Lecture 1.pdf

Lecture 2.pdf

Lecture 3.1.pdf

Lecture 3.2.pdf

Lecture 3.3.pdf

Lecture 3.4.pdf

Lecture 3.5.pdf

Lecture 4.1.pdf

Lecture 4.2.pdf

Lecture 4.3.pdf

Lecture 5.1.pdf

Lecture 5.pdf

Lecture 6.1.pdf

Lecture 6.2.pdf

Lecture 7.1.pdf

Lecture 7.2.pdf

Lecture 7.3.pdf

Lecture 8.1.pdf

Lecture 8.2.pdf

Lecture 8.3.pdf

Lecture 8.4.pdf

Finance Preliminaries

Introduction

Our objective is to learn the theory and application of time series methods.

- We will focus on financial time series applications.
- The methods in this course are broadly applicable to *any* type of time series.
- We will use the R programming environment to work with financial time series data.
- The quantmod library will be especially useful:

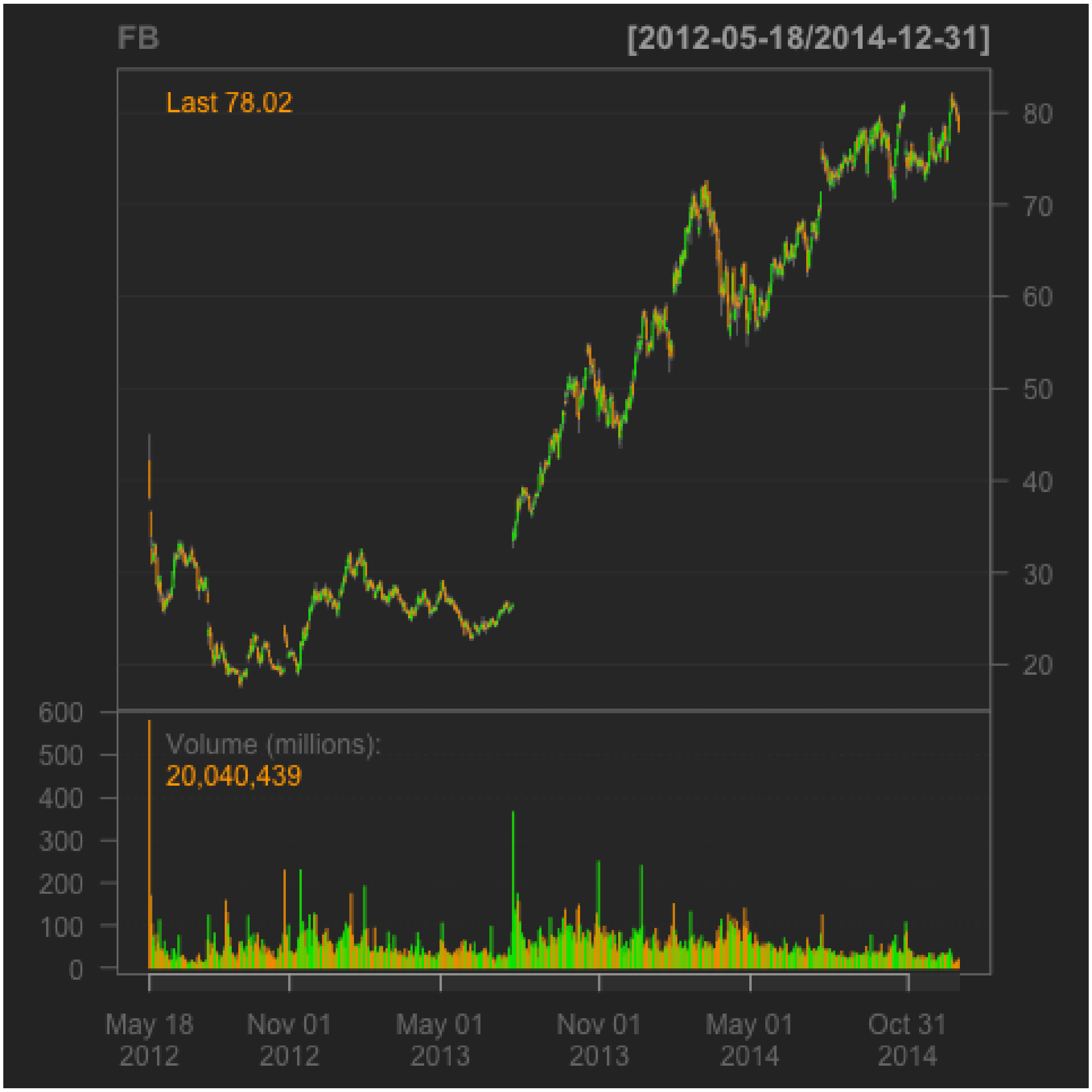
```
> install.packages("quantmod")
```

Time Series Example

Let’s plot the historical prices for Facebook (FB).

```
> library(quantmod)
> getSymbols("FB",src="google", from="2012-01-01", to="2014-12-31")
> png(filename="fb.png")
> chartSeries(FB)
> dev.off()
```

Plot of Facebook Price



Facebook Returns

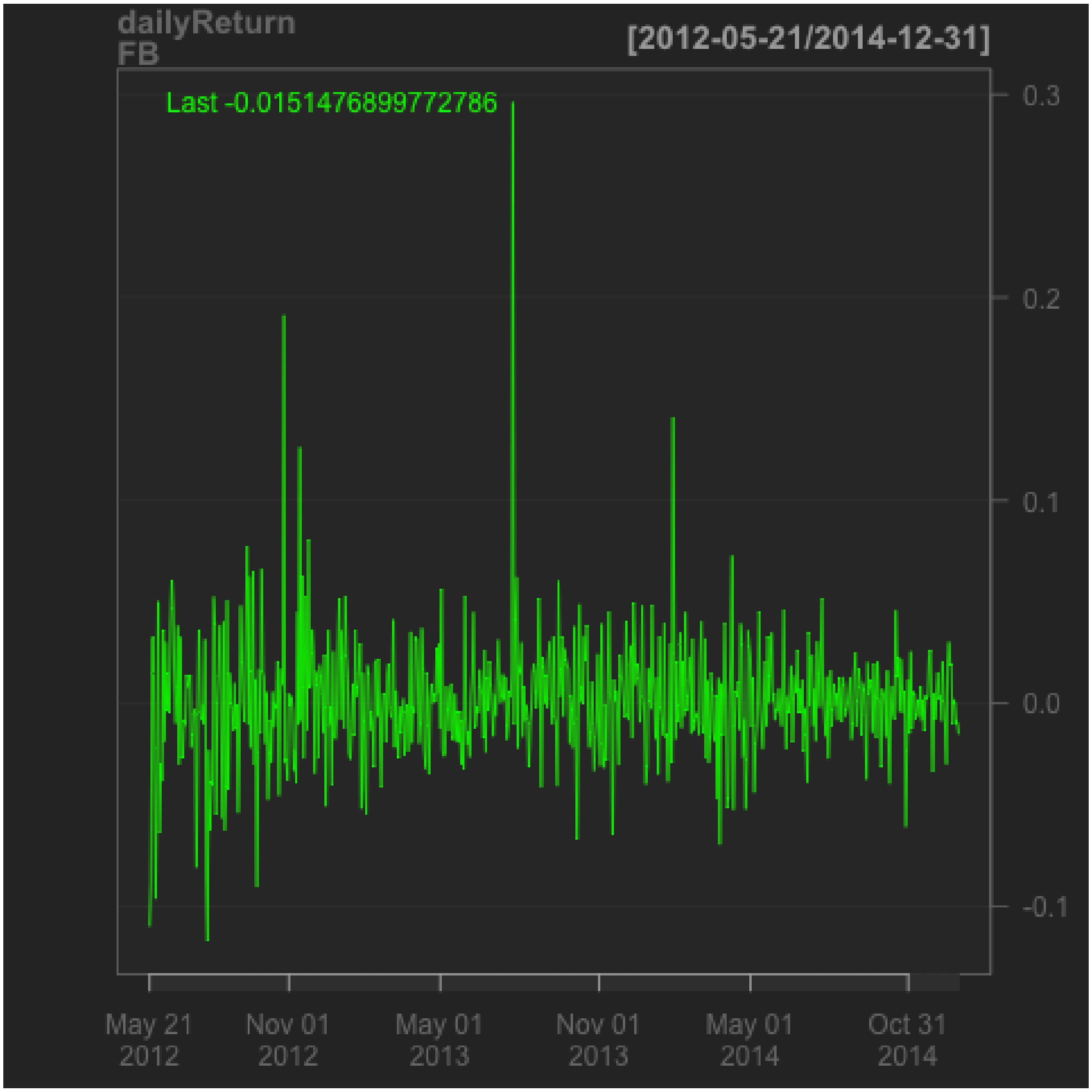
To plot just the closing prices:

```
> chartSeries(Cl(FB))
> chartSeries(FB$FB.Close)
```

Or daily returns:

```
> chartSeries(dailyReturn(Cl(FB)))
```

Plot of Facebook Returns



One-Period Return

Let P_t be the price of an asset at time t.

- The *gross return* of the asset between dates $t-1$ and t is:

$$R_t = \frac{P_t}{P_{t-1}} \quad \text{or} \quad P_t = P_{t-1} R_t.$$

- The *net return* is:

$$r_t = R_t - 1 = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

- Note that the return can be computed between any two dates (i.e. daily, weekly, monthly, etc).

Multi-Period Return

The k-period gross return between dates $t-k$ and t is:

$$
\begin{aligned}
R_t(k) = \frac{P_t}{P_{t-k}} &= \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \cdots \times \frac{P_{t-k+1}}{P_{t-k}} \\
&= R_t R_{t-1} \cdots R_{t-k+1} \\
&= \prod_{j=0}^{k-1} R_{t-j}.
\end{aligned}
$$

- The k -period net return is:

$$
r_t(k) = \frac{P_t - P_{t-k}}{P_{t-k}}.
$$

Logarithmic Approximation

In general, for any small value $\varepsilon > 0$:

$$
\ln(1 + \varepsilon) \approx \varepsilon.
$$

Thus,

$$
\ln(R_t) = \ln(1 + r_t) \approx r_t.
$$

Furthermore, by the definition of gross returns,

$$
r_t \approx \ln(R_t) = \ln(P_t / P_{t-1}) = \ln(P_t) - \ln(P_{t-1}).
$$

Approximation for Multiperiod Returns

A similar relationship holds for the k -period net return:

$$
r_t(k) \approx \ln(P_t) - \ln(P_{t-k}).
$$

Time Intervals

The interval of time for returns is of vital importance for understanding the data.

- Daily returns are very different from weekly, monthly, annual, etc. returns.

- Intra-day returns at various time scales (millisecond, second, minute) are very different from each other.

Aggregating Trading Intervals

When aggregating returns, we consider the following.

- There are approximately 250 trading days in a year.
- There are approximately 22 trading days in a month.
- There are 5 trading days in a week.
- U.S. equities markets are open from 9:30 am to 4:00 pm Eastern time - 6.5 hours each day.
- Thus there are approximately 6.5 hours, or 390 minutes or 23,400 seconds or 23,400,000 milliseconds in a trading day.
- Similarly, there are approximately 1625 trading hours, 97,500 trading minutes, 5,850,000 trading seconds and 5,850,000,000 trading milliseconds in a year.

Aggregating Returns

To aggregate net returns, we simply add them:

$$\begin{aligned} r_t(k) &= \ln(P_t) - \ln(P_{t-k}) \\ &= \ln(P_t) - \ln(P_{t-1}) + \ln(P_{t-1}) - \ln(P_{t-2}) + \ln(P_{t-2}) \\ &\qquad - \dots - \ln(P_{t-k+1}) + \ln(P_{t-k+1}) - \ln(P_{t-k}) \\ &= r_t + r_{t-1} + r_{t-2} + \dots + r_{t-k+1} \\ &= \sum_{j=0}^{k-1} r_{t-j}. \end{aligned}$$

For example, to annualize daily returns,

$$r_t(250) = \sum_{j=0}^{250} r_j.$$

Example of Aggregating Returns

Get Exxon Mobile equities data for the week of March 23rd, 2015.

```
> getSymbols("XOM", from="2015-03-23", to="2015-03-27")
[1] "XOM"
> XOM
           XOM.Open XOM.High XOM.Low XOM.Close XOM.Volume XOM.Adjusted
2015-03-23    85.02    85.78   85.01     85.43   17163200        85.43
2015-03-24    85.30    85.78   84.50     84.52   10099500        84.52
2015-03-25    85.05    85.57   84.77     84.86   11816000        84.86
2015-03-26    85.30    85.57   84.09     84.32   14388500        84.32
2015-03-27    84.04    84.05   83.33     83.58   11094600        83.58
```

- What are the daily returns?

- What is the weekly return?

Asset Classes

There are several broad classes of assets traded in financial markets.

- Equities.

- Futures.

- Options.

- Bonds.

- Currencies.

Indices

Indices are synthetic portfolios of assets that are not typically traded.

- The S&P 500 index is a portfolio of 500 equities *and is not traded*.

- To hold the S&P 500 index, one can:
  - Purchase the 500 component equities in the correct proportions.
  - Purchase shares in a mutual fund that tracks the index.
  - Purchase shares of the SPY exchange traded fund (ETF).
  - Purchase futures contracts on SPX.

Important Indices

- S&P 500 (SPX).

- VIX - portfolio of S&P 500 options which represents the expected value of a one-standard deviation move in the S&P 500 index over the next month (in annual terms).

- On March 30th, 2015, the closing value for VIX was 14.51 and the closing value for SPX 2086.24.

- Hence, the market expects the standard deviation of the SPX to be $14.51/\sqrt{12} = 4.19$ percent or $0.0419 \times 2086.24 = 87.39$ index points.

Important Assets

- SPY - SPX ETF.

- E-mini - Futures contract on the SPX.

- SPX Options.

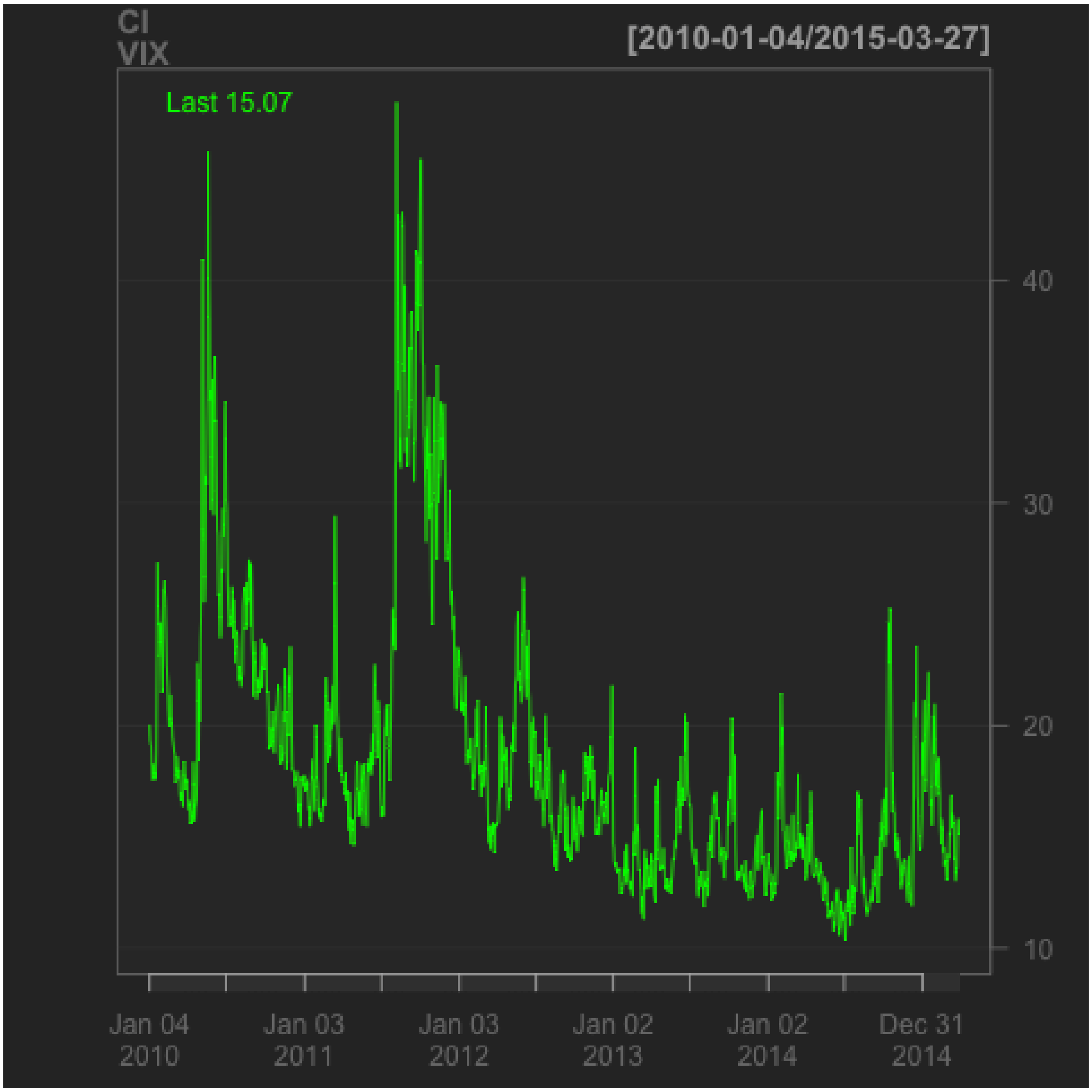
- SPY Options.

- VIX Options.

- VIX Futures.

