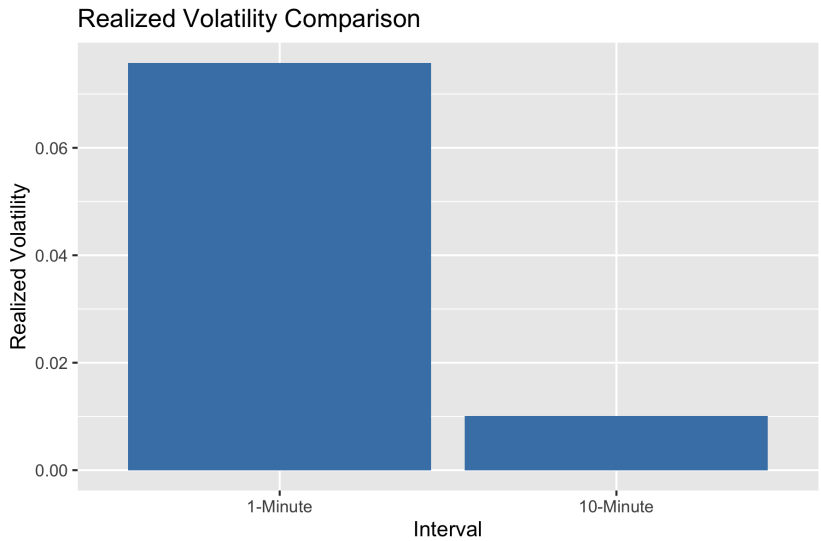
- Provides real-time and historical data for stocks, FX, crypto, etc.
- Access via free/premium API key (rate-limited).
- Get a key from: `https://www.alphavantage.co/support/#api-key`

Note: We requested the maximum number of available intraday data points from Alpha Vantage. Free accounts are subject to data and rate limits, so the exact sample depends on when the code is run. The plots shown here were generated on 20 November 2025; if re-run on a different date, the exact values may change, but the qualitative pattern should persist due to underlying microstructure noise.

AAPL Returns and RV: Plots



(a) AAPL log-returns (1-minute interval).



(b) Realized volatility at 1-min vs 10-min intervals.

Figure: Comparison of AAPL (Apple Inc.) Log-Returns and Realized Volatility

Measurement Error Model for Prices

- Observed log prices:

$$Y_{t_j} = X_{t_j} + \varepsilon_{t_j},$$

  where:

  - X_{t_j} latent efficient log-price,
  - ε_{t_j} i.i.d. noise, $\mathrm{E}\varepsilon_{t_j} = 0$, $\mathrm{Var}(\varepsilon_{t_j}) = \sigma_\varepsilon^2$,
  - ε_{t_j} independent of X_{t_j}.

- This is a stylised model of microstructure noise (Zhang, Mykland, and Aït-Sahalia 2005).
- Over long horizons, X dominates; at high frequencies, ε can dominate RV.

Effect of Noise on Realized Volatility

- RV based on observed Y:

$$RV_Y^n = \sum_{i=1}^{n-1} \left(Y_{\frac{i+1}{n}} - Y_{\frac{i}{n}}\right)^2$$
$$= \underbrace{\sum_{i=1}^{n-1} \left(X_{\frac{i+1}{n}} - X_{\frac{i}{n}}\right)^2}_{\text{efficient QV}} + \underbrace{\sum_{i=1}^{n-1} \left(\varepsilon_{\frac{i+1}{n}} - \varepsilon_{\frac{i}{n}}\right)^2}_{\text{noise term}}$$
$$+ 2 \sum_{i=1}^{n-1} \left(\varepsilon_{\frac{i+1}{n}} - \varepsilon_{\frac{i}{n}}\right)\left(X_{\frac{i+1}{n}} - X_{\frac{i}{n}}\right).$$

- By LLN:

$$\frac{1}{n} \sum_{i=1}^{n-1} \left(\varepsilon_{\frac{i+1}{n}} - \varepsilon_{\frac{i}{n}}\right)^2 \xrightarrow{p} 2\sigma_\varepsilon^2.$$

- Cross term $\to 0$ in probability (Cauchy–Schwarz).

Inconsistency of Naive RV with Noise

- Leading behaviour:

$$RV_Y^n \approx \underbrace{QV_X}_{O(1)} + \underbrace{2n\sigma_\varepsilon^2}_{O(n)}.$$

- As $n \to \infty$:

$$RV_Y^n \xrightarrow{p} \infty.$$

- Conclusion:
  - Naive realized volatility is **not** a consistent estimator of volatility in the presence of microstructure noise.
- Simple fix: downsample to lower frequency (but wastes data).
- Better: use noise-robust estimators such as TSRV, MSRV, realized kernels, pre-averaging.

Two-Scale Realized Volatility: Idea

- Partition high-frequency sample into K subsamples:
  - Subsample j: $\{Y_0, Y_{K/n}, \ldots, Y_{mK/n}\}$,
  - Each subsample has $m+1$ prices and m returns,
  - $K(m+1) = n$.
- For subsample j, define:

$$RV_{\text{sub}_j} = \sum_{i=1}^{m} \left(Y_{\frac{j+iK}{n}} - Y_{\frac{j+(i-1)K}{n}} \right)^2.$$

- These are realized volatilities at a coarser time scale.

Bias Decomposition and Correction

- With measurement error:

$$\frac{1}{m} RV_{\text{sub}_j} = \frac{1}{m} \sum_{i=1}^{m} (X\text{-increments})^2 + \frac{1}{m} \sum_{i=1}^{m} (\varepsilon\text{-increments})^2 + \text{cross term}.$$

- As $m \to \infty$:

$$\frac{1}{m} RV_{\text{sub}_j} \xrightarrow{p} 2\sigma_\varepsilon^2.$$

- Similarly, full-sample RV has leading noise term $2n\sigma_\varepsilon^2$.
- Bias-corrected combination:

$$RV_{\text{sub}_j} - \frac{m}{n} RV_n,$$

has *leading-order noise bias cancel*:

$$2m\sigma_\varepsilon^2 - \frac{m}{n} \cdot 2n\sigma_\varepsilon^2 = 0.$$

TSRV Estimator

Two-Scale Realized Volatility (TSRV)

Let $K(m+1) = n$ and RV_n be full-sample RV. Then

$$\widehat{\theta}_{\text{TSRV}} = \frac{1}{K}\sum_{j=1}^{K} RV_{\text{sub}_j} - \frac{m}{n} RV_n,$$

is the TSRV estimator of integrated volatility.