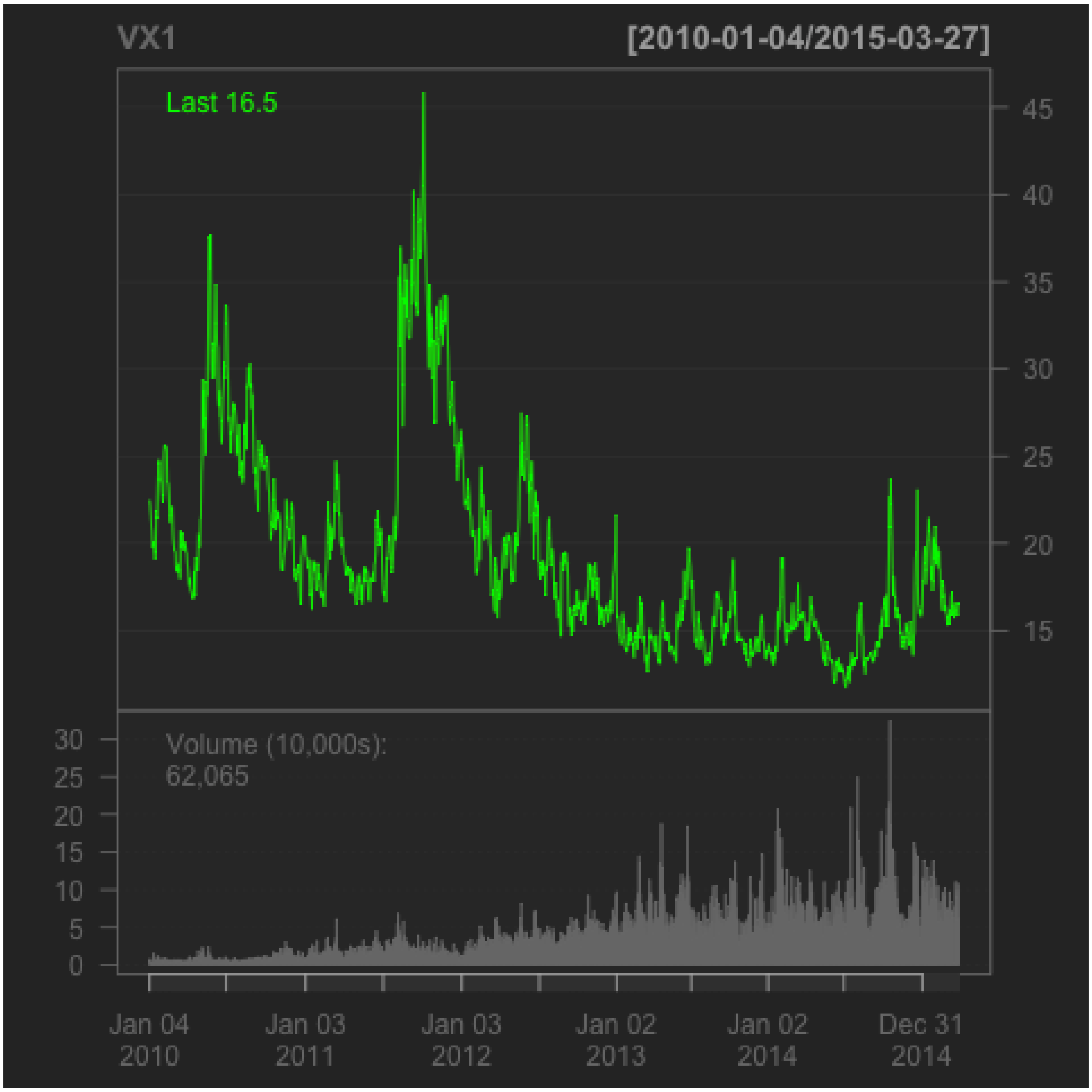
VIX

```
> getSymbols("^VIX", from="2014-01-01", to="2015-03-27")
> chartSeries(Cl(VIX))
```



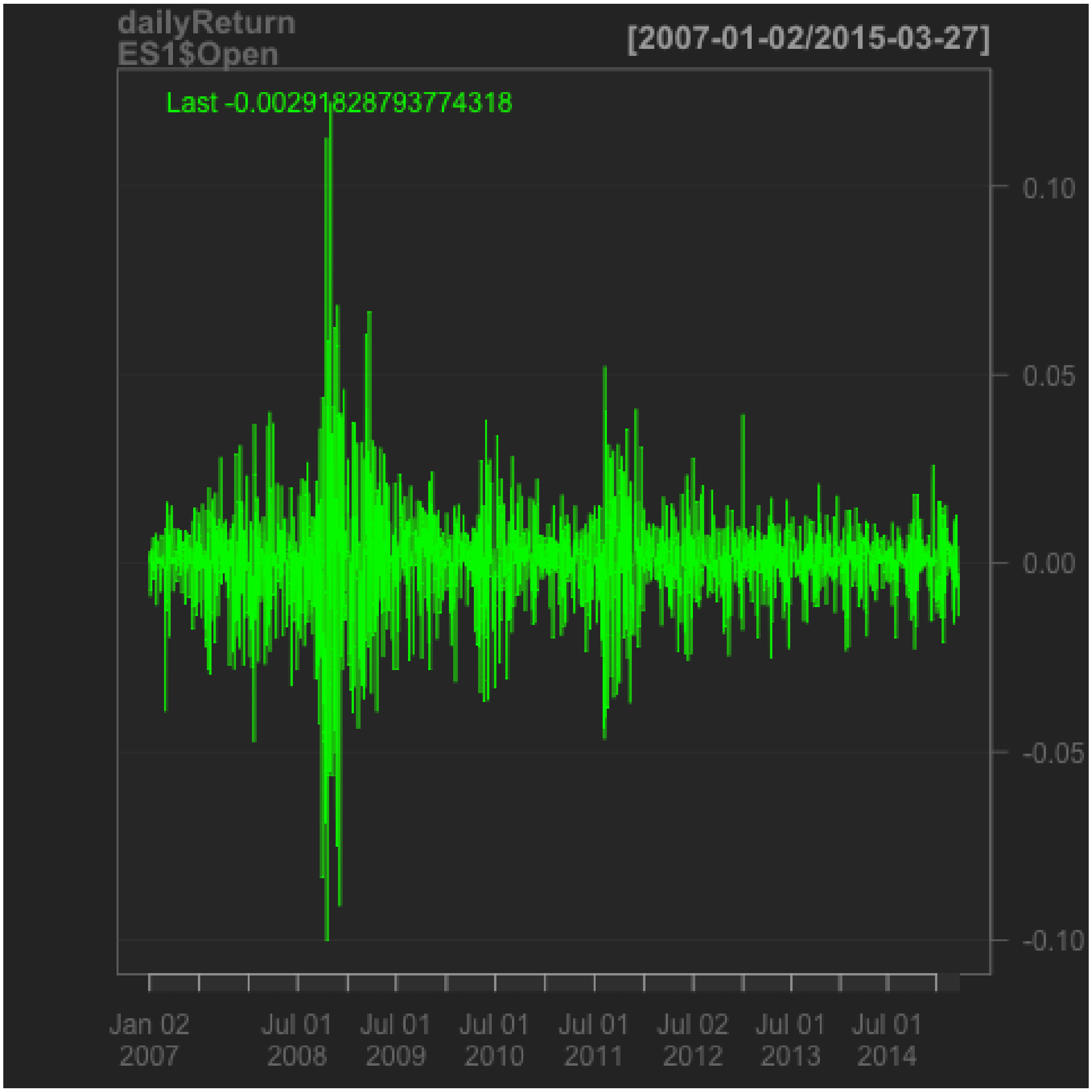
Near-Month VX Futures

```
> install.packages("Quandl")
> library(Quandl)
> VX1 = Quandl("OFDP/FUTURE_VX1",type="xts")
> chartSeries(VX1)
```



E-mini Near-Month Returns

```
> ES1 = Quandl("OFDP/FUTURE_ES1",start_date="2007-01-01",end_date="2015-03-27",type="xts")
> chartSeries(dailyReturn(ES1$Open))
```



Important Features of Returns

What do you notice about the E-mini returns?

Stationarity

Introduction

Time series analysis is concerned with dynamics.

- We may have complete knowledge of the *unconditional* distribution of a group of random variables but no understanding of their sequential dynamics.

- Time series is focused on understanding the sequential relationship of a group of random variables.

- Hence, the focus is *conditional* distributions and *autocovariances*.

Time Series

A time series is a stochastic process indexed by time:

$$Y_1, Y_2, Y_3, \ldots, Y_{T-1}, Y_T.$$

- *Stochastic* is a synonym for *random*.

- So a time series is a sequence of (potentially different) random variables ordered by time.

- We will let lower-case letters denote a realization of a time series.

$$y_1, y_2, y_3, \ldots, y_{T-1}, y_T.$$

Distributions

We will think of $\mathbf{Y}_T = \{Y_t\}_{t=1}^T$ as a random variable in its own right.

- $\mathbf{y}_T = \{y_t\}_{t=1}^T$ is a *single* realization of $\mathbf{Y}_T = \{Y_t\}_{t=1}^T$.

- The CDF is $F_{\mathbf{Y}_T}(\mathbf{y}_T)$ and the PDF is $f_{\mathbf{Y}_T}(\mathbf{y}_T)$.

- For example, consider $T = 100$:

$$F\left(\mathbf{y}_{100}\right) = P(Y_1 \leq y_1, \ldots, Y_{100} \leq y_{100}).$$

- Notice that $\mathbf{Y}_T$ is just a collection of random variables and $f_{\mathbf{Y}_T}(\mathbf{y}_T)$ is the joint density.

Time Series Observations

As statisticians and econometricians, we want many observations of $\mathbf{Y}_T$ to learn about its distribution:

$$\mathbf{y}_T^{(1)}, \quad \mathbf{y}_T^{(2)}, \quad \mathbf{y}_T^{(3)}, \quad \ldots$$

Likewise, if we are only interested in the marginal distribution of Y_{17}

$$F_{Y_{17}}(a) = P(Y_{17} \leq a)$$

we want many observations: $\left\{y_{17}^{(i)}\right\}_{i=1}^{N}$.

Time Series Observations

Unfortunately, we usually only have *one observation* of $\mathbf{Y}_T$.

- Think of the daily closing price of Harley-Davidson stock since January 2nd.

- Think of your cardiogram for the past 100 seconds.

In neither case can you repeat history to observe a new sequence of prices or electric heart signals.

- In time series econometrics we typically base inference on a single observation.

- Additional assumptions about the process will allow us to exploit information in the full sequence $\mathbf{y}_T$ to make inferences about the joint distribution $F_{\mathbf{Y}_T}(\mathbf{y}_T)$.

Moments

Since the stochastic process is comprised of individual random variables, we can consider moments of each:

$$E[Y_t] = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t) dy_t = \mu_t$$

$$Var(Y_t) = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t) dy_t = \gamma_{0t}$$

Moments

$$Cov(Y_t, Y_{t-j}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y_t - \mu_t)(y_{t-j} - \mu_{t-j}) \\ \times f_{Y_t, Y_{t-j}}(y_t, y_{t-j})dy_t dy_{t-j} = \gamma_{jt},$$

where f_{Y_t} and $f_{Y_t, Y_{t-j}}$ are the marginal distributions of $f_{\mathbf{Y}_T}$ obtained by integrating over the appropriate elements of $\mathbf{Y}_T$.

Autocovariance and Autocorrelation

- γ_{jt} is known as the j th autocovariance of Y_t since it is the covariance of Y_t with its own lagged value.

- The j th autocorrelation of Y_t is defined as

$$\begin{aligned}\rho_{jt} &= Corr(Y_t, Y_{t-j}) \\ &= \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-j})}} \\ &= \frac{\gamma_{jt}}{\sqrt{\gamma_{0t}}\sqrt{\gamma_{0t-j}}}.\end{aligned}$$

Sample Moments

If we had N observations $\mathbf{y}_T^{(1)}, \ldots, \mathbf{y}_T^{(N)}$, we could estimate moments of each (univariate) Y_t in the usual way:

$$\begin{aligned}\hat{\mu}_t &= \frac{1}{N}\sum_{i=1}^{N} y_t^{(i)}. \\ \hat{\gamma}_{0t} &= \frac{1}{N}\sum_{i=1}^{N}(y_t^{(i)} - \hat{\mu}_t)^2. \\ \hat{\gamma}_{jt} &= \frac{1}{N}\sum_{i=1}^{N}(y_t^{(i)} - \hat{\mu}_t)(y_{t-j}^{(i)} - \hat{\mu}_{t-j}).\end{aligned}$$

Example

Suppose each element of $\mathbf{Y}_T$ is described by

$$Y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_t^2), \forall t.$$

Example

In this case,

$$\begin{aligned} \mu_t &= E[Y_t] = \mu_t, \quad \forall t, \\ \gamma_{0t} &= Var(Y_t) = Var(\varepsilon_t) = \sigma_t^2, \quad \forall t \\ \gamma_{jt} &= Cov(Y_t, Y_{t-j}) = Cov(\varepsilon_t, \varepsilon_{t-j}) = 0, \quad \forall t, j \neq 0. \end{aligned}$$

- If $\sigma_t^2 = \sigma^2 \; \forall t, \varepsilon_T$ is known as a *Gaussian white noise* process.
- In this case, $\mathbf{Y}_T$ is a Gaussian white noise process with drift.
- μ_T is the drift vector.

White Noise

Generally speaking, ε_T is a *white noise process* if

$$\begin{aligned} E[\varepsilon_t] &= 0, \quad \forall t \\ E[\varepsilon_t^2] &= \sigma^2, \quad \forall t \\ E[\varepsilon_t \varepsilon_\tau] &= 0, \quad \text{for } t \neq \tau. \end{aligned}$$

White Noise

Notice there is no distributional assumption for ε_t.

- If ε_t and ε_τ are independent for $t \neq \tau, \varepsilon_T$ is *independent white noise.*
- Notice that independence $\Rightarrow E[\varepsilon_t \varepsilon_\tau] = 0$, but $E[\varepsilon_t \varepsilon_\tau] = 0 \not\Rightarrow$ independence.
- If $\varepsilon_t \sim \mathcal{N}(0, \sigma^2) \; \forall t$, as in the example above, ε_T is Gaussian white noise.

Weak Stationarity

Suppose the first and second moments of a stochastic process $\mathbf{Y}_T$ don't depend on $t \in T$:

$$E[Y_t] = \mu \quad \forall t$$
$$Cov(Y_t, Y_{t-j}) = \gamma_j \quad \forall t \text{ and any } j.$$

- In this case $\mathbf{Y}_T$ is *weakly stationary* or *covariance stationary*.

- In the previous example, if $Y_t = \mu + \varepsilon_t \, \forall t$, $\mathbf{Y}_T$ is weakly stationary.

- However if $\mu_t \neq \mu \, \forall t$, $\mathbf{Y}_T$ is *not* weakly stationary.

Autocorrelation under Weak Stationarity

If $\mathbf{Y}_T$ is weakly stationary

$$\begin{aligned}\rho_{jt} &= \frac{\gamma_{jt}}{\sqrt{\gamma_{0t}}\sqrt{\gamma_{0t-j}}} \\ &= \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} \\ &= \frac{\gamma_j}{\gamma_0} \\ &= \rho_j.\end{aligned}$$

- Note that $\rho_0 = 1$.

Weak Stationarity

Under weak stationarity, autocovariances γ_j only depend on the distance between random variables within a stochastic process:

$$Cov(Y_\tau, Y_{\tau-j}) = Cov(Y_t, Y_{t-j}) = \gamma_j.$$

This implies

$$\gamma_{-j} = Cov(Y_{t+j}, Y_t) = Cov(Y_t, Y_{t-j}) = \gamma_j.$$

Weak Stationarity

More generally,

$$\Sigma_{\mathbf{Y}_T} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}.$$

Strict Stationarity

$\mathbf{Y}_T$ is *strictly stationary* if for any set $\{j_1, j_2, \ldots, j_n\} \in T$

$$f_{Y_{j_1}, \ldots, Y_{j_n}}(a_1, \ldots, a_n) = f_{Y_{j_1+\tau}, \ldots, Y_{j_n+\tau}}(a_1, \ldots, a_n), \quad \forall \tau.$$

- Strict stationarity means that the joint distribution of any subset of random variables in $\mathbf{Y}_T$ is invariant to shifts in time, τ.

- Strict stationarity $\Rightarrow$ weak stationarity if the first and second moments of a stochastic process exist.

- Weak stationarity $\nRightarrow$ strict stationarity: invariance of first and second moments to time shifts (weak stationarity) does not mean that all higher moments are invariant to time shifts (strict stationarity).

Strict Stationarity

If $\mathbf{Y}_T$ is Gaussian then weak stationarity $\Rightarrow$ strict stationarity.

- If $\mathbf{Y}_T$ is Gaussian, all marginal distributions of $(Y_{j_1}, \ldots, Y_{j_n})$ are also Gaussian.

- Gaussian distributions are fully characterized by their first and second moments.

Ergodicity

Given N identically distributed weakly stationary stochastic processes $\left\{\mathbf{Y}_T\right\}_{i=1}^{N}$, the *ensemble average* is

$$\frac{1}{N}\sum_{i=1}^{N} Y_t^{(i)} \xrightarrow{p} \mu, \quad \forall t.$$

For a single stochastic process, we desire conditions under which the *time average*

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu.$$

Ergodicity

If $\mathbf{Y}_T$ is weakly stationary and

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty,$$

Then $\mathbf{Y}_T$ is *ergodic for the mean* and the time average converges.

- The equation above requires that the autocovariances fall to zero sufficiently quickly.
- i.e. a *long realization* of $\{y_t\}$ will have many segments that are uncorrelated and which can be used to approximate an ensemble average.

Ergodicity

A weakly stationary process is ergodic for the second moments if

$$\frac{1}{T-j}\sum_{t=j+1}^{T}(Y_t-\mu)(Y_{t-j}-\mu) \xrightarrow{p} \gamma_j.$$

- Separate conditions exist which cause the equation above to hold.
- If $\mathbf{Y}_T$ is Gaussian and stationary, then $\sum_{j=0}^{\infty}|\gamma_j| < \infty$ ensures that $\mathbf{Y}_T$ is ergodic for all moments.

Lag Operators

Lag Operator

Given a sequence of values, $y_1, y_2, \ldots, y_t$, indexed by time, the lag operator, L, is defined as

$$Ly_t = y_{t-1}.$$

- The lag operator shifts a time value y_t back by one period.
- y_t can be thought of as the input of the operator and y_{t-1} as the output.
- The lag operator can be applied to all values in a series $\{y_t\}_{t=1}^{T}$ and the result is a new series shifted back by one period: $\{y_t\}_{t=0}^{T-1}$.

Lag Operator

Applying the lag operator twice:

$$L(Ly_t) = Ly_{t-1} = y_{t-2}.$$

- We write $L(Ly_t)$ as $L^2 y_t$.
- Applying recursively:

$$L^k y_t = y_{t-k}.$$

- We will define $L^0 = 1$.

Useful Properties of the Lag Operator

- The lag operator is commutative:

$$L(\beta y_t) = \beta L y_t.$$

- The lag operator is distributive:

$$L(x_t + y_t) = Lz_t = z_{t-1} = x_{t-1} + y_{t-1} = Lx_t + Ly_t,$$

where $z_t = x_t + y_t$.

- The lag of a constant is the same constant:

$$Lc = c.$$

First-Order Difference Equation

Suppose we have a first-order difference equation:

$$y_t = \phi y_{t-1} + w_t \, .$$

In terms of the lag operator

$$(1 - \phi L) y_t = w_t \, .$$

We can write

$$\phi(L) y_t = w_t \, ,$$

where $\phi(L) = (1 - \phi L)$.

First-Order Difference Equation

Suppose the operator $\phi(L) = (1 - \phi L)$ has an inverse:

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

- The inverse is the operator such that

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

- If an inverse operator exists,

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

Recursive Substitution of First-Order Difference Equation

Applying recursive substitution to the first-order difference equation:

$$
\begin{aligned}
y_t &= \phi y_{t-1} + w_t \\
&= \phi(\phi y_{t-2} + w_{t-1}) + w_t \\
&= w_t + \phi w_{t-1} + \phi^2 y_{t-2} \\
&= w_t + \phi w_{t-1} + \phi^2 (\phi y_{t-3} + w_{t-2}) \\
&= w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 y_{t-3} \\
\vdots \\
&= \sum_{i=0}^{\infty} \phi^i w_{t-i} = \sum_{i=0}^{\infty} \phi^i L^i w_t .
\end{aligned}
$$

- The infinite recursive substitution can only be performed if $|\phi| < 1$.

Inverse of Lag Operator

Restating the previous result, for $|\phi| < 1$:

$$y_t = \left(\sum_{i=0}^{\infty} \phi^i L^i \right) w_t .$$

Substituting:

$$w_t = (1 - \phi L) y_t = (1 - \phi L) \left(\sum_{i=0}^{\infty} \phi^i L^i \right) w_t .$$

So when $|\phi| < 1$:

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

That is, $\sum_{i=0}^{\infty} \phi^i L^i$ is the inverse operator of $(1 - \phi L)$.

p th-Order Difference Equation

Suppose we have a p th-order difference equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + w_t .$$

In terms of the lag operator

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t = w_t .$$

We can write

$$\phi(L) y_t = w_t ,$$

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

Factoring Polynomials

In general, a p th-order, real-valued polynomial can be factored as

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = (1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_p z).$$

- $\left\{ \frac{1}{\lambda_i} \right\}_{i=1}^{p}$ are the p roots of the polynomial.

This is the value that makes the function zero because later on we will define z as the reciprocal, were then (1-1) equals zero

- Some of the roots may be complex and some may be identical.

Factoring p th-Order Difference Equation

If we factor the p th-order lag polynomial in the same way as a real-valued polynomial:

$$\begin{aligned}(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t \\ = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L) y_t = w_t.\end{aligned}$$

If $|\lambda_i| < 1$,

$$(1 - \lambda_i L)^{-1} = \sum_{j=0}^{\infty} \lambda_i^j L^j, \quad \forall i.$$

Factoring p th-Order Difference Equation

In this case,

$$\begin{aligned} y_t &= (1 - \lambda_1 L)^{-1} (1 - \lambda_2 L)^{-1} \cdots (1 - \lambda_p L)^{-1} w_t \\ &= \left(\sum_{j=0}^{\infty} \lambda_1^j L^j \right) \left(\sum_{j=0}^{\infty} \lambda_2^j L^j \right) \cdots \left(\sum_{j=0}^{\infty} \lambda_p^j L^j \right) w_t. \end{aligned}$$

Factoring p th-Order Difference Equation

If we define

$$\theta(L) = \left(\sum_{j=0}^{\infty} \lambda_1^j L^j \right) \left(\sum_{j=0}^{\infty} \lambda_2^j L^j \right) \cdots \left(\sum_{j=0}^{\infty} \lambda_p^j L^j \right)$$

then

$$y_t = \theta(L) w_t.$$

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

- Note that the inverse only exists when $|\lambda_i| < 1, \forall i$.

- This can also be stated as: the inverse only exists when the roots of $\phi(L)$ are greater than unity: $\frac{1}{|\lambda_i|} > 1, \forall i$.

Vector Difference Equation

We can rewrite the p th-order difference equation as

$$\mathbf{y}_t = \Phi \mathbf{y}_{t-1} + \mathbf{w}_t,$$

where

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} \quad \Phi = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad \mathbf{w}_t = \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Vector Difference Equation

It turns out that the values $\{\lambda_i\}_{i=1}^p$ are the p eigenvalues of Φ.

- So the eigenvalues of Φ are the inverses of the roots of the lag polynomial $\phi(L)$.

- Hence, $\phi(L)^{-1}$ exists if all p roots of $\phi(L)$ lie *outside* the unit circle or all p eigenvalues of Φ lie *inside* the unit circle.

Moving Average Processes

White Noise Revisited

White noise, $\{\varepsilon_t\}_{t=-\infty}^{\infty}$, is a fundamental building block of canonical time series processes.

$$\begin{aligned} E[\varepsilon_t] &= 0, \quad \forall t \\ E[\varepsilon_t^2] &= \sigma^2, \quad \forall t \\ E[\varepsilon_t \varepsilon_\tau] &= 0, \quad \text{for } t \neq \tau. \end{aligned}$$

E[(X-E(x))(Y-E(y))]
Expectations are zero
so cancel out

- We will often use the abbreviation $\{\varepsilon_t\}$.

$MA(1)$

Given white noise $\{\varepsilon_t\}$, consider the process

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

White noise has to come from a distribution that allows zero mean, so anything truncated above zero is not acceptable (ie: beta, gamma log normal). So distributions coming from the normal gaussian family is cool

where μ and θ are constants.

- This is a *first-order moving average* or $MA(1)$ process.
- We can rewrite in terms of the lag operator:

$$Y_t = \mu + \theta(L)\varepsilon_t, .$$

where $\theta(L) = (1 + \theta L)$.

$MA(1)$ Mean and Variance

The mean of the first-order moving average process is

$$\begin{aligned} \mathrm{E}[Y_t] &= \mathrm{E}[\mu + \varepsilon_t + \theta\varepsilon_{t-1}] \\ &= \mu + \mathrm{E}[\varepsilon_t] + \theta\mathrm{E}[\varepsilon_{t-1}] \\ &= \mu. \end{aligned}$$

Processing math: 100%

$MA(1)$ Autocovariances

$$
\begin{aligned}
\gamma_j &= \mathrm{E}\left[(Y_t-\mu)(Y_{t-j}-\mu)\right] \\
&= \mathrm{E}\left[(\varepsilon_t+\theta\varepsilon_{t-1})(\varepsilon_{t-j}+\theta\varepsilon_{t-j-1})\right] \\
&= \mathrm{E}[\varepsilon_t\varepsilon_{t-j}+\theta\varepsilon_t\varepsilon_{t-j-1}+\theta\varepsilon_{t-1}\varepsilon_{t-j}+\theta^2\varepsilon_{t-1}\varepsilon_{t-j-1}] \\
&= \mathrm{E}[\varepsilon_t\varepsilon_{t-j}]+\theta\mathrm{E}[\varepsilon_t\varepsilon_{t-j-1}]+\theta\mathrm{E}[\varepsilon_{t-1}\varepsilon_{t-j}]+\theta^2\mathrm{E}[\varepsilon_{t-1}\varepsilon_{t-j-1}].
\end{aligned}
$$

$MA(1)$ Autocovariances

- If $j = 0$

$$\gamma_0 = \mathrm{E}[\varepsilon_t^2]+\theta\mathrm{E}[\varepsilon_t\varepsilon_{t-1}]+\theta\mathrm{E}[\varepsilon_{t-1}\varepsilon_t]+\theta^2\mathrm{E}[\varepsilon_{t-1}^2] = (1+\theta^2)\sigma^2.$$

- If $j = 1$

$$\gamma_1 = \mathrm{E}[\varepsilon_t\varepsilon_{t-1}]+\theta\mathrm{E}[\varepsilon_t\varepsilon_{t-2}]+\theta\mathrm{E}[\varepsilon_{t-1}^2]+\theta^2\mathrm{E}[\varepsilon_{t-1}\varepsilon_{t-2}] = \theta\sigma^2.$$

- If $j > 1$, all of the expectations are zero: $\gamma_j = 0$.

$MA(1)$ Stationarity and Ergodicity

Since the mean and autocovariances are independent of time, an $MA(1)$ is weakly stationary.

- This is true *for all values of* θ.

The condition for ergodicity of the mean also holds:

$$
\begin{aligned}
\sum_{j=0}^{\infty} |\gamma_j| &= \gamma_0+\gamma_1 \\
&= (1+\theta^2)\sigma^2 + \left|\theta\sigma^2\right| < \infty.
\end{aligned}
$$

- If $\{\varepsilon_t\}$ is *Gaussian* then $\{Y_t\}$ is also ergodic *for all moments.*

$MA(1)$ Autocorrelations

The autocorrelations of an $MA(1)$ are

- $j = 0$: $\rho_0 = 1$ (*always*).

- $j = 1$:

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

- $j > 1: \rho_j = 0.$

- If $\theta > 0,$ first-order lags of Y_t are *positively* autocorrelated.

- If $\theta < 0,$ first-order lags of Y_t are *negatively* autocorrelated.

- $\max\ \{\rho_1\} = 0.5$ and occurs when $\theta = 1.$

- $\min\ \{\rho_1\} = -0.5$ and occurs when $\theta = -1.$

$MA(1)$ Example

```
##################################################################
# Simulate MA(1)
##################################################################

# Simulate MA(1)
N = 1000000;
sigma = 0.5;
eps = rnorm(N, 0, sigma);

# Simulate
mu = 0.61;
theta = 0.8;
y = mu + eps[2:N] + theta*eps[1:(N-1)];

# Plot
png(file="ma1ExampleSeries.png", height=800, width=1000)
plot(y, main=paste("MA(1), ",expression(theta)," = ",theta, sep=""),type="l")
dev.off()
```

Check the ACF plot to see if there are spikes beyond the 2 standard deviation blue lines r will generate automatically. The more spikes, and less convergence, the more evidence that the process isnt MA

$MA(1)$ Example

Unit1-ARMA/_static/ARMA/ma1ExampleSeries.png

$MA(1)$ Autocorrelations

```
##################################################################
# Plot ACF for MA(1)
##################################################################

# Plot the empirical acf
png(file="ma1ACF.png", height=800, width=1000)
acf(y, main="Autocorrelations for MA(1)")
dev.off()
```

rho=1/(n-1)(sum[(y_t-u)(y_t-j-u)])

The limits on the summation start from t=j+1 to n

Note that the larger j is, the less data you can use in your MA

Processing math: 100%

$MA(1)$ Autocorrelations

Unit1-ARMA/_static/ARMA/ma1ACF.png

$MA(1)$ Autocorrelations

```
##################################################################
# Plot lag 1 autocorrelation for different MA(1)
##################################################################

# Construct a grid of first-order coefficients
N = 10000;
thetaGrid = seq(-3, 3, length=N);

# Compute the lag 1 autocorrelations
rho1 = thetaGrid/(1+thetaGrid^2);

# Plot
png(file="ma1Lag1.png", height=600, width=1000)
plot(thetaGrid, rho1, type='l', xlab=expression(theta), ylab=expression(rho[1]),
     main="Lag 1 Autocorrelation for MA(1)")
abline(h=0);
abline(h=0.5, lty=3);
abline(h=-0.5, lty=3);
dev.off()
```

Take derivative of the ACF function and evaluate at theta*, (solving for optimimum) then plug in those values into the ACF function to find the values of the max and min of the ACF function

In this example, the MA(1) process will be within the range (-.5,.5)

This function doesnt pass the horizontal line test meaning that there will be a few MA processes that satisfy a given data set

$MA(1)$ Autocorrelations

Unit1-ARMA/_static/ARMA/ma1Lag1.png

$MA(1)$ Autocorrelations

From the figure above we see that there are two values of θ that generate each value of

$$\rho_1.$$

- In fact, θ and $1/\theta$ correspond to the same ρ_1:

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

$MA(1)$ Autocorrelations

Consider:

Processing math: 100%

$$Y_t = \varepsilon_t + 0.5\varepsilon_{t-1}$$
$$Y_t = \varepsilon_t + 2\varepsilon_{t-1}.$$

Then:

$$\rho_1 = \frac{0.5}{1+0.5^2} = \frac{2}{1+2^2} = 0.4.$$

$MA(q)$

A q th-order moving average or $MA(q)$ process is

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

where $\mu, \theta_1, \ldots, \theta_q$ are any real numbers.

- We can rewrite in terms of the lag operator:

$$Y_t = \mu + \theta(L)\varepsilon_t,$$

where $\theta(L) = (1 + \theta_1 L^1 + \ldots + \theta_q L^q)$.

$MA(q)$ Mean

As with the $MA(1)$:

$$\begin{aligned} \mathrm{E}[Y_t] &= \mathrm{E}[\mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}] \\ &= \mu + \mathrm{E}[\varepsilon_t] + \theta_1\mathrm{E}[\varepsilon_{t-1}] + \ldots + \theta_q\mathrm{E}[\varepsilon_{t-q}] \\ &= \mu. \end{aligned}$$

$MA(q)$ Autocovariances

$$\begin{aligned} \gamma_j &= \mathrm{E}\Big[(Y_t - \mu)(Y_{t-j} - \mu)\Big] \\ &= \mathrm{E}\Big[(\varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}) \\ &\qquad \times (\varepsilon_{t-j} + \theta_1\varepsilon_{t-j-1} + \ldots + \theta_q\varepsilon_{t-j-q})\Big]. \end{aligned}$$

- For $j > q$, all of the products result in zero expectations: $\gamma_j = 0$, for $j > q$.
- For $j = 0$, the squared terms result in nonzero expectations, while the cross products lead to zero expectations:

$$\gamma_0 = \mathrm{E}[\varepsilon_t^2] + \theta_1^2\mathrm{E}[\varepsilon_{t-1}^2] + \ldots + \theta_q^2\mathrm{E}[\varepsilon_{t-q}^2] = \left(1 + \sum_{j=1}^{q}\theta_j^2\right)\sigma^2.$$

$MA(q)$ Autocovariances

- For $j = \{1, 2, \ldots, q\}$, the nonzero expectation terms are

$$\begin{aligned}\gamma_j &= \theta_j \mathrm{E}[\varepsilon_{t-j}^2] + \theta_{j+1}\theta_1 \mathrm{E}[\varepsilon_{t-j-1}^2] \\ &\qquad + \theta_{j+2}\theta_2 \mathrm{E}[\varepsilon_{t-j-2}^2] + \ldots + \theta_q\theta_{q-j} \mathrm{E}[\varepsilon_{t-q}^2] \\ &= (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q\theta_{q-j})\sigma^2.\end{aligned}$$

The autocovariances can be stated concisely as

$$\gamma_j = \begin{cases} (\theta_j\theta_0 + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q\theta_{q-j})\sigma^2 & \text{for } j = 0, 1, \ldots, q \\ 0 & \text{for } j > q. \end{cases}$$

where $\theta_0 = 1$.

$MA(q)$ Autocorrelations

The autocorrelations can be stated concisely as

$$\rho_j = \begin{cases} \frac{\theta_j\theta_0 + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q\theta_{q-j}}{\theta_0^2 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2} & \text{for } j = 0, 1, \ldots, q \\ 0 & \text{for } j > q. \end{cases}$$

where $\theta_0 = 1$.

$MA(2)$ Example

For an $MA(2)$ process

$$\begin{aligned}\gamma_0 &= (1 + \theta_1^2 + \theta_2^2)\sigma^2 \\ \gamma_1 &= (\theta_1 + \theta_2\theta_1)\sigma^2 \\ \gamma_2 &= \theta_2\sigma^2 \\ \gamma_3 &= \gamma_4 = \ldots = 0.\end{aligned}$$

$MA(q)$ Stationarity and Ergodicity

Since the mean and autocovariances are independent of time, an $MA(q)$ is weakly stationary.

- This is true *for all values of* $\{\theta_j\}_{j=1}^q$.

The condition for ergodicity of the mean also holds:

$$\sum_{j=0} |\gamma_j| = \sum_{j=0} |\gamma_j| < \infty.$$

- If $\{\varepsilon_t\}$ is *Gaussian* then $\{Y_t\}$ is also ergodic *for all moments*.

$MA(\infty)$

If $\theta_0 = 1$, the $MA(q)$ process can be written as

$$Y_t = \mu + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}.$$

- If we take the limit $q \to \infty$:

$$Y_t = \mu + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} = \mu + \theta(L)\varepsilon_t,$$

where $\theta(L) = \sum_{j=0}^{\infty} \theta_j L^j$.