- Averaging across subsamples reduces noise variance:

$$T = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{m}\left\{(\varepsilon\text{-increments})^2 - 2\sigma_\varepsilon^2\right\} = O_p\left(\sqrt{m/K}\right).$$

- With suitable K, m (e.g. $K \asymp n^{1/2}$), TSRV is consistent.

Multi-Scale RV and Extensions

- **Multi-Scale Realized Volatility (MSRV)** zhang2006efficient:

$$\widehat{\theta}_{\text{MSRV}} = \sum_{\mathcal{L}=1}^{L} \alpha_{\mathcal{L}} \frac{1}{K_{\mathcal{L}}} \sum_{j=1}^{K_{\mathcal{L}}} RV_{\text{sub}_j}^{K_{\mathcal{L}}},$$

  with weights $\alpha_{\mathcal{L}}$ satisfying constraints.
- Achieves optimal convergence rate (matching Gaussian MLE) in constant-volatility case.
- Aït-Sahalia, Mykland, and Zhang (2011): TSRV and MSRV remain consistent under serially correlated noise.
- Other approaches:
  - Pre-averaging,
  - Realized kernels.

Table of Contents

Challenges in High-Frequency Covariance Estimation

- Multivariate setting: estimate covariance matrix of returns from HF data.
- Complications:
  1. **Bid–ask bounce** and noise:
     - Transaction prices differ from efficient prices,
     - Distorts both variances and covariances.
  2. **Price discreteness**:
     - Decimal/tick grid induces apparent jumps.
  3. **Asynchronicity**:
     - Assets trade at different times,
     - High-frequency sampling $\Rightarrow$ few truly synchronous trades.
- Consequence: **Epps effect**—off-diagonal covariances shrink toward 0 as frequency increases.

Synchronization and Refresh Times

- Need common time grid across assets for realized covariance.
- Two broad approaches:
  - Interpolation-based (fill missing prices) — often performs poorly,
  - Deletion/matching-based — more robust.
- **Refresh time method** (Harris et al. 1995):
  - Construct times at which each asset has traded at least once since last refresh,
  - Produces a sequence of synchronized timestamps.

Refresh Times (Two Assets)

Let X and Y observed at times Γ and Θ.

- First refresh: $\phi_1 = \max(\tau_1, \theta_1)$.
- Next: $\phi_{j+1} = \max(\tau_{N^X_{\phi_j}+1}, \theta_{N^Y_{\phi_j}+1})$.

TSX: Two-Scale Realized Covariance

- TSX (Zhang 2011) extends TSRV to covariance matrices.
- For variance of a log-price series X with n observations:

$$[X,X]_T^{(K)} = \frac{1}{K}\sum_{i=1}^{n-K+1}(X_{t_{i+K}} - X_{t_i})^2,$$

  and similarly $[X,X]_T^{(J)}$.
- Variance estimate:

$$\left(1-\frac{\bar{n}_K}{\bar{n}_J}\right)^{-1}\left([X,X]_T^{(K)} - \frac{\bar{n}_K}{\bar{n}_J}[X,X]_T^{(J)}\right),$$

  with $\bar{n}_K = (n-K+1)/K$, $\bar{n}_J = (n-J+1)/J$.

TSX Covariance via Refresh Times

- For covariance between X and Y:
  - Use refresh-time synchronization to obtain (X_{t_i}, Y_{s_i}) pairs.
  - Define

$$[X, Y]_T^{(K)} = \frac{1}{K} \sum_{i=1}^{M_N - K + 1} (X_{t_{i+K}} - X_{t_i})(Y_{s_{i+K}} - Y_{s_i}),$$

    where M_N is number of synchronized returns.

- TSX covariance:

$$c_N \left([X, Y]_T^{(K)} - \frac{\bar{n}_K}{\bar{n}_J} [X, Y]_T^{(J)} \right),$$

  with appropriate scaling constant c_N.

- TSX is not guaranteed PSD; use `makePsd=TRUE` in `highfrequency::rTSCov`.

Multivariate Realized Kernel (RKX)

Refresh Times for d Assets

$$\tau_1 = \max\left(t_1^{(1)}, \ldots, t_1^{(d)}\right),$$

$$\tau_{j+1} = \max\left(t^{(1)}_{N^{(1)}_{\tau_j}+1}, \ldots, t^{(d)}_{N^{(d)}_{\tau_j}+1}\right).$$

Multivariate Realized Kernel (RKX)

- Retention ratio:

$$p = \frac{dN}{\sum_{i=1}^{d} n^{(i)}},$$

  measures data kept after synchronization.

- Jittering at start/end:

$$X_0 = \frac{1}{m}\sum_{j=1}^{m} X(\tau_j), \quad X_n = \frac{1}{m}\sum_{j=1}^{m} X(\tau_{N-m+j}),$$

  with $m \approx 2$.

- Returns: $\boldsymbol{r}_j = X_j - X_{j-1}$.

RKX Estimator Definition

Multivariate Realized Kernel (RKX) barndorff2011multivariate

Given synchronized returns $\{\boldsymbol{r}_j\}$,

$$K(X) = \sum_{h=-n}^{n} k\Big(\frac{h}{H+1}\Big)\Gamma_h,$$

with

$$\Gamma_h = \begin{cases} \sum_{j=|h|+1}^{n} \boldsymbol{r}_j \boldsymbol{r}'_{j-h}, & h \geq 0, \\ \sum_{j=|h|+1}^{n} \boldsymbol{r}_{j-h} \boldsymbol{r}'_j, & h < 0. \end{cases}$$

RKX Estimator Definition

- Kernel $k(\cdot)$ must satisfy:
  - $k(0) = 1$, $k'(0) = 0$,
  - Smoothness and bounded moment conditions,
  - Nonnegative Fourier transform $\Rightarrow$ PSD.
- Bandwidth H chosen to minimise MSE.
- RKX has fast convergence rate $\widetilde{n}^{-1/5}$ and is PSD by construction.

Pre-Averaging Covariance Estimator (PAVX)

- PAVX (Christensen, Kinnebrock, and Podolskij 2010; Hautsch and Podolskij 2013) combines:
  - Noise reduction via pre-averaging (like kernels),
  - Bias correction (like TSX).
- Model:

$$Y_\tau = X_\tau + \epsilon_\tau,$$

  with latent Brownian semimartingale X and i.i.d. noise ϵ independent of X.
- Pre-averaging reduces impact of ϵ when it is i.i.d. with zero mean.