- It can be shown that an $MA(\infty)$ process is weakly stationary if

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

$MA(\infty)$

Since absolute summability implies square summability

$$\sum_{j=0}^{\infty} |\theta_j| \Rightarrow \sum_{j=0}^{\infty} \theta_j^2,$$

an $MA(\infty)$ process satisfying absolute summability is also weakly stationary.

- In general

$$\sum_{j=0}^{\infty} \theta_j^2 \nRightarrow \sum_{j=0}^{\infty} |\theta_j|.$$

$MA(\infty)$ Moments

Following the same reasoning as above,

$$\mathrm{E}[Y_t] = \mu$$

$$\gamma_j = \sigma^2 \sum_{i=0}^{\infty} \theta_{j+i}\theta_i.$$

- $\sum_{j=0}^{\infty} |\theta_j| \Rightarrow \sum_{j=0}^{\infty} |\gamma_j|$.

- So if the $MA(\infty)$ has absolutely summable coefficients, it is ergodic for the mean.

- Further, if $\{\varepsilon_t\}$ is *Gaussian* then $\{Y_t\}$ is also ergodic *for all moments*.

Autoregressive Processes

$AR(1)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

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

where c and ϕ are constants.

- This is a *first-order autoregressive* or $AR(1)$ process.
- We can rewrite in terms of the lag operator:

$$(1 - \phi L)Y_t = c + \varepsilon_t.$$

$AR(1)$ as $MA(\infty)$

From our discussion of lag operators, we know that if $|\phi| < 1$

$$\begin{aligned}
Y_t &= (1-\phi L)^{-1}c + (1-\phi L)^{-1}\varepsilon_t \\
&= \left(\sum_{i=0}^{\infty}\phi^i L^i\right)c + \left(\sum_{i=0}^{\infty}\phi^i L^i\right)\varepsilon_t \\
&= \left(\sum_{i=0}^{\infty}\phi^i\right)c + \left(\sum_{i=0}^{\infty}\phi^i L^i\right)\varepsilon_t \\
&= \frac{c}{1-\phi} + \theta(L)\varepsilon_t,
\end{aligned}$$

where

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

$AR(1)$ as $MA(\infty)$

Restating when $|\phi| < 1$

$$Y_t = \frac{c}{1-\phi} + \theta(L)\varepsilon_t = \frac{c}{1-\phi} + \sum_{i=0}^{\infty}\phi^i\varepsilon_{t-i}.$$

- This is an $MA(\infty)$ with $\mu = c/(1-\phi)$ and $\theta_i = \phi^i$.

- Note that $|\phi| < 1$ implies

$$\sum_{j=0}^{\infty}|\theta_j| = \sum_{j=0}^{\infty}|\phi|^j < \infty,$$

which means that Y_t is weakly stationary.

Expectation of $AR(1)$

Assume Y_t is weakly stationary: $|\phi| < 1$.

$$\begin{aligned} \mathrm{E}[Y_t] &= c + \phi\mathrm{E}[Y_{t-1}] + \mathrm{E}[\varepsilon_t] \\ &= c + \phi\mathrm{E}[Y_t] \\ \Rightarrow \mathrm{E}[Y_t] &= \frac{c}{1-\phi}. \end{aligned}$$

A Useful Property

If Y_t is weakly stationary,

$$Y_{t-j} - \mu = \sum_{i=0}^{\infty}\phi^i\varepsilon_{t-j-i}.$$

- That is, for $j \geq 1$, Y_{t-j} is a function of lagged values of ε_t and not ε_t itself.

- As a result, for $j \geq 1$

$$\mathrm{E}\left[(Y_{t-j} - \mu)\varepsilon_t\right] = \sum_{i=0}^{\infty}\phi^i\mathrm{E}[\varepsilon_t\varepsilon_{t-j-i}] = 0.$$

Variance of $AR(1)$

Given that $\mu = c/(1-\phi)$ for weakly stationary Y_t:

$$\begin{aligned} Y_t &= \mu(1-\phi) + \phi Y_{t-1} + \varepsilon_t \\ \Rightarrow (Y_t - \mu) &= \phi(Y_{t-1} - \mu) + \varepsilon_t. \end{aligned}$$

Squaring both sides and taking expectations:

$$
\begin{aligned}
\mathrm{E}\left[(Y_t-\mu)^2\right] &= \phi^2\mathrm{E}\left[(Y_{t-1}-\mu)^2\right] + 2\phi\mathrm{E}\left[(Y_{t-1}-\mu)\varepsilon_t\right] + \mathrm{E}[\varepsilon_t^2] \\
&= \phi^2\mathrm{E}\left[(Y_t-\mu)^2\right] + \sigma^2 \\
&\Rightarrow (1-\phi^2)\gamma_0 = \sigma^2 \\
&\Rightarrow \gamma_0 = \frac{\sigma^2}{1-\phi^2}
\end{aligned}
$$

Autocovariances of $AR(1)$

For $j \geq 1$,

$$
\begin{aligned}
\gamma_j &= \mathrm{E}\left[(Y_t-\mu)(Y_{t-j}-\mu)\right] \\
&= \phi\mathrm{E}[(Y_{t-1}-\mu)(Y_{t-j}-\mu)] + \mathrm{E}[\varepsilon_t(Y_{t-j}-\mu)] \\
&= \phi\gamma_{j-1} \\
&\vdots \\
&= \phi^j\gamma_0.
\end{aligned}
$$

Autocorrelations of $AR(1)$

The autocorrelations of an $AR(1)$ are

$$
\rho_j = \frac{\gamma_j}{\gamma_0} = \phi^j, \quad \forall j \geq 0.
$$

- Since we assumed $|\phi| < 1$, the autocorrelations decay exponentially as j increases.

- Note that if $\phi \in (-1, 0)$, the autocorrelations decay in an oscillatory fashion.

Examples of $AR(1)$ Processes

```
############################################################
# Simulate AR(1) processes for different values of phi
############################################################

# Number of simulated points
nSim = 1000000;

# Values of phi to consider
phi = c(-0.9, 0, 0.9, 0.99);

# Draw one set of shocks and use for each AR(1)
eps = rnorm(nSim, 0, 1);

# Matrix which stores each AR(1) in columns
y = matrix(0, nrow=nSim, ncol=length(phi));

# Each process is intialized at first shock
y[1,] = eps[1];

# Loop over each value of phi
for(j in 1:length(phi)){

    # Loop through the series, simulating the AR(1) values
    for(i in 2:nSim){
        y[i,j] = phi[j]*y[i-1,j]+eps[i]
    }
 }
```

Examples of $AR(1)$ Processes

```
############################################################
# Plot the AR(1) realizations for each phi
############################################################

# Only plot a subset of the whole simulation
plotInd = 1:1000

# Specify a plot grid
png(file="ar1ExampleSeries.png", height=600, width=1000)
par(mfrow=c(2,2))

# Loop over each value of phi
for(j in 1:length(phi)){
    plot(plotInd,y[plotInd,j], type='l', xlab='Time Index',
         ylab="Y", main=paste(expression(phi), " = ", phi[j], sep=""))
    abline(h=0)
 }
 graphics.off()
```

Examples of $AR(1)$ Processes

Unit1-ARMA/_static/ARMA/ar1ExampleSeries.png

$AR(1)$ Autocorrelations

```
############################################################
# Plot the sample ACFs for each AR(1) simulation
# For large nSim, sample ACFs are close to true ACFs
############################################################

# Specify a plot grid
png(file="ar1ExampleACF.png", height=600, width=1000)
par(mfrow=c(2,2))

# Loop over each value of phi
for(j in 1:length(phi)){
    acf(y[,j], main=paste(expression(phi), " = ", phi[j], sep=""))
}
graphics.off()
```

$AR(1)$ Autocorrelations

Unit1-ARMA/_static/ARMA/ar1ExampleACF.png

$AR(p)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

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

where c and $\{\phi\}_{i=1}^{p}$ are constants.

- This is a p *th-order autoregressive* or $AR(p)$ process.

- We can rewrite in terms of the lag operator:

$$\phi(L)Y_t = c + \varepsilon_t.$$

where

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

$AR(p)$ as $MA(\infty)$

From our discussion of lag operators,

$$Y_t = \phi(L)^{-1}c + \phi(L)^{-1}\varepsilon_t,$$

if the roots of $\phi(L)$ all lie outside the unit circle.

- In this case, $\phi(L) = (1 - \lambda_1 L)(1 - \lambda_2 L)\cdots(1 - \lambda_p L)$.

- If the roots, $\frac{1}{|\lambda_i|} > 1, \forall i$ then $|\lambda_i| < 1$,

  $\forall i$ and

$$\phi(L)^{-1} = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}\cdots(1 - \lambda_p L)^{-1}$$

$$= \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right)\left(\sum_{j=0}^{\infty} \lambda_2^j L^j\right)\cdots\left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right).$$

$AR(p)$ as $MA(\infty)$

For $|\lambda_i| < 1, \forall i$

- Y_t is an $MA(\infty)$ with $\mu = \phi(L)^{-1}c$ and :math:`smash{theta(L) =

  phi(L)^{-1}}`.

- It can be shown that $\sum_{i=1}^{\infty} |\theta_i| < \infty$.

- As a result, Y_t is weakly stationary.

Vector Autoregressive Process

We can rewrite the $AR(p)$ as

$$\mathbf{Y}_t = \mathbf{c} + \Phi \mathbf{Y}_{t-1} + \varepsilon_t,$$

where

$$\mathbf{Y}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{bmatrix} \quad \Phi = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad \varepsilon_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and $\mathbf{c} = (c, c, \ldots, c)'_{1 \times p}$.

Vector Autoregressive Process

It turns out that the values $\{\lambda_i\}_{i=1}^{p}$ are the p eigenvalues of Φ.

- So the eigenvalues of Φ are the inverses of the roots of the lag polynomial $\phi(L)$.
- Hence, $\phi(L)^{-1}$ exists if all p roots of $\phi(L)$ lie *outside* the unit circle or all p eigenvalues of Φ lie *inside* the unit circle.
- These conditions ensure weak stationarity of the $AR(p)$ process.

Expectation of $AR(p)$

Assume Y_t is weakly stationary: the roots of $\phi(L)$ lie outside the unit circle.

$$\begin{aligned} \mathrm{E}[Y_t] &= c + \phi_1 \mathrm{E}[Y_{t-1}] + \ldots + \phi_p \mathrm{E}[Y_{t-p}] + \mathrm{E}[\varepsilon_t] \\ &= c + \phi_1 \mathrm{E}[Y_t] + \ldots + \phi_p \mathrm{E}[Y_t] \\ \Rightarrow \mathrm{E}[Y_t] &= \frac{c}{1 - \phi_1 - \ldots - \phi_p} = \mu. \end{aligned}$$

Autocovariances of $AR(p)$

Given that $\mu = c/(1 - \phi_1 - \ldots - \phi_p)$ for weakly stationary Y_t:

$$\begin{aligned} Y_t &= \mu(1 - \phi_1 - \ldots - \phi_p) + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \\ \Rightarrow (Y_t - \mu) &= \phi_1 (Y_{t-1} - \mu) + \ldots + \phi_p (Y_{t-p} - \mu) + \varepsilon_t. \end{aligned}$$

Thus,

$$\begin{aligned} \gamma_j &= \mathrm{E}\left[(Y_t - \mu)(Y_{t-j} - \mu)\right] \\ &= \phi_1 \mathrm{E}\left[(Y_{t-1} - \mu)(Y_{t-j} - \mu)\right] + \ldots \\ &\qquad + \phi_p \mathrm{E}\left[(Y_{t-p} - \mu)(Y_{t-j} - \mu)\right] + \mathrm{E}\left[\varepsilon_t (Y_{t-j} - \mu)\right] \\ &= \begin{cases} \phi_1 \gamma_{j-1} + \ldots + \phi_p \gamma_{j-p} & \text{for } j = 1, \ldots \\ \phi_1 \gamma_1 + \ldots + \phi_p \gamma_p + \sigma^2 & \text{for } j = 0. \end{cases} \end{aligned}$$

Processing math: 100%

riances of $AR(p)$

For $j = 0, 1, \ldots, p$, the equations above are a system of $p + 1$ equations with $p + 1$ unknowns: $\{\gamma_j\}_{j=0}^{p}$.

- $\{\gamma_j\}_{j=0}^{p}$ can be solved for as functions of $\{\phi_j\}_{j=1}^{p}$ and σ^2.
- It can be shown that $\{\gamma_j\}_{j=0}^{p}$ are the first p elements of the first column of $\sigma^2[I_{p^2} - \Phi \otimes \Phi]^{-1}$, where

    $\otimes$ denotes the Kronecker product.

- $\{\gamma_j\}_{j=p+1}^{\infty}$ can then be determined using prior values of γ_j and $\{\phi_j\}_{j=1}^{p}$.

Autocorrelations of $AR(p)$

Dividing the autocovariances by γ_0,

$$\rho_j = \phi_1 \rho_{j-1} + \ldots + \phi_p \rho_{j-p} \quad \text{for } j = 1, \ldots$$

ARMA Processes

$ARMA(p,q)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

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

where c, $\{\phi_i\}_{i=1}^p$ and $\{\theta_i\}_{i=1}^q$ are constants.

- This is an $ARMA(p,q)$ process.

$ARMA(p,q)$ Process

We can rewrite in terms of lag operators:

$$\phi(L) Y_t = c + \theta(L) \varepsilon_t,$$

where

$$\begin{aligned} \phi(L) &= 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p \\ \theta(L) &= 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q. \end{aligned}$$

$ARMA(p,q)$ as $MA(\infty)$

Recall

- $\phi(L) = (1-\lambda_1 L)(1-\lambda_2 L)\cdots(1-\lambda_p L)$.
- If the roots, $\frac{1}{|\lambda_i|} > 1, \forall i$ then $|\lambda_i| < 1, \forall i$ and

$$\begin{aligned} \phi(L)^{-1} &= (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1}\cdots(1-\lambda_p L)^{-1} \\ &= \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right)\left(\sum_{j=0}^{\infty} \lambda_2^j L^j\right)\cdots\left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right). \end{aligned}$$

$ARMA(p,q)$ as $MA(\infty)$

Thus, if the roots of $\phi(L)$ all lie outside the unit circle,

$$Y_t = \mu + \psi(L)\varepsilon_t,$$

where $\mu = \phi(L)^{-1}c$ and $\psi(L) = \phi(L)^{-1}\theta(L)$.

- This restriction on the roots of $\phi(L)$ results in
$$\sum_{i=1}^{\infty} |\psi_i| < \infty.$$
- Hence, Y_t is an $MA(\infty)$ process and is weakly stationary.

- The stationarity of an $ARMA(p, q)$ depends only on $\{\phi_i\}_{i=1}^{p}$ and not on $\{\theta_i\}_{i=1}^{q}$.

Expectation of $ARMA(p, q)$

Assume Y_t is weakly stationary: the roots of $\phi(L)$ lie outside the unit circle.

$$\begin{aligned} \mathrm{E}[Y_t] &= c + \phi_1\mathrm{E}[Y_{t-1}] + \ldots + \phi_p\mathrm{E}[Y_{t-p}] \\ &\qquad + \mathrm{E}[\varepsilon_t] + \theta_1\mathrm{E}[\varepsilon_{t-1}] + \ldots + \theta_q\mathrm{E}[\varepsilon_{t-q}] \\ &= c + \phi_1\mathrm{E}[Y_t] + \ldots + \phi_p\mathrm{E}[Y_t] \\ \Rightarrow \mathrm{E}[Y_t] &= \frac{c}{1 - \phi_1 - \ldots - \phi_p} = \mu. \end{aligned}$$

- This is the same mean as an $AR(p)$ process with parameters c and $\{\phi_i\}_{i=1}^{p}$.

Autocovariances of $ARMA(p, q)$

Given that $\mu = c/(1 - \phi_1 - \ldots - \phi_p)$ for weakly stationary Y_t:

$$\begin{aligned} Y_t &= \mu(1 - \phi_1 - \ldots - \phi_p) + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} \\ &\qquad + \varepsilon_t + \theta_1\varepsilon_1 + \ldots\theta_q\varepsilon_{t-q} \\ \Rightarrow (Y_t - \mu) &= \phi_1(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) \\ &\qquad + \varepsilon_t + \theta_1\varepsilon_1 + \ldots\theta_q\varepsilon_{t-q}. \end{aligned}$$

$$\begin{aligned}\gamma_j &= \mathrm{E}\Big[(Y_t-\mu)(Y_{t-j}-\mu)\Big] \\ &= \phi_1\mathrm{E}\Big[(Y_{t-1}-\mu)(Y_{t-j}-\mu)\Big] + \ldots \\ &\qquad + \phi_p\mathrm{E}\Big[(Y_{t-p}-\mu)(Y_{t-j}-\mu)\Big] \\ &\qquad\qquad + \mathrm{E}\Big[\varepsilon_t(Y_{t-j}-\mu)\Big] + \theta_1\mathrm{E}\Big[\varepsilon_{t-1}(Y_{t-j}-\mu)\Big] \\ &\qquad\qquad\qquad + \ldots + \theta_q\mathrm{E}\Big[\varepsilon_{t-q}(Y_{t-j}-\mu)\Big]\end{aligned}$$

Autocovariances of $ARMA(p, q)$

- For $j > q, \gamma_j$ will follow the same law of motion as for an $AR(p)$ process:

$$\gamma_j = \phi_1\gamma_{j-1} + \ldots + \phi_p\gamma_{j-p} \quad \text{for } j = q+1, \ldots$$

- These values will not be the same as the $AR(p)$ values for $j = q+1, \ldots$, since the initial $\gamma_0, \ldots, \gamma_q$ will differ.

- The first q autocovariances, $\gamma_0, \ldots, \gamma_q$, of an $ARMA(p, q)$ will be more complicated than those of an $AR(p)$.

Redundancy of $ARMA(p, q)$

Factoring the polynomials $\phi(L)$ and $\theta(L)$, an $ARMA(p, q)$ can be written as

$$(1-\lambda_1 L)\cdots(1-\lambda_p L)(Y_t-\mu) = (1-\eta_1 L)\cdots(1-\eta_q L)\varepsilon_t.$$

- If two of the roots are identical, $\lambda_i = \eta_j$, both polynomials can be divided by $(1-\lambda_i L)$.

- The result would be an $ARMA(p-1, q-1)$:

$$(1-\phi_1^* L - \ldots - \phi_{p-1}^* L^{p-1})(Y_t-\mu) = (1+\theta_1^* L + \ldots + \theta_{q-1}^* L^{q-1})\varepsilon_t.$$

Causality and Invertibility

Causality

Suppose $\{Y_t\}$ is an $AR(1)$ process:

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

- We have shown that when $|\phi| < 1$, $\{Y_t\}$ is stationary.
- What if $|\phi| > 1$?

Causality

Let's run the AR recursion forward:

$$\begin{aligned} Y_{t-1} &= \frac{1}{\phi}Y_t - \frac{1}{\phi}\varepsilon_t = \frac{1}{\phi}\left(\frac{1}{\phi}Y_{t+1} - \frac{1}{\phi}\varepsilon_{t+1}\right) - \frac{1}{\phi}\varepsilon_t \\ &= -\frac{1}{\phi}\varepsilon_t - \left(\frac{1}{\phi}\right)^2\varepsilon_{t+1} + \left(\frac{1}{\phi}\right)^2 Y_{t+1} \\ &\vdots \\ &= -\sum_{j=0}^{\infty}\left(\frac{1}{\phi}\right)^{j+1}\varepsilon_{t+j} \\ &= -\left(\sum_{j=0}^{\infty}\left(\frac{1}{\phi}\right)^{j+1} L^{-j}\right)\varepsilon_t. \end{aligned}$$

Causality

The previous sum converges, so Y_t is stationary.

- However it is not a function of past ε_t.

Causality

A process $\{X_t\}$ is a causal function of $\{W_t\}$ if $\exists \psi(L) = \psi_0 + \psi_1 L^1 + \ldots$ such that $x_t = \psi(L) w_t$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

- An $AR(1)$ is causal only if $|\phi| < 1$.
- However it is stationary as long as $|\phi| \neq 1$.

Causality of $AR(p)$

Suppose $\{Y_t\}$ is an $AR(p)$ process with lag polynomial $\phi(L)$.

- If all roots of $\phi(L)$ are inside or outside the unit circle, $\{Y_t\}$ is stationary.
- If any root of $\phi(L)$ is on the unit circle, $\{Y_t\}$ is not stationary.
- If all roots of $\phi(L)$ are outside the unit circle, $\phi(L)^{-1}$ exists and $\{Y_t\}$ is stationary and causal.

Invertibility

Suppose $\{Y_t\}$ is an $MA(q)$ process:

$$Y_t = \mu + \theta(L)\varepsilon_t,$$

where $\theta(L) = 1 + \theta_1 L^1 + \ldots + \theta_q L^q$.

- We say $\{Y_t\}$ is *invertible* if $\theta(L)^{-1}$ exists.

Invertibility

The $MA(q)$ lag polynomial can be factored as

$$\theta(L) = 1 + \theta_1 L^1 + \ldots + \theta_q L^q = (1 - \eta_1 L) \cdots (1 - \eta_q L).$$

- $\left\{\frac{1}{\eta_i}\right\}_{i=1}^{q}$ are the roots of $\theta(L)$.

Suppose $|\eta_i| < 1 \ \forall i$. Then

$$(1 - \eta_i L)^{-1} = \sum_{j=0}^{\infty} \eta_i^j L^j \quad \forall i$$

$$\theta(L)^{-1} = \left(\sum_{j=0}^{\infty} \eta_1^j L^j\right) \cdots \left(\sum_{j=0}^{\infty} \eta_q^j L^j\right).$$

Stationarity/Invertibility

We previously showed that an $MA(q)$ is *always* stationary, regardless of the roots of $\theta(L)$.

- It is only invertible if all of the roots of $\theta(L)$ lie outside the unit circle.

- In this case

$$\varepsilon_t = \theta(L)^{-1} Y_t.$$

- That is, $\{\varepsilon_t\}$ is a causal function of $\{Y_t\}$.

Inverting an $MA(q)$

Given an $MA(q)$ process,

$$Y_t = \mu + \theta(L)\varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma^2),$$

suppose, without loss of generality,

- $|\eta_i| < 1$ for $i = 1, \ldots, m$
- $|\eta_i| > 1$ for $i = m+1, \ldots, q$.

Inverting an $MA(q)$

Create a new process

$$\tilde{Y}_t = \mu + \tilde{\theta}(L)\tilde{\varepsilon}_t, \quad \tilde{\varepsilon}_t \overset{i.i.d.}{\sim} WN(0, \sigma^2 \eta_{m+1}^2 \cdots \eta_q^2),$$

where

$$\begin{aligned} \tilde{\theta}(L) &= 1 + \tilde{\theta}_1 L^1 + \ldots + \tilde{\theta}_q L^q \\ &= \left(1 - \eta_1 L\right) \cdots \left(1 - \eta_m L\right) \cdot \left(1 - \frac{1}{\eta_{m+1}} L\right) \cdots \left(1 - \frac{1}{\eta_q} L\right). \end{aligned}$$

Inverting an $MA(q)$

It can be shown that $\tilde{Y}_t$ has the same first and second moments as Y_t.

- $\tilde{Y}_t$ is known as the invertible represenation of the $MA(q)$ process.

- Note that every $MA(q)$ process has an invertible representation so long as none of the roots of $\theta(L)$ lie on the unit circle.

- If an invertible representation exists, it is unique.

Causality and Invertibility of an $ARMA(p, q)$

The notions of stationarity, causality and invertibility extend to an $ARMA(p, q)$ process:

$$\phi(L)Y_t = c + \theta(L)\varepsilon_t.$$

- If none of the roots of $\phi(L)$ lie on the unit circle, $\{Y_t\}$ is stationary.

- If all of the roots of $\phi(L)$ lie outside the unit circle, $\{Y_t\}$ is causal.

- If none of the roots of $\theta(L)$ lie on the unit circle, $\{Y_t\}$ has a unique invertible representation.

Maximum Likelihood Estimation

Estimating Parameters of Distributions

We almost never know the true distribution of a data sample.

- We might hypothesize a family of distributions that capture broad characteristics of the data (locations, scale and shape).

- However, there may be a set of one or more parameters of the distribution that we don't know.

- Typically we use the data to estimate the unknown parameters, $\boldsymbol{\theta}$.

Joint CDF

Suppose we have a collection of random variables $\mathbf{Y}_T = (Y_1, \dots, Y_T)'$.

- We view a data sample of size T as one realization of each random variable: $\mathbf{y}_T = (y_1, \dots, y_T)'$.

- The *joint cumulative density* of $\mathbf{Y}_T$ is

$$F_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta}) = P(Y_1 \leq y_1, \dots, Y_T \leq y_T).$$

Joint Density

- The *joint probability density* of $\mathbf{Y}_T$ is

$$f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta}) = \frac{\partial^T F_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta})}{\partial Y_1 \dots \partial Y_T}.$$

since

$$F_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta}) = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_T} f_{\mathbf{Y}_T}(\mathbf{a}_T | \boldsymbol{\theta}) \; da_1 \dots da_T.$$

Independence

When $Y_1, \ldots, Y_T$ are independent of each other and have identical distributions:

- We say that they are *independent and identically distributed*, or i.i.d.

- When $Y_1, \ldots, Y_T$ are i.i.d., they have the same marginal densities:

$$f_{Y_1}(y \mid \boldsymbol{\theta}) = \ldots = f_{Y_T}(y \mid \boldsymbol{\theta}).$$

Joint Density Under Independence

Further, when $Y_1, \ldots, Y_T$ are i.i.d.

$$f_{\mathbf{Y}_T}(\mathbf{y}_T \mid \boldsymbol{\theta}) = f_{Y_1}(y_1 \mid \boldsymbol{\theta}) \cdot f_{Y_2}(y_2 \mid \boldsymbol{\theta}) \cdots f_{Y_T}(y_T \mid \boldsymbol{\theta}) = \prod_{i=1}^{T} f_{Y_i}(y_i \mid \boldsymbol{\theta}).$$

- This is analogous to the computation of joint probabilities.

- For independent events A, B and C,

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Bayes' Rule

Recall Bayes' Rule:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ \Rightarrow P(A \cap B) &= P(A \mid B)P(B). \end{aligned}$$

Iterating:

$$\begin{aligned} P(A \cap B \cap C) &= P((A \cap B) \cap C) \\ &= P(A \cap B \mid C)P(C) \\ &= P(A \mid B, C)P(B \mid C)P(C). \end{aligned}$$

Bayes' Rule

The previous iteration generalizes to be:

$$
\begin{aligned}
P\left(\bigcap_{i=1}^{n} A_i\right) &= P(A_n | A_{n-1}, \ldots, A_1) \\
&\quad \times P(A_{n-1} | A_{n-2}, \ldots, A_1) \cdots P(A_2 | A_1) P(A_1) \\
&= P(A_1) \prod_{i=2}^{n} P(A_i | \mathbf{A}_{i-1})
\end{aligned}
$$

where $\mathbf{A_i} = (A_1, \ldots, A_i)'$

General Joint Density

Suppose $Y_1, \ldots, Y_T$ depend on each other sequentially.

- We use Bayes' Rule to obtain the joint density:

$$f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta}) = f_{Y_1}(y_1 | \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t | \mathbf{Y}_{t-1}}(y_t | \mathbf{y}_{t-1}, \boldsymbol{\theta}).$$

Maximum Likelihood Estimation

One of the most important and powerful methods of parameter estimation is *maximum likelihood estimation*. It requires

- A data sample: $\mathbf{y} = (y_1, \ldots, y_T)'$.

- A joint probability density:

$$f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta}) = f_{Y_1}(y_1 | \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t | \mathbf{Y}_{t-1}}(y_t | \mathbf{y}_{t-1}, \boldsymbol{\theta}).$$

Likelihood

$f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta})$ is loosely interpreted as the probability of observing data sample $\mathbf{y}_T$, given a functional form for the density of $Y_1, \ldots, Y_T$ and given a set of parameters $\boldsymbol{\theta}$.

- We can reverse the notion and think of $\mathbf{y}_T$ as being fixed and $\boldsymbol{\theta}$ some unknown variable.

- In this case we write $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T) = f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta})$.

- We refer to $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T)$ as the likelihood.

- Fixing $\mathbf{y}_T$, maximum likelihood estimation chooses the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T) = f_{\mathbf{Y}_T}(\mathbf{y}_T | \boldsymbol{\theta})$.

Likelihood Maximization

Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, we maximize $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T)$ by

- Differentiating with respect to each θ_i, $i = 1, \ldots, p$.

- Setting the resulting derivatives equal to zero.

- Solving for the values $\hat{\theta}_i$, $i = 1, \ldots, p$, that make all of the derivatives zero.

Log Likelihood

It is often easier to work with the logarithm of the likelihood function.

- By the properties of logarithms

$$\begin{aligned}
\ell(\boldsymbol{\theta} | \mathbf{y}_T) &= \log\left(\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T)\right) \\
&= \log\left(f_{Y_1}(y_1 | \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t | \mathbf{Y}_{t-1}}(y_t | \mathbf{y}_{t-1}, \boldsymbol{\theta})\right) \\
&= \log\left(f_{Y_1}(y_1 | \boldsymbol{\theta})\right) + \sum_{t=2}^{T} \log\left(f_{Y_t | \mathbf{Y}_{t-1}}(y_t | \mathbf{y}_{t-1}, \boldsymbol{\theta})\right)
\end{aligned}$$

Log Likelihood

- Maximizing $\ell(\boldsymbol{\theta} | \mathbf{y}_T)$ is the same as maximizing $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_T)$ since $\log$ is a monotonic transformation.

- A derivative of $\mathcal{L}$ will involve many product-rule terms, whereas a derivative of ℓ will simply be a sum of derivatives.

MLE Example

Suppose we have a dataset $\mathbf{y}_n = (y_1, \ldots, y_n)$, where

$$Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2).$$

- We will assume μ is *unknown* and σ is *known*.

- So, $\boldsymbol{\theta} = \mu$ (it is a single value, rather than a vector).

MLE Example

- The likelihood is

$$\begin{aligned}
\mathcal{L}(\mu \mid \mathbf{y}_n) &= f_{\mathbf{Y}_n}(\mathbf{y}_n \mid \mu) \\
&= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.
\end{aligned}$$

MLE Example

The log likelihood is

$$\ell(\mu \mid \mathbf{y}_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

MLE Example

- The MLE, $\hat{\mu}$, is the value that sets $\frac{d}{d\mu}\ell(\mu \mid \mathbf{y}) = 0$:

$$\begin{aligned}
\frac{d}{d\mu}\ell(\mu \mid \mathbf{y}_n)\bigg|_{\hat{\mu}} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \\
&\Rightarrow \sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \\
&\Rightarrow \sum_{i=1}^{n}\hat{\mu} = \sum_{i=1}^{n} y_i \\
&\Rightarrow \hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.
\end{aligned}$$

MLE Example: $n = 1$, Unknown μ

Suppose we have only one observation: y_1.

- If we specialize the previous result:

$$\hat{\mu} = y_1.$$

- The density $f_{Y_1}(y_1 | \mu)$ gives the probability of observing some data value y_1, conditional on some *known* parameter μ.

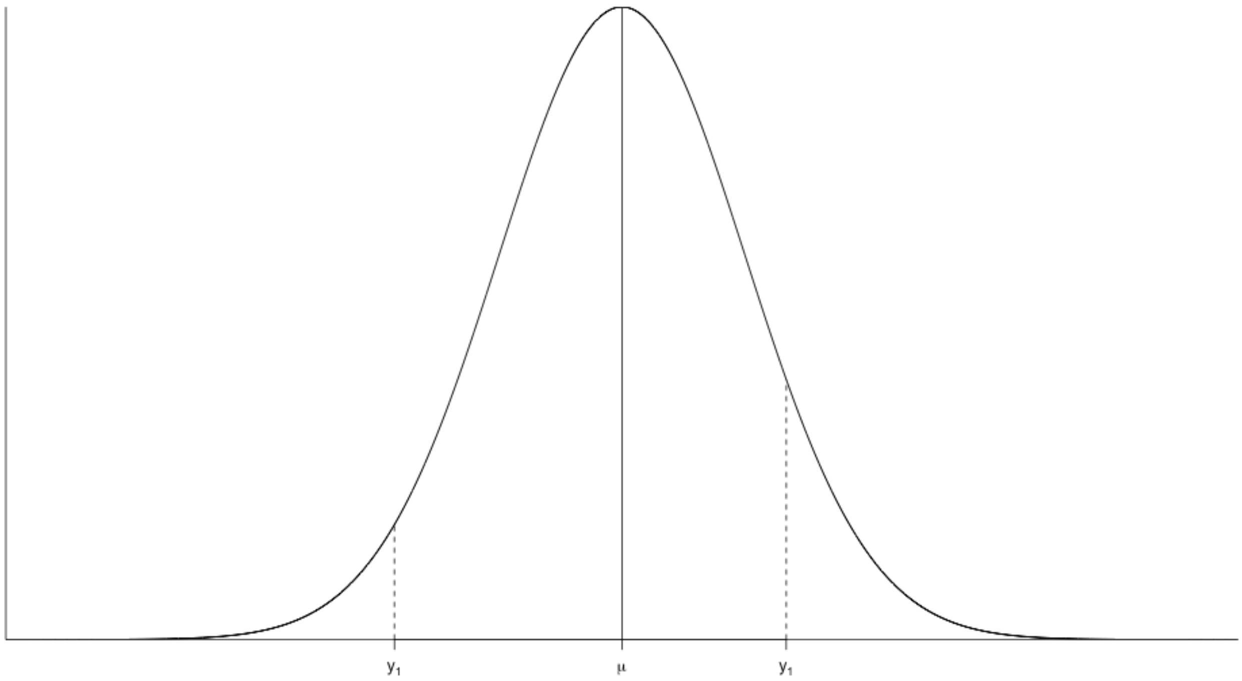
- This is a normal distribution with mean μ and variance σ^2.

MLE Example: $n = 1$, Unknown μ

- The likelihood $\mathcal{L}(\mu | y_1)$ gives the probability of μ, conditional on some observed data value y_1.

- This is a normal distribution with mean y_1 and variance σ^2.