Univariate Pre-Averaging

- Suppose N equidistant returns r_{τ_i} over interval $[0,1]$.
- Define pre-averaged returns:

$$\bar{r}_{\tau_j} = \sum_{h=1}^{k_N-1} g\Big(\frac{h}{k_N}\Big)\, r_{\tau_{j+h}},$$

where $g(x) = \min(x, 1-x)$ and $k_N = \lfloor \theta N^{1/2} \rfloor$.

- Variance estimator (univariate):

$$\hat{C} = \frac{N^{-1/2}}{\theta\psi_2} \sum_{i=0}^{N-k_N+1} \bar{r}_{\tau_i}^2 - \frac{\psi_1^{k_N} N^{-1}}{2\theta^2 \psi_2^{k_N}} \sum_{i=0}^{N} r_{\tau_i}^2,$$

with

$$\psi_1^{k_N} = k_N \sum_{j=1}^{k_N} \big(g(\tfrac{j+1}{k_N}) - g(\tfrac{j}{k_N})\big)^2, \quad \psi_2^{k_N} = \frac{1}{k_N} \sum_{j=1}^{k_N-1} g^2\Big(\frac{j}{k_N}\Big), \quad \psi_2 = \frac{1}{12}.$$

Multivariate PAVX and PSD Variant

- Multivariate PAVX estimator:

$$\mathrm{PAVX} = \frac{N}{N - k_N + 2} \cdot \frac{1}{\psi_2 k_N} \sum_{i=0}^{N-k_N+1} \overline{\boldsymbol{r}}_{\tau_i} \overline{\boldsymbol{r}}'_{\tau_i} - \frac{\psi_1^{k_N}}{\theta^2 \psi_2^{k_N}} \widehat{\Psi}_N,$$

  where $\overline{\boldsymbol{r}}_{\tau_i}$ are pre-averaged returns and

$$\widehat{\Psi}_N = \frac{1}{2N} \sum_{i=1}^{N} \boldsymbol{r}_{\tau_i} \boldsymbol{r}'_{\tau_i}.$$

- Bias-corrected PAVX is not guaranteed PSD.
- PSD PAVX:

$$\mathrm{PAVX}^{\delta} = \frac{N}{N - k_N + 2} \cdot \frac{1}{\psi_2 k_N} \sum_{i=0}^{N-k_N+1} \overline{\boldsymbol{r}}_{\tau_i} \overline{\boldsymbol{r}}'_{\tau_i},$$

  with $k_N = \lfloor \theta N^{1/2+\delta} \rfloor$ and $\delta > 0$ small.

TSX, RKX, PAVX: Comparison

- **TSX**:
  - Bias correction across multiple scales (Epps effect + noise),
  - Not necessarily PSD.
- **RKX**:
  - Kernel-based estimator, PSD by design,
  - Robust to endogenous, serially correlated noise,
  - Convergence rate $\tilde{n}^{-1/5}$.
- **PAVX**:
  - Combines pre-averaging and TSX-type bias correction,
  - Achieves fast convergence rate,
  - PSD variant available via bandwidth choice.
- In R:
  - `highfrequency::rTSCov` (TSX),
  - `highfrequency::rKernelCov` (RKX),
  - `highfrequency::rMRCov` (PAVX); often use `makePsd = TRUE`.

Table of Contents

Continuous-Time and High-Frequency Toolkit: Summary (1)

- We built a continuous-time framework based on:
  - Brownian motion and Itô diffusions,
  - Itô calculus and Itô's Lemma,
  - Standard financial SDEs (GBM, OU/Vasicek, CIR, etc.).
- Estimation:
  - MLE using exact transition densities when available,
  - Aït-Sahalia's Hermite expansions for approximate likelihoods when densities are intractable.

Continuous-Time and High-Frequency Toolkit: Summary (2)

- Volatility from high-frequency data:
  - Quadratic variation (QV) as cumulative volatility,
  - Realized volatility (RV) as discrete QV estimator,
  - Mixed-normal CLTs with integrated quarticity.
- Microstructure noise:
  - Measurement error model $Y_t = X_t + \varepsilon_t$,
  - Naive RV diverges at ultra-high frequencies,
  - Noise-robust estimators (TSRV, MSRV, realized kernels, pre-averaging) restore consistency.

Continuous-Time and High-Frequency Toolkit: Summary (3)

- Multivariate aspects:
  - Epps effect and asynchronicity challenge covariance estimation,
  - Refresh-time sampling provides a common grid,
  - High-frequency covariance estimators: TSX, RKX, PAVX.
- Portfolio applications:
  - With an estimated covariance matrix Σ, one can construct, e.g., the global minimum variance portfolio:

$$\min_{w} w'\Sigma w \quad \text{s.t. } \mathbf{1}'w = 1,$$

  - Possible additional constraints: no shorting, caps, etc.
- R implementations (e.g. `highfrequency`, `MLEMVD`) bridge theory and practice.

Table of Contents

References I

- Aït-Sahalia, Yacine (2002). "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach". In: *Econometrica* 70.1, pp. 223–262.
- Ait-Sahalia, Yacine (1996). "Testing continuous-time models of the spot interest rate". In: *The Review of Financial Studies* 9.2, pp. 385–426.
- Ait-Sahalia, Yacine, Jianqing Fan, and Heng Peng (2009). "Nonparametric transition-based tests for jump diffusions". In: *Journal of the American Statistical Association* 104.487, pp. 1102–1116.
- Aït-Sahalia, Yacine, Per A Mykland, and Lan Zhang (2011). "Ultra high frequency volatility estimation with dependent microstructure noise". In: *Journal of Econometrics* 160.1, pp. 160–175.
- Bachelier, Louis (1900). "Théorie de la spéculation". In: *Annales Scientifiques de l'École Normale Supérieure*. Vol. 17, pp. 21–86.

References II

- Barndorff-Nielsen, Ole E and Neil Shephard (2002). "Estimating quadratic variation using realized variance". In: *Journal of Applied Econometrics* 17.5, pp. 457–477.
- Chen, Song Xi, Jiti Gao, and Cheng Yong Tang (2008). "A test for model specification of diffusion processes". In: *The Annals of Statistics* 36.1.
- Christensen, Kim, Silja Kinnebrock, and Mark Podolskij (2010). "Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data". In: *Journal of Econometrics* 159.1, pp. 116–133.
- Feller, William (1991). *An introduction to probability theory and its applications, Volume 2*. Vol. 81. John Wiley & Sons.
- Harris, Frederick H. deB et al. (1995). "Cointegration, error correction, and price discovery on informationally linked security markets". In: *Journal of Financial and Quantitative Analysis* 30.4, pp. 563–579.

References III

- Hautsch, Nikolaus and Mark Podolskij (2013). "Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence". In: *Journal of Business & Economic Statistics* 31.2, pp. 165–183.
- Jacod, Jean and Philip Protter (1998). "Asymptotic error distributions for the Euler method for stochastic differential equations". In: *The Annals of Probability* 26.1, pp. 267–307.
- Zhang, Lan (2011). "Estimating covariation: Epps effect, microstructure noise". In: *Journal of Econometrics* 160.1, pp. 33–47.
- Zhang, Lan, Per A Mykland, and Yacine Aït-Sahalia (2005). "A tale of two time scales: Determining integrated volatility with noisy high-frequency data". In: *Journal of the American Statistical Association* 100.472, pp. 1394–1411.

Chapter 10 — Selected Machine Learning Tools for Econometrics in R

Yongmiao Hong†, Oliver Linton‡, Jiajing Sun§

Academy of Mathematics and Systems Science, Chinese Academy of Sciences†
Faculty of Economics, University of Cambridge‡
School of Economics and Management, University of Chinese Academy of Sciences§

November 27, 2025

Chapter Overview

- Earlier chapters: classical econometrics
  - Linear regression, ARMA/VAR, GARCH, kernel/nonparametric methods.
  - Researcher specifies a parametric or semi-parametric model, then estimates a relatively small number of parameters.
- Growing empirical challenges:
  - High-dimensional panels (many potential predictors),
  - High-frequency financial data,
  - Unstructured data (text, images, etc.),
  - Evaluation based on out-of-sample performance.
- Machine learning (statistical learning) offers:
  - Rich model classes + regularization,
  - Focus on prediction performance on unseen data,
  - Tools that complement, not replace, classical methods.

Four Recurring Tasks in Applied Work

Machine learning tools are especially useful for:

- **Measurement**:
  - Constructing sentiment, risk, or quality indices from text, images, or high-frequency records.
- **Prediction**:
  - Forecasting future events or continuous outcomes.
- **Support for causal inference**:
  - Flexibly modeling high-dimensional nuisance components.
- **Pattern discovery**:
  - Discovering robust empirical regularities to guide theory and model specification.

Table of Contents

Prediction vs Estimation and Inference

- Classical regression:
  - Parameters (e.g. β in linear models) are the main objects of interest,
  - Interpretation: marginal effects, elasticities, risk exposures,
  - Quality judged by unbiasedness, efficiency, consistency, and validity of tests and confidence intervals.
- Statistical learning:
  - Focus shifts to the **prediction function** $m(x)$,
  - Central question: how well can we predict y for new x?
  - Parameters are a means to construct a prediction rule, not necessarily of direct substantive interest.

Bias–Variance Trade-off

- Model complexity vs prediction error:
  - Simple models: high bias, low variance,
  - Very flexible models: low bias, high variance.
- Overfitting:
  - Adding regressors improves in-sample fit,
  - But may worsen out-of-sample performance by fitting noise.
- Many earlier issues rephrased:
  - Multicollinearity $\Rightarrow$ high variance,
  - Overfitting in nonparametrics,
  - Unstable high-order AR/VAR estimates.
- Machine-learning methods:
  - Provide systematic ways to control model complexity,
  - Explicitly manage the bias–variance trade-off.

Training, Validation, and Cross-Validation

- To assess prediction performance, must evaluate on data not used for estimation.
- Basic sample split:

  Training set used to fit the model.
  Validation set used to tune hyperparameters (e.g. penalties).
  Test set used once at the end to assess performance.
- K-fold cross-validation:
  - Split data into K folds,
  - For each fold: train on $K-1$ folds, validate on the remaining one,
  - Average validation errors; repeat over tuning parameters.

Cross-Validation for Time Series

- Time-series data are dependent over time:
  - Random shuffling destroys temporal structure.
- Use **rolling** or **expanding** windows:
  1. Estimate model on an initial block,
  2. Forecast the next period,
  3. Move window forward, repeat.
- This yields genuine out-of-sample forecasts and respects time ordering.
- Same idea used to tune penalty parameters or complexity controls.

Connections to Earlier Chapters

- Multicollinearity (Chapter 1 Regression Models):
  - Ridge regression shrinks coefficients, reduces variance.
- Nonparametric methods (Chapter 5 Nonparametric Methods):
  - Bandwidth controls smoothness and complexity,
  - Regularization plays an analogous role.
- Time-series (Chapters 2 Univariate Time Series, 3 Multivariate Linear Time Series, 4 Volatility Models):
  - Many lags and variables can overfit,
  - Shrinkage and tree-based algorithms handle many predictors while controlling overfitting.
- Machine learning extends familiar ideas about balancing fit vs complexity.

Table of Contents

Classical Linear Regression Recap

- Model:

$$y_t = X_t'\beta + u_t, \quad t = 1, \ldots, T,$$

where y_t scalar, X_t is $p \times 1$, β is $p \times 1$, u_t error.

- Matrix form:

$$y = \mathbf{X}\beta + u,$$

with y $(T \times 1)$, u $(T \times 1)$, $\mathbf{X}$ $(T \times p)$.

- OLS:

$$\widehat{\beta}^{\mathrm{OLS}} \in \arg\min_{\beta} \|y - \mathbf{X}\beta\|_2^2.$$

- Works well if p small, $\mathbf{X}'\mathbf{X}$ nonsingular.

High-Dimensional Issues

- When p is large relative to T:
  - $\mathbf{X}'\mathbf{X}$ can be ill-conditioned,
  - OLS becomes unstable: small data changes $\Rightarrow$ large changes in $\widehat{\beta}^{\text{OLS}}$.
- If $p > T$:
  - $\mathbf{X}'\mathbf{X}$ is singular,
  - OLS coefficients not unique,
  - Fitted values interpolate data: $\mathbf{X}\widehat{\beta} = y$.
- Overfitting leads to poor out-of-sample performance.
- Regularized linear models add penalties:
  - Control complexity,
  - Exploit sparsity (many predictors may be irrelevant).

Penalized Least Squares: General Form

- Consider:

$$\widehat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{T} \sum_{t=1}^{T} (y_t - X_t'\beta)^2 + \lambda \|\beta\|_q \right\}.$$

- $\lambda \geq 0$: tuning (penalty) parameter.
- $\|\cdot\|_q$: norm (or quasi-norm) on $\mathbb{R}^p$.
- Different $q \Rightarrow$ different estimators:
  - $q = 2^2 \Rightarrow$ ridge regression,
  - $q = 1 \Rightarrow$ lasso,
  - nonconvex penalties (SCAD) use folded-concave $P_T(|\beta_j|)$.