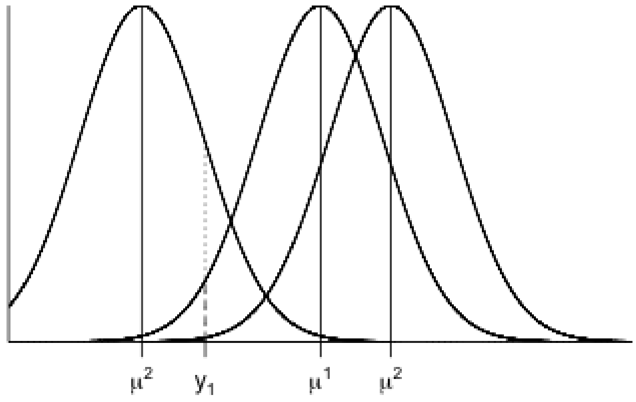
MLE Example: $n = 1$

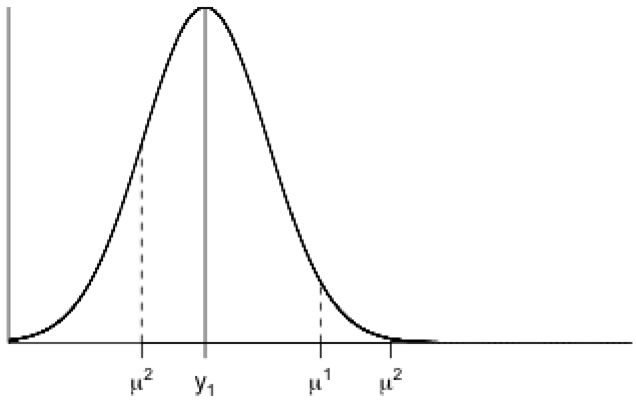


MLE Example: $n = 1$

```
#######################################################################################
# Plot likelihood for normal data with unknown mean
#######################################################################################

# Generate the true normal density
mu = 10;
sig = 15;
xGrid = seq(mu-5*sig, mu+5*sig, length=10000)
trueDens = dnorm(xGrid, mu, sig)

# A couple of possible data values
y11 = 30;
y12 = -17.7;

# Plot the true normal distribution with possible data values
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu), expression(y[1]), expression(y[1])),
     at=c(mu, y11, y12))
abline(v=mu)
segments(y11, 0, y11, dnorm(y11, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
dev.copy(png, file="densExample.png", height=600, width=1000)
dev.off()
```

MLE Example: $n = 1$, Unknown μ





MLE Example: $n = 1$, Unknown μ

```
# Plot several densities for fixed data observation
mu1 = -33
mu2 = 27
dens1 = dnorm(xGrid, mu1, sig)
dens2 = dnorm(xGrid, mu2, sig)
par(mfrow=c(2,1))
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu^1), expression(mu^2), expression(mu^2), expression(y[1])),
     at=c(mu, mu1, mu2, y12))
lines(xGrid, dens1)
lines(xGrid, dens2)
abline(v=mu)
abline(v=mu1)
abline(v=mu2)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu1, sig), lty=3)
segments(y12, 0, y12, dnorm(y12, mu2, sig), lty=4)

# Plot the resulting likelihood
like = dnorm(xGrid, y12, sig)
plot(xGrid, like, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu^1), expression(mu^2), expression(mu^2), expression(y[1])),
     at=c(mu, mu1, mu2, y12))
abline(v=y12)
segments(mu, 0, mu, dnorm(mu, y12, sig), lty=2)
segments(mu1, 0, mu1, dnorm(mu1, y12, sig), lty=2)
segments(mu2, 0, mu2, dnorm(mu2, y12, sig), lty=2)
dev.copy(png, file="likeExample.png", height=600, width=400)
dev.off()
```

MLE Example: $n = 1$, Unknown σ

Let’s continue with the assumption of one data observation, y_1.

- If μ is known but σ is unknown, the density of the data, y_1, is still normal.

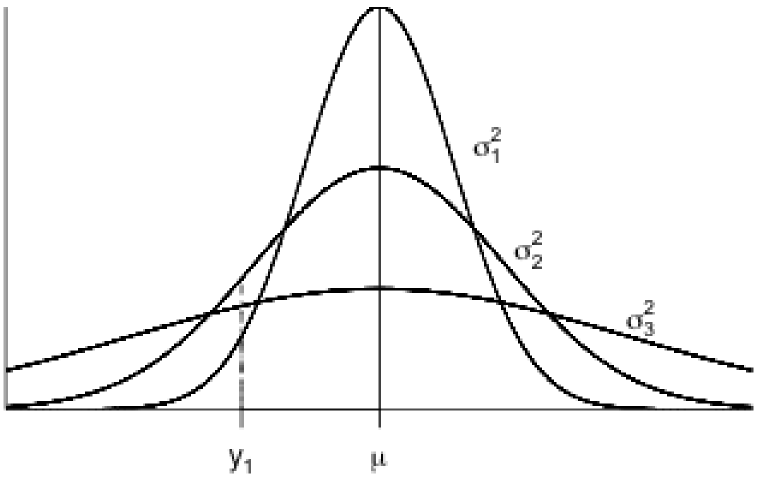
- However, the likelihood is

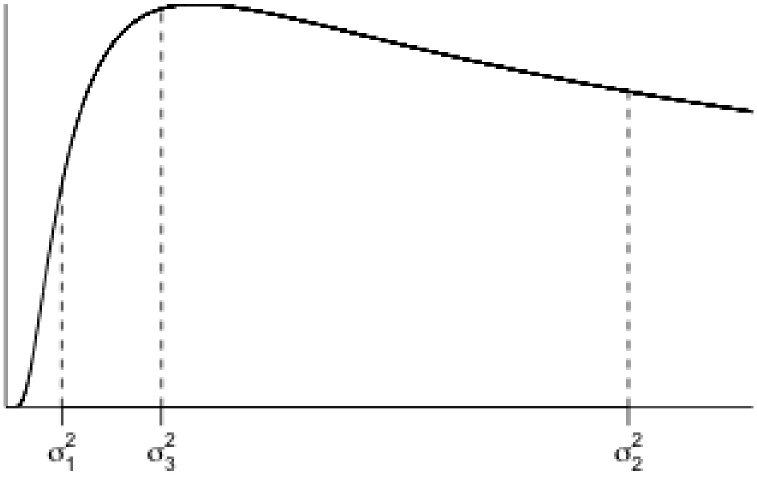
$$\mathcal{L}(\sigma^2 | y_1) = \frac{\alpha}{\sigma^2} \exp \left\{ -\frac{\beta}{\sigma^2} \right\}$$

$$\alpha = \frac{1}{\sqrt{2\pi}}, \qquad \beta = \frac{(y_1 - \mu)^2}{2}.$$

- The likelihood for σ^2 is *not* normal, but *inverse gamma*.

MLE Example: $n = 1$, Unknown σ





MLE Example: $n = 1$, Unknown σ

```
#######################################################################################
# Plot likelihood for normal data with unknown sd
#######################################################################################

# Plot several densities for fixed data observation
sig1 = 50
sig2 = 25
dens1 = dnorm(xGrid, mu, sig1)
dens2 = dnorm(xGrid, mu, sig2)
par(mfrow=c(2,1))
yMax = max(max(trueDens), max(dens1), max(dens2))
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L', ylim=c(0,yMax))
axis(1, labels=c(expression(mu), expression(y[1])), at=c(mu, y12))
lines(xGrid, dens1)
lines(xGrid, dens2)
abline(v=mu)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu, sig1), lty=3)
segments(y12, 0, y12, dnorm(y12, mu, sig2), lty=4)
xDist = max(xGrid)-mu
text(c(mu+0.29*xDist, mu+0.4*xDist, mu+0.7*xDist), c(0.66, 0.4, 0.23)*yMax,
     labels=c(expression(sigma[1]^2), expression(sigma[2]^2), expression(sigma[3]^2)))

# Plot the resulting likelihood (which is an inverse gamma)
alpha = -0.5
beta = ((y12 - mu)^2)/2
scale = 1/sqrt(2*pi)
sigGrid = seq(0.01,3000,length=10000)
like = scale*(sigGrid^(-alpha-1))*exp(-beta/sigGrid)
likeTrue = scale*(sig^(-2*alpha-2))*exp(-beta/sig^2)
like1 = scale*(sig1^(-2*alpha-2))*exp(-beta/sig1^2)
like2 = scale*(sig2^(-2*alpha-2))*exp(-beta/sig2^2)
plot(sigGrid, like, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(sigma[1]^2), expression(sigma[2]^2), expression(sigma[3]^2)),
     at=c(sig^2, sig1^2, sig2^2))
segments(sig^2, min(like), sig^2, likeTrue, lty=2)
segments(sig1^2, min(like), sig1^2, like1, lty=2)
segments(sig2^2, min(like), sig2^2, like2, lty=2)
dev.copy(png, file="likeExample2.png", height=600, width=400)
dev.off()
```

MLE Accuracy

Maximum likelihood estimation results in estimates of true unknown parameters.

- What is the probability that our estimates are identical to the true population parameters?

- Our estimates are imprecise and contain error.

- We would like to quantify the precision of our estimates with standard errors.

- We will use the *Fisher Information* to compute standard errors.

Fisher Information

Let $\mathcal{H}(\boldsymbol{\theta}|\mathbf{y}_t) = \frac{d^2}{d\boldsymbol{\theta}^2}\ell(\boldsymbol{\theta}|\mathbf{y}_T)$, the matrix of second derivatives of the log likelihood.

- The Fisher Information is

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left[\mathcal{H}(\boldsymbol{\theta}|\mathbf{y}_t)\right].$$

- The observed Fisher Information is

$$\tilde{\mathcal{I}}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta}|\mathbf{y}_t).$$

Fisher Information

- Observed Fisher Information does not take an expectation, which may be difficult to compute.

- Since $\ell(\boldsymbol{\theta}|\mathbf{y}_T)$ is often a sum of many terms, $\tilde{\mathcal{I}}(\boldsymbol{\theta})$ will converge to $\mathcal{I}(\boldsymbol{\theta})$ for large samples.

MLE Central Limit Theorem

Under certain conditions, a central limit theorem holds for the MLE, $\hat{\boldsymbol{\theta}}$.

- For infinitely large samples $\mathbf{y}_T$,

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta})^{-1}).$$

- For large samples, $\hat{\boldsymbol{\theta}}$ is normally distributed *regardless* of the distribution of the data, $\mathbf{y}_T$.

MLE Central Limit Theorem

- $\hat{\boldsymbol{\theta}}$ is also normally distributed for large samples even if $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}_T)$ is some other distribution.

- Hence, for large samples,

$$Var(\hat{\theta}_i) = \frac{1}{\mathcal{I}(\boldsymbol{\theta})_{ii}} \qquad \Rightarrow \qquad Std(\hat{\theta}_i) = \frac{1}{\sqrt{\mathcal{I}(\boldsymbol{\theta})_{ii}}}$$

MLE Standard Errors

Since we don't know the true $\boldsymbol{\theta}$, we approximate

$$Std(\hat{\theta}_i) \approx \frac{1}{\sqrt{\mathcal{I}(\hat{\boldsymbol{\theta}})_{ii}}}.$$

- Alternatively, to avoid computing the expectation, we could use the approximation

$$Std(\hat{\theta}_i) \approx \frac{1}{\sqrt{\tilde{\mathcal{I}}(\hat{\boldsymbol{\theta}})_{ii}}}.$$

MLE Standard Errors

- In reality, we never have an infinite sample size.
- For finite samples, these values are approximations of the standard errors of the components of $\hat{\boldsymbol{\theta}}$.

MLE Variance Example

Let's return to the example where $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, with known σ.

- The log likelihood is

$$\ell(\mu \,|\, \mathbf{y}_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

- The resulting derivatives are

$$\frac{\partial \ell(\mu \,|\, \mathbf{y}_n)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu), \qquad \frac{\partial^2 \ell(\mu \,|\, \mathbf{y}_n)}{\partial \mu^2} = -\frac{n}{\sigma^2}.$$

MLE Variance Example

In this case the Fisher Information is identical to the observed Fisher Information:

$$\mathcal{I}(\mu) = -E\left[-\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2} = \tilde{\mathcal{I}}(\mu).$$

- Since $\mathcal{I}(\mu)$ doesn't depend on μ, we don't need to resort to an approximation with $\hat{\mu} = \bar{\mathbf{y}}$.

- The result is

$$Std(\hat{\mu}) = \frac{1}{\sqrt{\mathcal{I}(\mu)}} = \frac{\sigma}{\sqrt{n}}.$$

ARMA Maximum Likelihood Estimation

$AR(p)$ Likelihood

Recall a Gaussian $AR(p)$ process:

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

- In this case $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \sigma^2)$.
- We will suppose that $\{Y_t\}$ is stationary and causal.

$AR(p)$ Likelihood

Suppose we know that $Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots, Y_{t-p} = y_{t-p}$ for $t > p$. Then

$$\begin{aligned} Y_t &= c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t \\ \mathrm{E}[Y_t \,|\, Y_{t-1}, \ldots, Y_{t-p}, \boldsymbol{\theta}] &= c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} \\ \mathrm{Var}(Y_t \,|\, Y_{t-1}, \ldots, Y_{t-p}, \boldsymbol{\theta}) &= \sigma^2. \end{aligned}$$

$AR(p)$ Likelihood

Thus,

$$Y_t \,|\, Y_{t-1}, \ldots, Y_{t-p} \sim \mathcal{N}(c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}, \sigma^2),$$

which means

$$\begin{aligned} & f_{Y_t | Y_{t-1}, \ldots, Y_{t-p}}(y_t \,|\, y_{t-1}, \ldots, y_{t-p}, \boldsymbol{\theta}) \\ & \qquad = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2 \right\}. \end{aligned}$$

$AR(p)$ Likelihood

The likelihood of $\boldsymbol{Y}_T = \{Y_t\}$ is

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}_T) &= f_{\boldsymbol{Y}_T}(\boldsymbol{y}_T|\boldsymbol{\theta}) \\
&= f_{\boldsymbol{Y}_p}(\boldsymbol{y}_p|\boldsymbol{\theta})\prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(y_t|y_{t-1},\ldots,y_{t-p},\boldsymbol{\theta})
\end{aligned}
$$

where $f_{\boldsymbol{Y}_p}(\boldsymbol{y}_p|\boldsymbol{\theta})$ is the joint density of $\boldsymbol{Y}_T = \{Y_t\}_{t=1}^{p}$.

- Maximizing this likelihood results in a set of nonlinear equations in $\boldsymbol{\theta}$ and $\boldsymbol{y}_T$, and must be solved numerically.

$AR(p)$ Conditional Likelihood

We can approximate the $AR(p)$ likelihood with only the product of conditional densities:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}_T) &\approx \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(y_t|y_{t-1},\ldots,y_{t-p},\boldsymbol{\theta}) \\
&= \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-\frac{T-p}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2\right\}.
\end{aligned}
$$

$AR(p)$ Conditional Log Likelihood

The conditional log likelihood of the $AR(p)$ is

$$
\ell(\boldsymbol{\theta}|\boldsymbol{y}_T) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2.
$$

- Maximizing the conditional log likelihood with respect to $c, \phi_1, \ldots, \phi_p$, conditional on σ^2, is the same as minimizing
$$
\sum_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2.
$$
- Hence, the MLEs are the same as the least squares estimates.

$AR(p)$ Conditional MLEs

Since the MLEs and LS estimates are the same, we can solve for the MLEs by simply running a regression

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e},$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} c \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} X = \begin{bmatrix} 1 & y_{T-1} & y_{T-2} & \dots & y_{T-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_p & y_{p-1} & \dots & y_1 \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} y_T \\ \vdots \\ y_{p+1} \end{bmatrix} \boldsymbol{e} = \begin{bmatrix} e_T \\ \vdots \\ e_{p+1} \end{bmatrix}.$$

$AR(p)$ Conditional MLEs

Differentiating the log likelihood with respect to σ^2,

$$\begin{aligned} \frac{\partial l}{\partial \sigma^2} \Big|_{\hat{\sigma}^2} &= -\frac{T-p}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{t=p+1}^{T} (y_t - c - \phi y_{t-1} - \dots - \phi_p y_{t-p})^2 = 0 \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - c - \phi y_{t-1} - \dots - \phi_p y_{t-p})^2 \\ &\approx \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi} y_{t-1} - \dots - \hat{\phi}_p y_{t-p})^2. \end{aligned}$$

- This is the usual regression estimator of σ^2.

$AR(p)$ Conditional MLEs

- Assuming Gaussianity doesn't impact the consistency of our estimates.

- If ε is not Gaussian, then $\hat{\boldsymbol{\beta}}$ is the Quasi Maximum Likelihood Estimate because the model is misspecified.

$MA(q)$ Conditional Likelihood

Recall a Gaussian $MA(q)$ process:

$$Y_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots + \psi_q \varepsilon_{t-q}, \quad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

- Now, $\boldsymbol{\theta} = (\mu, \psi_1, \dots, \psi_q, \sigma^2)'$.

$MA(q)$ Conditional Likelihood

If $\varepsilon_{t-q}, \ldots, \varepsilon_{t-1}$ are known with certainty then

$$
\begin{aligned}
& Y_t \sim \mathcal{N}(\mu + \psi_1\varepsilon_{t-1} + \ldots + \psi_q\varepsilon_{t-q}, \sigma^2) \\
\Rightarrow \; & f_{Y_t|\varepsilon_{t-q},\ldots,\varepsilon_{t-1}}(y_t|\varepsilon_{t-q}, \ldots, \varepsilon_{t-1}) \\
& \qquad = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \mu - \psi_1\varepsilon_{t-1} - \ldots - \psi_q\varepsilon_{t-q})^2\right\} \\
& \qquad = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\varepsilon_t^2\right\}.
\end{aligned}
$$

$MA(q)$ Conditional Likelihood

Assume $\varepsilon_0 = \varepsilon_{-1} = \varepsilon_{-2} = \ldots = \varepsilon_{-q+1} = 0$ and iteratively compute

$$
\varepsilon_t = y_t - \mu - \psi_1\varepsilon_{t-1} - \ldots - \psi_q\varepsilon_{t-q}, \quad \text{for} \quad t = 1, \ldots, T.
$$

$MA(q)$ Conditional Likelihood

Then the likelihood is

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}_T, \boldsymbol{\varepsilon}_0 = \mathbf{0}) &= f_{Y_1,\ldots,Y_T|\boldsymbol{\varepsilon}_0}(y_1, \ldots, y_T|\boldsymbol{\varepsilon}_0, \boldsymbol{\theta}) \\
&= f_{Y_1|\boldsymbol{\varepsilon}_0}(y_1|\boldsymbol{\varepsilon}_0, \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t|Y_1,\ldots,Y_t,\boldsymbol{\varepsilon}_0}(y_t|y_1, \ldots, y_t, \boldsymbol{\varepsilon}_0, \boldsymbol{\theta}) \\
&= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\varepsilon_t^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2\right\}
\end{aligned}
$$

where $\boldsymbol{\varepsilon}_0 = \{\varepsilon_t\}_{t=-q+1}^{0}$.