Ridge Regression

- Ridge regression:

$$\widehat{\beta}_{\lambda}^{\text{ridge}} \in \arg\min_{\beta} \left\{ \frac{1}{T} \sum_{t=1}^{T} (y_t - X_t'\beta)^2 + \lambda \|\beta\|_2^2 \right\}.$$

- Closed-form solution:

$$\widehat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda T I_p)^{-1} \mathbf{X}' y.$$

- Properties:
  - Shrinks all coefficients towards zero,
  - Stabilises estimation when columns of $\mathbf{X}$ are highly collinear,
  - Well-defined even when $p > T$ (unlike OLS).

Lasso Regression

- Lasso:

$$\widehat{\beta}_{\lambda}^{\text{lasso}} \in \arg\min_{\beta} \left\{ \frac{1}{T} \sum_{t=1}^{T} (y_t - X_t'\beta)^2 + \lambda \|\beta\|_1 \right\}.$$

- Objective is convex but nondifferentiable at zero; no closed form, but efficient algorithms exist.
- Key feature: **shrinkage + variable selection**.
- Simple orthogonal case:

$$y_t = \beta_t + u_t, \quad t = 1, \ldots, T,$$

$\Rightarrow$ lasso solution:

$$\widehat{\beta}_{\lambda,t}^{\text{lasso}} = \text{sign}(y_t)(|y_t| - \lambda)_+.$$

- Many coefficients set exactly to zero as λ increases.

SCAD Penalty: Motivation

- Lasso:
  - Encourages sparsity,
  - But shrinks large coefficients as strongly as small ones,
  - Can introduce bias for strong signals.
- SCAD (Smoothly Clipped Absolute Deviation) (**fan2001scad**):
  - Folded-concave penalty,
  - Aims to combine sparsity with reduced bias on large coefficients,
  - Penalty function $P_T(|\beta_j|)$ has derivative:

$$P'_T(u) = \lambda \left\{ \mathbf{1}(u \leq \lambda) + \frac{(a\lambda - u)_+}{(a-1)u} \mathbf{1}(u > \lambda) \right\}, \quad u > 0,\ a > 2.$$

SCAD Thresholding Rule

Orthogonal model $y_t = \beta_t + u_t$:

$$\widehat{\beta}_{\lambda,t}^{\mathrm{SCAD}} = \begin{cases} \mathrm{sign}(y_t)(|y_t| - \lambda)_+, & |y_t| \le 2\lambda, \\ \dfrac{(a-1)y_t - \mathrm{sign}(y_t)a\lambda}{a-2}, & 2\lambda < |y_t| < a\lambda, \\ y_t, & |y_t| \ge a\lambda. \end{cases}$$

Behavior:

- Like lasso near zero $\Rightarrow$ sparsity,
- Much weaker shrinkage for large $|y_t|$,
- Under suitable conditions, SCAD enjoys **oracle** properties: selects true nonzero coefficients and estimates them efficiently.

Nonconvex optimisation: more delicate, but practical algorithms exist.

Choosing the Tuning Parameter

- All three methods (ridge, lasso, SCAD) depend on λ:
  - Larger $\lambda \Rightarrow$ stronger shrinkage,
  - Smaller $\lambda \Rightarrow$ closer to OLS.
- SCAD also depends on an extra parameter $a > 2$.
- In practice:
  - λ chosen by cross-validation,
  - For time series, use blocked/rolling cross-validation.
- We choose λ that minimises out-of-sample prediction error.

Further Applications of Regularized Linear Models

- **Return prediction with many predictors**:
  - Firm characteristics, technical indicators, macro variables,
  - Lasso/SCAD can select informative predictors, reduce noise.
- **Volatility modeling**:
  - Regress squared returns or realized volatility on a rich set of lags and exogenous variables,
  - Use ridge/lasso/SCAD to control complexity,
  - Flexible linear alternative to GARCH-type models.
- Regularized linear models:
  - Extend classical regression,
  - Straightforward to implement in R,
  - Serve as building blocks for more advanced high-dimensional and causal ML methods.

Links to Earlier Chapters

- As extensions of OLS:
  - Address multicollinearity (ridge),
  - Perform variable selection (lasso, SCAD),
  - Handle $p \gg T$ situations.
- Relation to nonparametrics:
  - Penalty parameter λ plays role similar to bandwidth,
  - Larger $\lambda \Rightarrow$ smoother (less complex) estimators.
- High-dimensional time series:
  - AR/VAR with many lags and variables become feasible,
  - Cross-validation guards against overfitting.

Table of Contents

Why Tree-Based Methods?

- Regularized linear models remain linear in regressors.
- In many settings:
  - Nonlinearities,
  - Threshold effects,
  - Interactions among variables,
  - Unknown functional forms.
- Tree-based methods:
  - Approximate $m(x) = \mathrm{E}(y \mid X = x)$ by a step function,
  - Automatically capture interactions and nonlinearities,
  - Nonparametric and data-driven.

Regression Trees: Basic Idea

- Observations: (y_t, X_t), $t = 1, \ldots, T$.
- Aim: approximate $m(x) = \mathrm{E}(y_t \mid X_t = x)$.
- Tree partitions covariate space into regions:

$$R_1, \ldots, R_J \subset \mathbb{R}^p, \quad R_j \cap R_\ell = \emptyset, \quad \cup_j R_j = \mathbb{R}^p.$$

- Prediction in region R_j:

$$\widehat{m}_{\text{tree}}(x) = \sum_{j=1}^{J} \widehat{\mu}_j \mathbf{1}\{x \in R_j\}, \quad \widehat{\mu}_j = \frac{1}{T_j} \sum_{t: X_t \in R_j} y_t.$$

- Step function: constant within each leaf region.

CART Algorithm: Recursive Binary Splitting

- Start with root node containing all data.
- At each node:
  - For each regressor k and threshold s, consider split:

$$R_\ell(k,s) = \{x : x_k \leq s\}, \quad R_r(k,s) = \{x : x_k > s\}.$$

  - Choose (k,s) that minimises within-node RSS:

$$\text{RSS}(k,s) = \sum_{t:X_t \in R_\ell} (y_t - \widehat{\mu}_\ell)^2 + \sum_{t:X_t \in R_r} (y_t - \widehat{\mu}_r)^2.$$

- Repeat recursively in child nodes until stopping criteria:
  - Minimum node size,
  - Minimal RSS improvement, etc.

Pruning and Complexity Control

- Fully grown tree tends to overfit:
  - Low bias, high variance.
- CART uses **cost-complexity pruning**:

$$\mathrm{RSS}(T) + \alpha |T|,$$

  where $|T|$ is number of leaves, $\alpha \geq 0$.
- Starting from large tree, sequentially remove branches to obtain nested sequence of trees.
- Use cross-validation to choose α (hence tree size):
  - Small $\alpha \Rightarrow$ large trees,
  - Large $\alpha \Rightarrow$ smaller, simpler trees.

Interpretation and Nonparametric View

- Regression trees are:
  - Easy to visualise (tree diagrams),
  - Naturally handle interactions and threshold effects,
  - Flexible with mixtures of continuous and categorical regressors.
- From nonparametric perspective:
  - Trees are adaptive step-function estimators,
  - Regions R_j analogous to data-driven bins,
  - Similar spirit to histograms and kernel regressions (Chapter 5) but with data-driven partitioning.
- Limitation: a single tree often high-variance and unstable.

A Simple Regression Tree: Partition of the Covariate Space

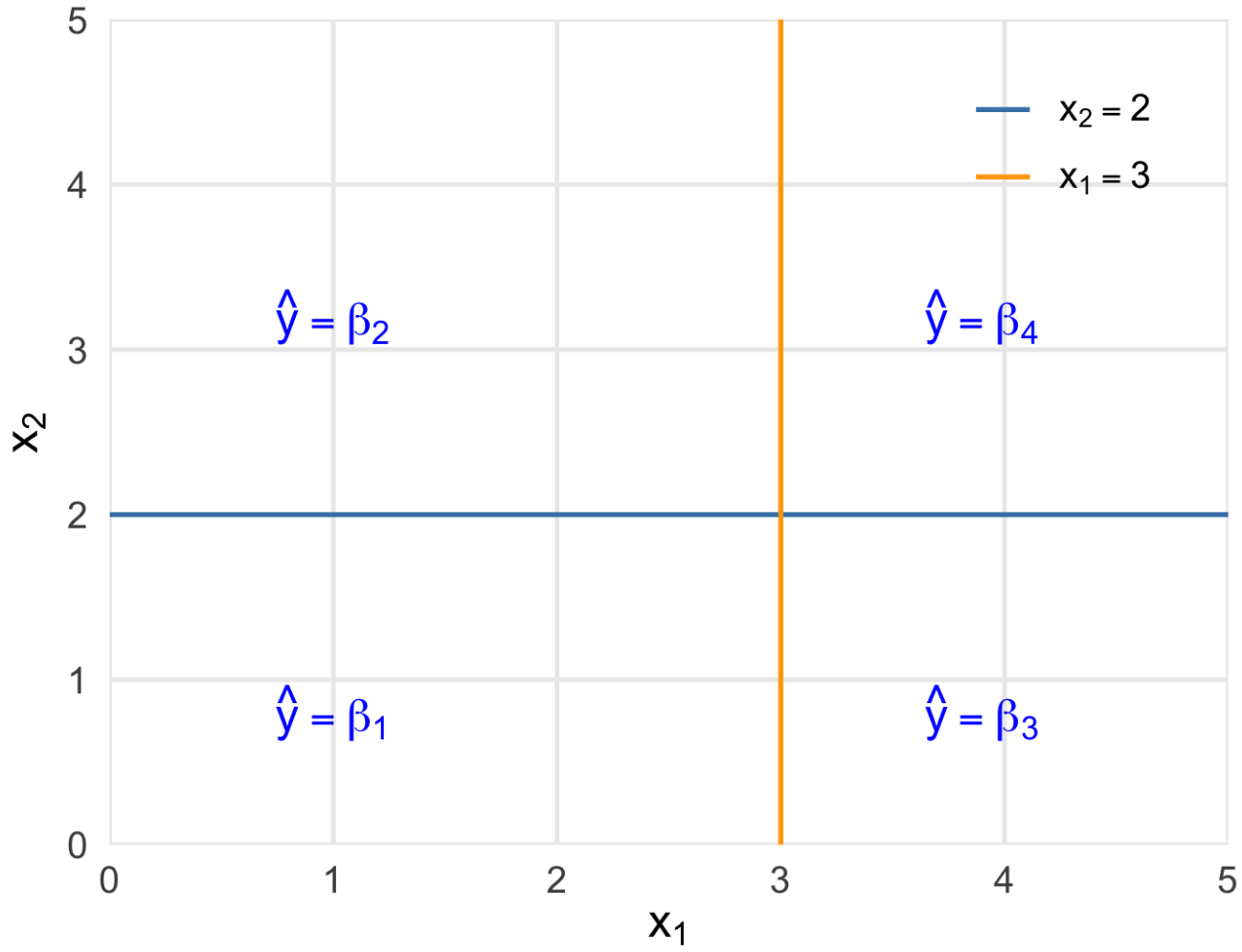


Figure: Partition of the covariate space induced by a simple regression tree.

A Simple Regression Tree: Tree Representation

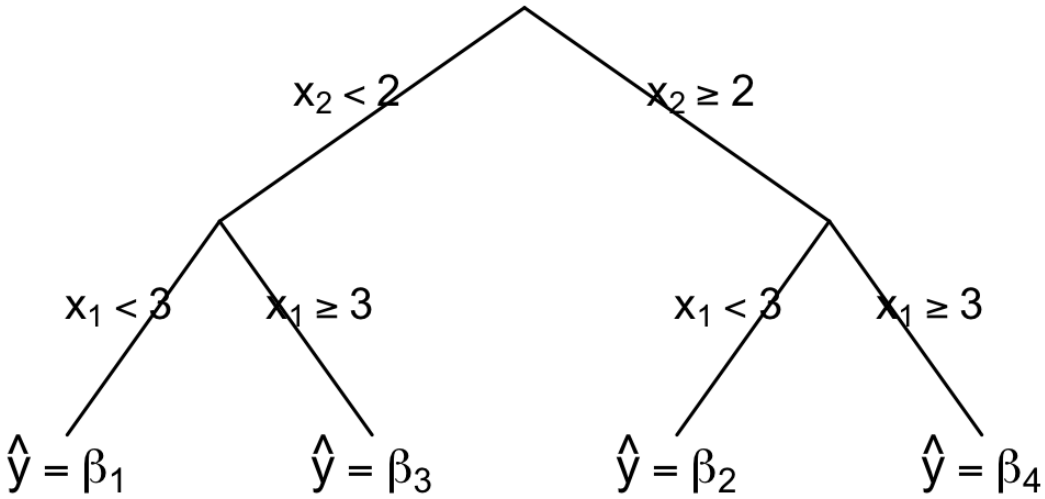


Figure: Tree representation of the same regression model.