$MA(q)$ Conditional Log Likelihood

The log likelihood is

$$
\ell(\boldsymbol{\theta}|\boldsymbol{y}_T, \boldsymbol{\varepsilon}_0 = \mathbf{0}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2.
$$

- The MLEs cannot be found analytically.

The rough numerical algorithm is

1. Guess values for $\boldsymbol{\theta} = (\mu, \psi_1, \ldots, \psi_q, \sigma^2)'$.

2. Assume $\varepsilon_0 = \varepsilon_{-1} = \varepsilon_{-2} = \ldots = \varepsilon_{-q+1} = 0$.

3. Iteratively compute $\{\varepsilon\}_{t=1}^{T}$.

4. Evaluate the log likelihood for $\{\varepsilon\}_{t=1}^{T}$.

5. Update $\boldsymbol{\theta}$ and return to step 2 until convergence.

$MA(q)$ Conditional Log Likelihood

The conditional log likelihood function can only be used with the invertible representation of the $MA(q)$.

- If a non-invertible representation is used, it can be shown (via backward recursion on ε_t) that the assumption of $\boldsymbol{\varepsilon}_0 = \mathbf{0}$ is explosive.

$ARMA(p, q)$ Cond. Likelihood

Recall a Gaussian $ARMA(p, q)$ process:

$$
\begin{aligned}
Y_t = c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} \\
+ \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots + \psi_q \varepsilon_{t-q}, \quad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

$ARMA(p, q)$ Cond. Likelihood

To form the conditional likelihood, we combine the methods of the $AR(p)$ and $MA(q)$:

1. Condition on $y_0 = y_{-1} = \ldots = y_{-p+1} = \mu = \frac{c}{1-\phi_1 - \ldots - \phi_p}$.

2. Condition on $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$.

3. Recursively compute $\{\varepsilon_t\}_{t=1}^{T}$ using $\{y_t\}_{t=1}^{T}$, $\{\varepsilon_t\}_{t=-q+1}^{0}$ and $\{y_t\}_{t=-p+1}^{0}$.

4. Compute the log likelihood as

$$
\ell(\boldsymbol{\theta} | \boldsymbol{y}_T, \boldsymbol{\varepsilon}_0 = \mathbf{0}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} \varepsilon_t^2.
$$

The MA polynomial must be invertible in order to use the conditional log likelihood for estimation.

$ARMA(p, q)$ Cond. Likelihood

Alternatively, we could start the recursions at $t = p + 1$ without an initial condition on $\{y_t\}_{t=-p+1}^{0}$.

1. Condition on $\varepsilon_p = \varepsilon_{p-1} = \ldots = \varepsilon_{p-q+1} = 0$.

2. Recursively compute $\{\varepsilon_t\}_{t=p+1}^{T}$ using $\{y_t\}_{t=1}^{T}$ and $\{\varepsilon_t\}_{t=p-q+1}^{0}$.

3. Compute the log likelihood as

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{y}_T, \boldsymbol{\varepsilon}_0 = \mathbf{0}) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}\varepsilon_t^2.$$

Numerical Optimization

Numerical Maximum Likelihood

Given, $\boldsymbol{\theta}$ and $\boldsymbol{y}$, suppose we can compute the value of a

likelihood or log likelihood.

- Likelhood optimization is often very challenging.
- We may not be able to obtain analytical expressions for the MLEs, $\hat{\boldsymbol{\theta}}$.
- Numerical optimization techniques will often help us find an approximate (not exact) MLE.
- We will need to set a tolerance level for the quality of our approximation.

Grid Search

Let $\boldsymbol{\theta} \in \mathrm{R}^k$.

- We can define a univariate grid of m_i points $\theta_i \in \Theta^{(i)} = \{\theta_{i,1}, \dots, \theta_{i,m_i}\}$ for $i = 1, \dots, k$.
- Define $\Theta = \Theta^{(1)} \otimes \Theta^{(2)} \otimes \cdots \otimes \Theta^{(k)}$, which is the cartesian product of the k univariate grids.
- Often such grids are equally spaced, but this is certainly not required.
- Optimal location of grid points is an extremely important way to improve numerical efficiency.

Grid Search

To implement grid search, we simply evaluate the likelihood at each value in Θ.

- Each value in Θ defines a set of candidate parameter values.
- The approximated MLE is the point that achieves the highest likelihood or log likelihood value.

Grid search is ineffective for large k because the number of grid points in Θ grows exponentially.

- Doubling the number of points (for more accuracy) in each dimension results in 2^k extra points.

- This is called the curse of dimensionality.

$AR(1)$ Grid Search

Suppose $c = 0$ and $\sigma^2 = 1$.

- In this case, $\boldsymbol{\theta} = \phi$ and $k = 1$.

- Under stationarity, we know $|\phi| < 1$, so we might define an equally-spaced grid of values $\{-0.99, -0.98, \ldots, 0.98, 0.99\}$.

- Given data $\boldsymbol{y}$, we can compute the exact or conditional likelihood for each ϕ_i in the grid.

- The ϕ_i that results in the highest likelihood value is the approximate MLE, which we denote ϕ^*.

- We can iteratively refine the grid around ϕ^* until our tolerance is reached.

Binary Search

Binary search is an optimization method that is far more efficient than grid search, for *univariate* problems.

- It can only be used if the criterion function is concave.

- The algorithm is

1. Pick two adjacent points θ_j and θ_{j+1} in the middle of the grid and evaluate the likelihood.

2. If $\mathcal{L}(\theta_{j+1}) < \mathcal{L}(\theta_j)$, set the upper bound of the grid to be θ_{j+1} and otherwise set the lower bound to be θ_j.

3. If the lower and upper bounds are separated by more than one grid point, return to step 1. Otherwise, stop.

- Golden search is similar to binary search. See Heer and Maussner (2009) for details.

Newton's Method

Newton's method is an iterative root finding algorithm, that uses derivative/gradient information:

$$x^{(i+1)} = x^{(i)} - f(x^{(i)})/f'(x^{(i)}).$$

The value $x^{(n)}$ for large n is an approximation of the function root, $x : f(x) = 0$.

Newton-Raphson

Newton's method can also be used for optimization (not just root finding).

- Optimization is the same as root finding for the derivative function.
- The Newton-Raphson algorithm is

$$x^{(i+1)} = x^{(i)} - f'(x^{(i)})/f''(x^{(i)}).$$

Newton-Raphson

Define

$$\boldsymbol{g}(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$
$$\mathcal{H}(\boldsymbol{\theta}) = \nabla^2 \ell(\boldsymbol{\theta}) = \nabla \boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2},$$

where $\mathcal{H}(\boldsymbol{\theta})$ is positive definite:

$$\boldsymbol{x}^T \mathcal{H}(\boldsymbol{\theta}) \boldsymbol{x} > 0 \quad \forall \boldsymbol{x} \in \mathrm{R}^k.$$

Newton-Raphson

We approximate $\ell(\boldsymbol{\theta})$ with a second-order Taylor expansion around $\boldsymbol{\theta}^{(0)}$:

$$\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}^{(0)}) + \boldsymbol{g}(\boldsymbol{\theta}^{(0)})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^T \mathcal{H}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}).$$

The Newton-Raphson method chooses $\boldsymbol{\theta}^{(1)}$ to maximize $\tilde{\ell}(\boldsymbol{\theta})$:

$$\nabla \tilde{\ell}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}} = \boldsymbol{g}(\boldsymbol{\theta}^{(0)}) + \mathcal{H}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}) = 0.$$
$$\Rightarrow \quad \boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \mathcal{H}(\boldsymbol{\theta}^{(0)})^{-1} \boldsymbol{g}(\boldsymbol{\theta}^{(0)}).$$

Newton-Raphson

The Newton-Raphson method begins with $\boldsymbol{\theta}^{(0)}$ and iteratively computes

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathcal{H}(\boldsymbol{\theta}^{(i)})^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(i)})$$

until $||\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}|| < \tau$, where τ is some tolerance level.

Newton-Raphson

- Newton-Raphson converges fast if the likelihood is concave and the initial guess is good enough.

- A modified version of Newton-Raphson computes:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - s\mathcal{H}(\boldsymbol{\theta}^{(i)})^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(i)})$$

for various values of s at each iteration and chooses $\boldsymbol{\theta}^{(i+1)}$ that yields the largest likelihood value.

Quasi Newton-Raphson

Various modified Newton-Raphson methods have been proposed which substitute other positive definite matrices for $\mathcal{H}(\boldsymbol{\theta}^{(i)})^{-1}$.

- These are useful if $\mathcal{H}(\boldsymbol{\theta}^{(i)})^{-1}$ is not possible to compute or invert.

- Typically these are slower but more robust.

Numerical Differentiation

If analytical derivatives are not possible, numerical derivatives are an option.

- The i th element of $\boldsymbol{g}(\boldsymbol{\theta})$ can be approximated with:

$$g_i(\boldsymbol{\theta}) = \frac{1}{\Delta}\Big(\ell(\theta_1, \ldots, \theta_i + \Delta, \ldots, \theta_k) - \ell(\theta_1, \ldots, \theta_i, \ldots, \theta_k)\Big),$$

for some small Δ.

- The hessian can be computed numerically from $\boldsymbol{g}(\boldsymbol{\theta})$ in a similar manner.

Forecasting ARMA Models

Forecasting with Infinite Data

Consider an $ARMA$ process with $MA(\infty)$ representation:

$$Y_t - \mu = \psi(L)\varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma^2)$$

where

$$\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$$
$$\sum_{j=0}^{\infty} |\psi_j| < \infty$$
$$\psi_0 = 1.$$

Forecasting with Infinite Data

Suppose

- we observe an infinite history of $\{\varepsilon_t\}$ up to date t: $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$.
- we know the MA parameters μ, σ, $\{\psi_j\}_{j=0}^{\infty}$.

Then

$$Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1} + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \ldots$$

Optimal Forecast

The optimal forecast of Y_{t+s} in terms of MSE is:

$$E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \ldots$$

Note, this is different from

$$Y_t = \mu + \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \ldots$$

Forecast Error

The forecast error is:

$$\begin{aligned} Y_{t+s} - E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots] & \\ &= \mu + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \ldots + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t+1} + \ldots \\ &\qquad - \mu - \psi_s\varepsilon_t - \psi_{s+1}\varepsilon_{t-1} - \ldots \\ &= \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1} \end{aligned}$$

Since $E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots]$ is linear in $\{\varepsilon_\tau\}_{\tau=-\infty}^{t}$ it is both the optimal forecast and optimal linear forecast.

Forecast as Linear Projection

Hamilton refers to optimal linear forecasts as $\hat{E}[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots]$.

- In this case

$$\begin{aligned} E[Y_{t+s} | \varepsilon_t, \ldots] &= \hat{E}[Y_{t+s} | \varepsilon_t, \ldots] \\ \implies Y^*_{t+s|t} &= \hat{Y}_{t+s|t} \end{aligned}$$

which is also a linear projection $\hat{p}(Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots)$.

- Clearly, the linear projection condition is satisfied for $j = t, t-1, \ldots$

$$\begin{aligned} E[(Y_{t+s} - E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots])\varepsilon_j] & \\ &= E[(\varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1})\varepsilon_j] = 0. \end{aligned}$$

Forecast MSE

The forecast MSE is:

$$\begin{aligned} E[(Y_{t+s} - E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \ldots])^2] & \\ &= E[(\varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1})^2] \\ &= \sigma^2 \sum_{j=0}^{s-1} \psi_j^2. \end{aligned}$$

Processing math: 100%

Forecasting Conditional on Lagged Y_t

Suppose we don’t observe the full history of ε_t.

- Instead, we observe the full history of $y_t: y_t, y_{t-1}, y_{t-2}, \dots$.
- We have an $ARMA$ process with the same MA representation as before.

If the $MA(\infty)$ representation is invertible, we can write it as an $AR(\infty)$:

$$\eta(L)(Y_t - \mu) = \varepsilon_t,$$

where $\eta(L) = \psi^{-1}(L)$.

Computing Historical Values

The history of ε_t can be constructed with the history of y_t:

$$\begin{aligned} \varepsilon_t &= \eta(L)(y_t - \mu) \\ \varepsilon_{t-1} &= \eta(L)(y_{t-1} - \mu) \\ \varepsilon_{t-2} &= \eta(L)(y_{t-2} - \mu) \\ &\vdots \\ \Rightarrow \quad E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \dots] &= E[Y_{t+s} | y_t, y_{t-1}, \dots] \\ &= \mu + (\psi_s + \psi_{s+1}L + \psi_{s+2}L^2 + \dots)\varepsilon_t \\ &= \mu + (\psi_s + \psi_{s+1}L + \psi_{s+2}L^2 + \dots)\eta(L)(y_t - \mu). \end{aligned}$$

Example: $AR(1)$

For an $AR(1)$ with $|\phi| < 1$:

$$Y_t - \mu = \psi(L)\varepsilon_t,$$

where

$$\psi(L) = (1 + \phi L + \phi^2 L^2 + \dots) = (1 + \psi_1 L + \psi_2 L^2 + \dots).$$

Example: $AR(1)$

The optimal forecast s -periods ahead is

$$\begin{aligned}E[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] &= \mu + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \ldots \\ &= \mu + \phi^s\varepsilon_t + \phi^{s+1}\varepsilon_{t-1} + \phi^{s+2}\varepsilon_{t-2} + \ldots \\ &= \mu + \phi^s(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \ldots) \\ &= \mu + \phi^s(y_t - \mu)\end{aligned}$$

- The forecast decays toward μ as s increases.
- The MSE is $\sigma^2\sum_{j=0}^{s-1}\phi^{2j}$.
- As $s \to \infty, MSE \to \frac{\sigma^2}{1-\phi^2} = Var(Y_t)$.

Forecasting with Finite Data

In reality, we don't observe an infinite history of $y_t, y_{t-1}, y_{t-2}, \ldots$.

- Suppose we have only a finite set of m past observations of $y_t : y_t, y_{t-1}, \ldots, y_{t-m+1}$.
- The optimal $AR(p)$ forecast only makes use of the past p observations if available (i.e. $p < m$).
- If we want to forecast an MA or $ARMA$ (of arbitrary order), we need an infinite history to construct an optimal forecast.

Approximate Optimal Forecasts

Start by setting all ε‘s prior to time $t - m + 1$ equal to zero.

$$E[Y_{t+s}|y_t, y_{t-1}, \ldots] \approx E[Y_{t+s}|y_t, y_{t-1}, \ldots, y_{t-m+1}, \varepsilon_{t-m} = 0, \varepsilon_{t-m-1} = 0, \ldots].$$

Example $MA(q)$

Start with

$$\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \ldots = \hat{\varepsilon}_{t-m-q+1} = 0.$$

Calculate forward recursively

$$\begin{aligned}\hat{\varepsilon}_{t-m+1} &= (y_{t-m+1} - \mu) \\ \hat{\varepsilon}_{t-m+2} &= (y_{t-m+2} - \mu) - \theta_1\hat{\varepsilon}_{t-m+1} \\ \hat{\varepsilon}_{t-m+3} &= (y_{t-m+3} - \mu) - \theta_1\hat{\varepsilon}_{t-m+2} - \theta_2\hat{\varepsilon}_{t-m+1} \\ &\vdots\end{aligned}$$

Example $MA(q)$

With $\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-m+1}$ in hand we can compute forecasts

$$\hat{Y}_{t+s} = \theta_s \hat{\varepsilon}_t + \theta_{s+1} \hat{\varepsilon}_{t-1} + \ldots + \theta_q \hat{\varepsilon}_{t-q+s}.$$

Exact Finite Sample Forecasts

Another forecast approximation method is to simply project $Y_{t+1} - \mu$ on $\boldsymbol{X}_t = (Y_t - \mu, Y_{t-1} - \mu, \ldots, Y_{t-m+1} - \mu)^T$.

That is

$$\begin{aligned} \hat{Y}^{(m)}_{t+1|t} - \mu &= \boldsymbol{X}_t^{'} \boldsymbol{\beta}^{(m)} \\ &= \beta_1^{(m)}(Y_t - \mu) + \beta_2^{(m)}(Y_{t-1} - \mu) + \ldots + \beta_m^{(m)}(Y_{t-m+1} - \mu). \end{aligned}$$

Exact Finite Sample Forecasts

$$\boldsymbol{\beta}^{(m)} = E[\boldsymbol{X}_t \boldsymbol{X}_t^{'}]^{-1} E[\boldsymbol{X}_t (Y_{t+1} - \mu)] = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{m-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{m-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{m-1} & \cdots & \cdots & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}.$$

Exact Finite Sample Forecasts

Similarly,

$$\begin{aligned} Y^{(m)}_{t+s|t} - \mu &= \boldsymbol{X}_t^{'} \boldsymbol{\beta}^{(m,s)} \\ &= \beta_1^{(m,s)}(Y_t - \mu) + \beta_2^{(m,s)}(Y_{t-1} - \mu) + \ldots + \beta_m^{(m,s)}(Y_{t-m+1} - \mu). \end{aligned}$$

Exact Finite Sample Forecasts

$$\boldsymbol{\beta}^{(m,s)} = E[\boldsymbol{X}_t\boldsymbol{X}_t^{'}]^{-1}E[\boldsymbol{X}_t(Y_{t+s} - \mu)]$$

$$= \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{m-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{m-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{m-1} & \cdots & \cdots & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{bmatrix}.$$

Example $AMRA(1, 1)$

Let $\{Y_t\}$ be an $ARMA(1, 1)$ process with $|\phi| < 1$ and $|\theta| < 1$ (causal and invertible). Then:

$$\begin{aligned} (1 - \phi L)(Y_t - \mu) &= (1 + \theta L)\varepsilon_t \\ \implies Y_t - \mu &= \psi(L)\varepsilon_t \end{aligned}$$

where $\psi(L) = (1 - \phi L)^{-1}(1 + \theta L)$.