A Simple Regression Tree: Tree Representation

Figure 2 shows the same model drawn as a tree, with:

- internal nodes corresponding to splits on covariates,
- leaves corresponding to terminal regions with constant predictions.

From Bagging to Random Forests

- Bagging (bootstrap aggregation):

$$\widehat{m}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{m}_b(x),$$

  where each $\widehat{m}_b$ is a tree fit on a bootstrap sample.
- Reduces variance by averaging many noisy trees.
- Random forests add extra randomness:
  - At each split: choose a random subset of regressors of size $m < p$,
  - Only search over those m variables.
- Random forest predictor:

$$\widehat{m}_{\text{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{m}_b(x),$$

  but with trees decorrelated via random feature selection.

Practical Features of Random Forests

- Typically strong out-of-the-box performance:
  - Few hyperparameters,
  - Robust to overfitting (to some extent).
- Out-of-bag (OOB) error:
  - For each tree, about $1/3$ of observations not used in bootstrap,
  - Use these as validation to estimate prediction error,
  - No separate test set needed for rough performance assessment.
- Automatic modeling of:
  - Nonlinearities,
  - Interactions,
  - Heteroskedasticity.
- Interpretation aids:
  - Variable importance,
  - Partial dependence plots.

Variable Importance and Partial Dependence

- **Permutation variable importance**:
  - Permute values of regressor j in OOB data,
  - Measure increase in prediction error,
  - Larger increase $\Rightarrow$ more important variable.
- **Partial dependence plots**:
  - For regressor X_j, plot:

$$\widehat{f}_j(z) = \frac{1}{T}\sum_{t=1}^{T} \widehat{m}_{\mathrm{RF}}(z, X_{t,-j}),$$

  - Shows average predicted response as X_j varies,
  - Must be interpreted carefully if regressors are correlated.
- Gives some insight into forest structure without full interpretability.

Table of Contents

Data for the Case Study

- Asset: SPY (S&P 500 ETF)
  - Highly liquid equity index ETF,
  - Daily adjusted closing prices from 1 January 2020 to 31 December 2024.
- Return series:
  - Let P_t be the adjusted close on trading day t,
  - Define daily log return (in percent):

$$r_t = 100\big(\log P_t - \log P_{t-1}\big).$$

Volatility Forecasting Case Study: Setup

- Volatility proxy (forecast target):
  - Squared return:
$$y_t = r_t^2,$$
  - Interpreted as a simple proxy for next-day conditional variance,
  - Links directly to volatility models in Chapter 4 Volatility Models.
- Forecast design:
  - Split into training and evaluation periods,
  - Rolling or expanding windows to ensure out-of-sample predictions,
  - One-step-ahead forecasts $\widehat{y}_{t+1|t}$.
- Evaluation:
  - RMSE, MAE,
  - Possibly quasi-likelihood loss (not pursued here).
- Note: any other liquid asset (broad index, FX rate, commodity future) could be used; data are obtained from Yahoo Finance via `quantmod`.

Classical Benchmark Models

- **AR model on y_t**:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t.$$

  - Often AR(1) or AR(p) chosen by information criteria,
  - Forecast:

$$\widehat{y}^{\mathrm{AR}}_{t+1|t} = \widehat{\phi}_0 + \sum_{j=1}^{p} \widehat{\phi}_j y_{t+1-j}.$$

- **GARCH(1,1)**:
  - Return model: $r_t = \sigma_t \varepsilon_t$,
  - Variance recursion:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2.$$

  - Forecast:

$$\widehat{\sigma}^2_{t+1|t} = \widehat{\omega} + \widehat{\alpha} r_t^2 + \widehat{\beta} \widehat{\sigma}_t^2.$$

ML Models in the Case Study

- **Feature construction**:
  - Lags of y_t (squared returns): $y_t, \ldots, y_{t-L}$,
  - Lags of $|r_t|$,
  - Rolling averages of volatility measures,
  - Calendar dummies (day-of-week, etc.).
- **Regularized linear model**:

$$\widehat{y}_{t+1|t}^{\text{Lasso/Ridge}} = X_t' \widehat{\beta}_\lambda,$$

  with λ chosen via time-series cross-validation.
- **Random forest**:

$$\widehat{y}_{t+1|t}^{\text{RF}} = \widehat{m}_{\text{RF}}(X_t),$$

  using the same feature set.

Case Study: RMSE Comparison

- Rolling forecast results for SPY (example):

$$\begin{aligned} \text{AR(1) on } v_t &: \text{RMSE} \approx 1.33, \\ \text{GARCH(1,1)} &: \text{RMSE} \approx 1.20, \\ \text{Lasso} &: \text{RMSE} \approx 1.28, \\ \text{Random forest} &: \text{RMSE} \approx 0.66. \end{aligned}$$

- Interpretation:
  - GARCH improves on AR(1) by modeling conditional heteroskedasticity,
  - Lasso on richer features similar to AR(1) in this design,
  - Random forest substantially reduces RMSE, capturing nonlinearities and interactions.
- Caveat: results are sample- and design-specific; other data sets may yield different rankings.

Realised Volatility vs Lasso Forecast

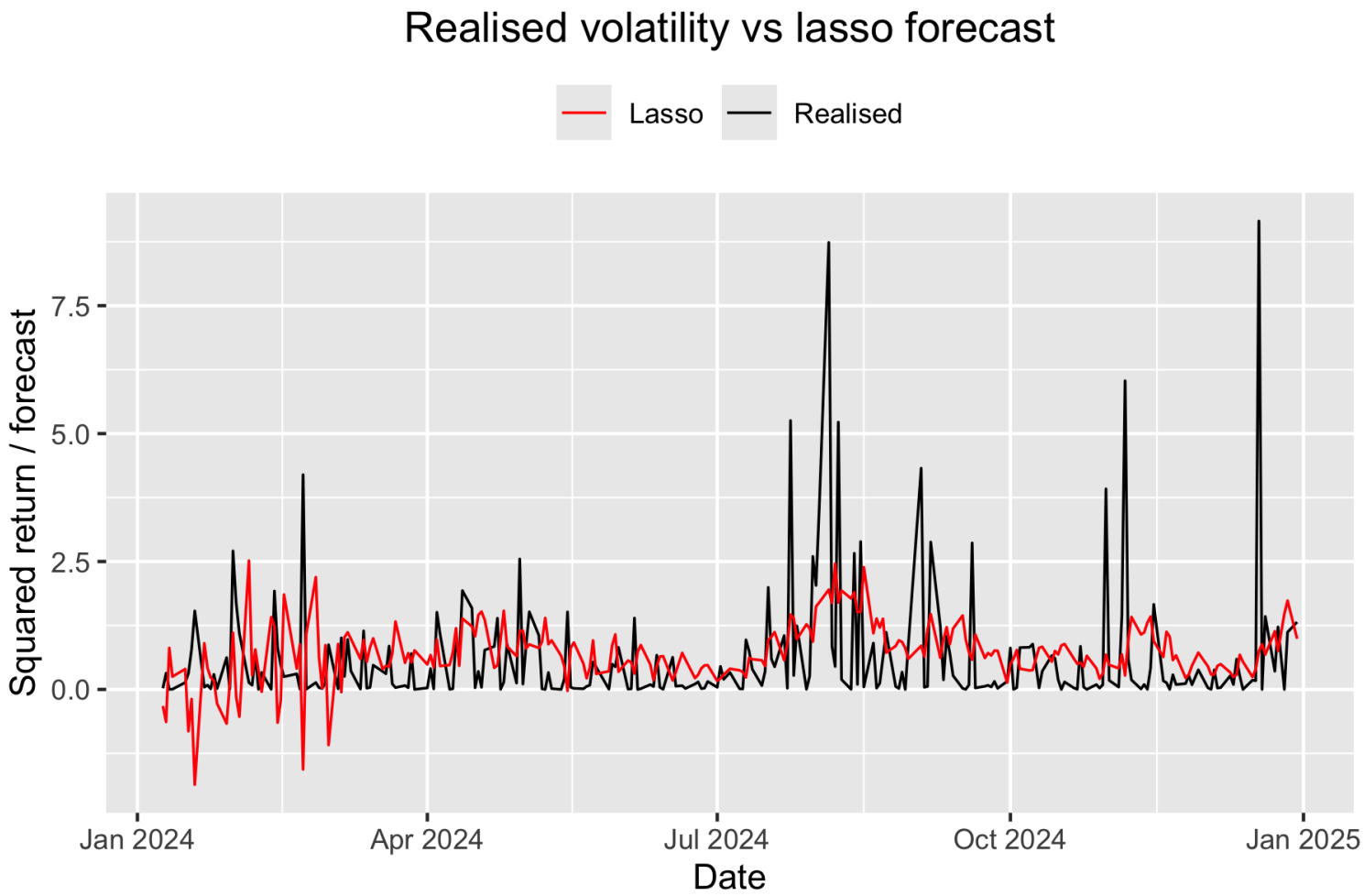


Figure: Realised volatility (squared returns) and one–step–ahead lasso forecasts for SPY daily data.

Realised vs Lasso Forecast: Qualitative Behavior

- Realised volatility (squared returns) vs lasso forecast:
  - Lasso forecast is smoother,
  - Tracks broad level of volatility reasonably well,
  - Underestimates sudden volatility spikes.
- Typical for many volatility models:
  - Good at gradual volatility changes,
  - Less able to anticipate extreme moves.

Takeaways from the Case Study

- Classical models:
  - AR(1) and GARCH(1,1) are simple but not useless,
  - GARCH can substantially improve over simple AR for volatility.
- Regularized linear models:
  - Lasso on richer features may or may not outperform classical models,
  - Gains depend on genuine predictive signal in the predictors.
- Random forest:
  - In this example, delivers the lowest RMSE,
  - Shows clear potential of flexible nonlinear ML tools.
- Interpretation vs prediction:
  - AR/GARCH parameters have clear economic meaning,
  - Lasso/ridge/forests are "black-box" forecasters with limited structural interpretability.

Classical vs ML: Complementary Roles

- Classical models:
  - Transparent and interpretable,
  - Grounded in time-series theory,
  - Rich diagnostic toolkit (residuals, ARCH tests, etc.).
- ML models:
  - Handle high-dimensional predictors,
  - Capture nonlinearities interactions,
  - Often better pure predictors.
- Good practice:
  - Use both families of models,
  - Compare out-of-sample performance,
  - Let economic questions determine required interpretability.

Table of Contents

Chapter Summary (1)

- Reframed classical econometrics in statistical learning language:
  - Prediction functions,
  - Bias–variance trade-off,
  - Training/validation/test split, cross-validation.
- Regularized linear models:
  - Ridge: stabilises multicollinearity,
  - Lasso: shrinkage + variable selection,
  - SCAD: oracle-like variable selection with reduced bias.
- Implementation in R:
  - `glmnet` for ridge/lasso,
  - `ncvreg` for SCAD/lasso.

Chapter Summary (2)

- Tree-based methods:
  - Regression trees (CART) as adaptive step-function estimators,
  - Random forests as ensembles of trees that:
    - Reduce variance via bagging,
    - Decorrelate trees via random feature selection.
- Interpretation aids:
  - Variable importance,
  - Partial dependence plots.
- Volatility-forecasting case study:
  - Classical AR/GARCH vs lasso vs random forest,
  - RF can substantially improve RMSE for this dataset,
  - Classical models remain competitive and more interpretable.

Further Topics Not Covered

- Deep learning:
  - RNNs, LSTMs, transformers,
  - Growing role in high-frequency and text-based finance.
- Causal machine learning:
  - Double/debiased ML, causal forests, etc.,
  - Semiparametric efficiency and high-dimensional inference.
- Boosting and XGBoost:
  - Powerful ensemble methods beyond bagging and forests.
- Asymptotic theory for high-dimensional estimators and ensembles.

Recommended Reading

- Statistical learning:
  - James et al. (2021) (ISLR2): accessible, R-based,
  - Hastie, Tibshirani, and Friedman (2009): advanced, theoretical.
- Econometrics and ML:
  - Varian (2014), Mullainathan and Spiess (2017),
  - Prado (2020) for asset management.
- Forecasting:
  - Hyndman and Athanasopoulos (2018): R examples, rolling-window evaluation.
- Software:
  - Friedman, Hastie, and Tibshirani (2010), Wright and Ziegler (2017),
  - Kuhn and Silge (2022).

Table of Contents

References I

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.
- Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.
- James, Gareth et al. (2021). *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. New York: Springer.
- Kuhn, Max and Julia Silge (2022). *Tidy Modeling with R*. Sebastopol, CA: O'Reilly Media.
- Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

References II


Prado, Marcos López de (2020). *Machine Learning for Asset Managers*. Cambridge: Cambridge University Press.


Varian, Hal R. (2014). "Big Data: New Tricks for Econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.


Wright, Marvin N. and Andreas Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17.