- We can also write

$$\varepsilon_t = (1 + \theta L)^{-1}(1 - \phi L)(Y_t - \mu) = \psi(L)^{-1}(Y_t - \mu).$$

Example $AMRA(1, 1)$

Expanding the MA representation

$$\begin{aligned} \psi(L) &= (1 + \phi L + \phi^2 L^2 + \ldots)(1 + \theta L) \\ &= 1 + (\phi + \theta)L + (\phi^2 + \phi\theta)L^2 + (\phi^3 + \phi^2\theta)L^3 + \ldots \\ &= 1 + \sum_{j=1}^{\infty}(\phi^j + \phi^{j-1}\theta)L^j \\ \implies \psi_m &= \phi^m + \phi^{m-1}\theta. \end{aligned}$$

Example $AMRA(1, 1)$

Let’s define $\psi_s(L)$ as the polynomial

$$\psi_s(L) = \psi_s + \psi_{s+1}L + \psi_{s+2}L^2 + \ldots$$

This is different from $\psi_s L^s + \psi_{s+1}L^{s+1} + \ldots$

Example $AMRA(1, 1)$

For the $ARMA(1, 1)$,

$$\begin{aligned}\psi_s(L) &= (\phi^s + \phi^{s-1}\theta) + (\phi^{s+1} + \phi^s\theta)L + (\phi^{s+2} + \phi^{s+1}\theta)L^2 + \dots \\ &= \sum_{j=s}^{\infty}(\phi^j + \phi^{j-1}\theta)L^{j-s} \\ &= (\phi^s + \phi^{s-1}\theta)\sum_{j=0}^{\infty}\phi^j L^j \\ &= (\phi^s + \phi^{s-1}\theta)(1 - \phi L)^{-1}.\end{aligned}$$

Example $AMRA(1, 1)$

Recall, for an $MA(\infty)$, the optimal forecast is

$$\begin{aligned}\hat{Y}_{t+s|t} - \mu &= E[Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \dots] \\ &= \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \dots = \psi_s(L)\varepsilon_t\end{aligned}$$

So, for the $ARMA(1, 1)$.

$$\begin{aligned}\hat{Y}_{t+s|t} - \mu &= (\phi^s + \phi^{s-1}\theta)(1 - \phi L)^{-1}\varepsilon_t \\ &= (\phi^s + \phi^{s-1}\theta)(1 - \phi L)^{-1}(1 - \phi L)(1 + \theta L)^{-1}(Y_t - \mu) \\ &= (\phi^s + \phi^{s-1}\theta)(1 + \theta L)^{-1}(Y_t - \mu).\end{aligned}$$

Example $AMRA(1, 1)$

Notice

$$\begin{aligned}\hat{Y}_{t+s|t} - \mu &= (\phi^s + \phi^{s-1}\theta)(1 + \theta L)^{-1}(Y_t - \mu) \\ &= \phi(\phi^{s-1} + \phi^{s-2}\theta)(1 + \theta L)^{-1}(Y_t - \mu) \\ &= \phi(\hat{Y}_{t+s-1|t} - \mu), \quad \text{if } s \geq 2,\end{aligned}$$

which means the forecast decays toward μ.

Example $AMRA(1, 1)$

For $s = 1$,

$$\begin{aligned}
\hat{Y}_{t+s|t} - \mu &= (\phi + \theta)(1 + \theta L)^{-1}(Y_t - \mu) \\
&= (\phi + \phi\theta L - \phi\theta L + \theta)(1 + \theta L)^{-1}(Y_t - \mu) \\
&= [\phi(1 + \theta L) + \theta(1 - \phi L)](1 + \theta L)^{-1}(Y_t - \mu) \\
&= \phi(Y_t - \mu) + \theta(1 - \phi L)(1 + \theta L)^{-1}(Y_t - \mu) \\
&= \phi(Y_t - \mu) + \theta\varepsilon_t.
\end{aligned}$$

Linear Predictors

Forecasting

Suppose we are interested in forecasting a random variable Y_{t+1} based on a set of variables $\boldsymbol{X}_t$.

- $\boldsymbol{X}_t$ might be comprised of m lags of Y_{t+1}: $Y_t, Y_{t-1}, \ldots, Y_{t-m+1}$.
- We can denote $Y^*_{t+1|t}$ as the forecast of Y_{t+1} based on $\boldsymbol{X}_t$.
- We can choose $Y^*_{t+1|t}$ to minimize some loss function, $L\left(Y^*_{t+1|t}\right)$, which evaluates the quality of $Y^*_{t+1|t}$.
- A common choice is the quadratic loss function:

$$L\left(Y^*_{t+1|t}\right) = \mathrm{E}\left[\left(Y_{t+1} - Y^*_{t+1|t}\right)^2\right].$$

Mean Squared Error Loss

Quadratic loss is also known as *mean squared error*.

$$MSE\left(Y^*_{t+1|t}\right) = \mathrm{E}\left[\left(Y_{t+1} - Y^*_{t+1|t}\right)^2\right].$$

- The conditional expectation, $\mathrm{E}\left[Y_{t+1}|\boldsymbol{X}_t\right]$ minimizes $MSE\left(Y^*_{t+1|t}\right)$.

MSE Minimizer

Let $Y^*_{t+1|t} = g(\boldsymbol{X}_t)$. Then

$$
\begin{aligned}
\mathrm{E}\left[\left(Y_{t+1} - g(\boldsymbol{X}_t)\right)^2\right] &= \mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right.\right. \\
&\qquad \left.\left. + \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)^2\right] \\
&= \mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)^2\right] \\
&\quad + 2\mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)\right. \\
&\qquad \left. \times \left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)\right] \\
&\qquad + \mathrm{E}\left[\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)^2\right]
\end{aligned}
$$

MSE Minimizer

By the law of iterated expectations

$$
\begin{aligned}
&\mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)\right] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)|\boldsymbol{X}_t\right]\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)\right] \\
&\quad = \mathrm{E}\left[\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)\right] \\
&\quad = 0.
\end{aligned}
$$

- This means that the second term of the equation on the previous slide is zero.

MSE Minimizer

Substituting the previous result:

$$
\begin{aligned}
\mathrm{E}\left[\left(Y_{t+1} - g(\boldsymbol{X}_t)\right)^2\right] &= \mathrm{E}\left[\left(Y_{t+1} - \mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right)^2\right] \\
&\quad + \mathrm{E}\left[\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)^2\right]
\end{aligned}
$$

- Clearly the the MSE is minimized when

$$
\mathrm{E}\left[\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] - g(\boldsymbol{X}_t)\right)^2\right] = 0.
$$

- This occurs when $\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t] = g(\boldsymbol{X}_t)$.

Linear Projection

We can restrict our forecast to be a linear function of $\boldsymbol{X}_t$:

$$Y^*_{t+1|t} = \boldsymbol{X}_t^{'}\boldsymbol{\beta}.$$

- Let $\boldsymbol{\beta}^*$ be the value of $\boldsymbol{\beta}$ so that the forecast error is *orthogonal* to or *uncorrelated* with $\boldsymbol{X}_t$:

$$\mathrm{E}\left[\boldsymbol{X}_t\left(\underbrace{Y_{t+1} - \boldsymbol{X}_t^{'}\boldsymbol{\beta}^*}_{\text{forecast error}}\right)\right] = \boldsymbol{0}.$$

- This is a *system* of equations.

- $\boldsymbol{\beta}^*$ minimizes the MSE.

Linear Projection

We can use the same steps as before to show that $\boldsymbol{\beta}^*$ minimizes MSE.

- Begin with an arbitrary linear forecasting rule, $Y^*_{t+1|t} = \boldsymbol{X}_t^{'}\boldsymbol{\gamma}$.

- Show that

$$\begin{aligned} MSE\left(Y^*_{t+1|t}\right) &= \mathrm{E}\left[\left(Y_{t+1} - \boldsymbol{X}_t^{'}\boldsymbol{\gamma}\right)^2\right] \\ &= \mathrm{E}\left[\left(Y_{t+1} - \boldsymbol{X}_t^{'}\boldsymbol{\beta}^* + \boldsymbol{X}_t^{'}\boldsymbol{\beta}^* - \boldsymbol{X}_t^{'}\boldsymbol{\gamma}\right)^2\right] \\ &= \mathrm{E}\left[(Y_{t+1} - \boldsymbol{X}_t^{'}\boldsymbol{\beta}^*)^2\right] + \mathrm{E}\left[\left(\boldsymbol{X}_t^{'}\boldsymbol{\beta}^* - \boldsymbol{X}_t^{'}\boldsymbol{\gamma}\right)^2\right]. \end{aligned}$$

- Hence, MSE is minimized when $\boldsymbol{\gamma} = \boldsymbol{\beta}^*$.

Linear Projection

$Y^*_{t+1|t} = \boldsymbol{X}_t^{'}\boldsymbol{\beta}^*$ is referred to as the *linear projection* of Y_{t+1} on $\boldsymbol{X}_t$.

- We will denote the linear projection as

$$\hat{P}(Y_{t+1}|\boldsymbol{X}_t) = \boldsymbol{X}_t^{'}\boldsymbol{\beta}^* \quad \text{or} \quad \hat{Y}_{t+1|t} = \boldsymbol{X}_t^{'}\boldsymbol{\beta}^*.$$

- Clearly

$$MSE\left(\hat{P}(Y_{t+1}|\boldsymbol{X}_t)\right) \geq MSE\left(\mathrm{E}[Y_{t+1}|\boldsymbol{X}_t]\right).$$

Linear Projection Solution

Using the orthogonality condition:

$$\boldsymbol{\beta}^* = \mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}\left[\boldsymbol{X}_tY_{t+1}\right].$$

- Least squares projection is the sample analogue of the equation above.

Linear Projection MSE

Using our solution for $\boldsymbol{\beta}^*$, we can solve for the MSE of the linear projection:

$$\begin{aligned}
MSE(Y_{t+1|t}^*) &= \mathrm{E}\left[\left(Y_{t+1} - \boldsymbol{X}_t'\boldsymbol{\beta}^*\right)^2\right] \\
&= \mathrm{E}[Y_{t+1}^2] - 2\mathrm{E}[Y_{t+1}\boldsymbol{X}_t'\boldsymbol{\beta}^*] + \mathrm{E}[\boldsymbol{\beta}^{*'}\boldsymbol{X}_t\boldsymbol{X}_t'\boldsymbol{\beta}^*] \\
&= \mathrm{E}[Y_{t+1}^2] - 2\mathrm{E}[Y_{t+1}\boldsymbol{X}_t']\mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}\left[\boldsymbol{X}_tY_{t+1}\right] \\
&\qquad + \mathrm{E}\left[Y_{t+1}\boldsymbol{X}_t'\right]\mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t'] \\
&\qquad\qquad \times \mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}\left[\boldsymbol{X}_tY_{t+1}\right] \\
&= \mathrm{E}[Y_{t+1}^2] - \mathrm{E}[Y_{t+1}\boldsymbol{X}_t']\mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}\left[\boldsymbol{X}_tY_{t+1}\right].
\end{aligned}$$

Vector Linear Projection

Let $\boldsymbol{Y}_{t+1}$ be an $(n \times 1)$ vector and $\boldsymbol{X}_t$ an $(m \times 1)$ vector.

- The linear projection of $\boldsymbol{Y}_{t+1}$ on $\boldsymbol{X}_t$ is

$$\hat{P}(\boldsymbol{Y}_{t+1}'|\boldsymbol{X}_t) = \hat{\boldsymbol{Y}}_{t+1|t}' = \boldsymbol{X}_t'\boldsymbol{\beta}^*.$$

where $\boldsymbol{\beta}^*$ is the $(m \times n)$ matrix such that

$$\mathrm{E}\left[\boldsymbol{X}_t\left(\boldsymbol{Y}_{t+1}' - \boldsymbol{X}_t'\boldsymbol{\beta}^*\right)\right] = \boldsymbol{0}.$$

- As in the univariate case

$$\boldsymbol{\beta}^* = \mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{X}_t'\right]^{-1}\mathrm{E}\left[\boldsymbol{X}_t\boldsymbol{Y}_{t+1}'\right].$$

Vector Autoregression

Definition

A p th order vector autoregression generalizes a scalar $AR(p)$:

$$\boldsymbol{Y}_t = \boldsymbol{c} + \Phi_1 \boldsymbol{Y}_{t-1} + \Phi_2 \boldsymbol{Y}_{t-2} + \ldots + \ + \Phi_p \boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

- $Y_t = (Y_{1t}, Y_{2t}, \ldots, Y_{nt})^{'}$ is an $n \times 1$ vector of random variables.
- $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^{'}$ is an $n \times 1$ vector of constants.
- Φ_j is an $n \times n$ matrix of autoregressive coefficients for $j = 1, \ldots, p$.
- $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})^{'}$ is a vector white noise process:

$$E[\boldsymbol{\varepsilon}_t] = \boldsymbol{0} \ \text{ and } \ E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_\tau^{'}] = \begin{cases} \Omega & t = \tau \\ 0 & \text{o/w} \end{cases}.$$

- $\boldsymbol{Y}_t$ is referred to as a $VAR(p)$.

$AR(p)$ as $VAR(1)$

Recall, an $AR(p)$ can be written as a $VAR(1)$

$$\boldsymbol{Y}_t = \Phi \boldsymbol{Y}_{t-1} + \boldsymbol{v}_t$$

where

$$\boldsymbol{Y}_t = \begin{bmatrix} Y_t \\ \vdots \\ Y_{t-p+1} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Lag Operator Notation

In lag operator notation,

$$\begin{aligned}\Phi(L)\boldsymbol{Y}_t &= \left[I_n - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p\right]\boldsymbol{Y}_t \\ &= \boldsymbol{c} + \boldsymbol{\varepsilon}_t.\end{aligned}$$

- $\Phi(L)$ is a matrix where each component is a scalar lag polynomial.

Weak Stationarity

The concept of weak stationarity is unchanged: $\boldsymbol{Y}_t$ is weakly stationary if

$$E[\boldsymbol{Y}_t] \text{ and } E[\boldsymbol{Y}_t\boldsymbol{Y}_{t-j}^{'}]$$

are independent of $t \ \forall \ j$

Mean

By weak stationarity,

$$\begin{aligned}E[\boldsymbol{Y}_t] = \boldsymbol{\mu} &= \boldsymbol{c} + \Phi_1\boldsymbol{\mu} + \ldots + \Phi_p\boldsymbol{\mu} \\ \implies \boldsymbol{\mu} &= [I_n - \Phi_1 - \ldots - \Phi_p]^{-1}\boldsymbol{c}\end{aligned}$$

Note that

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{'} \neq (\mu, \mu, \ldots, \mu)^{'}$$

Alternatively, we can re-express as a zero-mean process:

$$(\boldsymbol{Y}_t - \boldsymbol{\mu}) = \Phi_1(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}) + \ldots + \Phi_p(\boldsymbol{Y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t.$$

$VAR(p)$ as $VAR(1)$

We can write a $VAR(p)$ as a $VAR(1)$:

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

where

$$\xi_t = \begin{bmatrix} \boldsymbol{Y}_t - \boldsymbol{\mu} \\ \boldsymbol{Y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{Y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}, \quad F = \begin{bmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & 0 & \ldots & 0 \\ 0 & I_n & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_n & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$VAR(p)$ as $VAR(1)$

Clearly,

$$\Omega \quad t = \tau$$

$$E[\boldsymbol{v}_t\boldsymbol{v}_\tau] = \begin{cases} 0 & \text{o/w} \end{cases}$$

where

$$Q = \begin{bmatrix} \Omega & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}_{np \times np}.$$

Recursive Iteration

Recursively iterating on the $VAR(1)$:

$$\xi_{t+s} = \boldsymbol{v}_{t+s} + F\boldsymbol{v}_{t+s-1} + F^2\boldsymbol{v}_{t+s-2} + \dots + F^{s-1}\boldsymbol{v}_{t+1} + F^s\xi_t.$$

Assuming F is nonsingular, it can be decomposed as

$$F = T\Lambda T^{-1}.$$

- Λ is a diagonal matrix comprised of the np eigenvalues of F.
- T is a matrix of eigenvectors as columns.

Recursive Iteration

Substituting the decomposition,

$$F^2 = FF = (T\Lambda T^{-1})\, T\Lambda T^{-1} = T\Lambda^2 T^{-1}$$
$$\implies F^s = T\Lambda^s T^{-1} \to 0 \quad \text{if} \quad |\lambda_k| < 1 \quad \text{for} \quad k = 1, \dots, np.$$

If $F^s \to 0$ as $s \to \infty$,

- The effect of $\boldsymbol{\varepsilon}_t$ on ξ_{t+s} dies out as $s \to \infty$.
- ξ_t (and hence $\boldsymbol{Y}_t$) is stationary and causal.
- Alternatively, $\boldsymbol{Y}_t$ is stationary and causal if the roots of $[I_n - \Phi_1 z - \Phi_2 z^2 - \dots - \Phi_p z^p]$ all lie outside the unit circle.

Vector $MA(\infty)$ Representation

If $F^s \to 0$ as $s \to \infty$, then

$$\xi_{t+s} = \boldsymbol{v}_{t+s} + F\boldsymbol{v}_{t+s-1} + F^2\boldsymbol{v}_{t+s-2} + F^3\boldsymbol{v}_{t+s-3} + \dots$$

which is a vector $MA(\infty)$ process.

Vector $MA(\infty)$ Representation

We can also write $\boldsymbol{Y}_t$ alone as a vector $MA(\infty)$. First, recognize

$$\begin{aligned}\boldsymbol{Y}_{t+s} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t+s} + \Psi_1\boldsymbol{\varepsilon}_{t+s-1} + \Psi_2\boldsymbol{\varepsilon}_{t+s-2} + \ldots + \Psi_{s-1}\boldsymbol{\varepsilon}_{t+1} \\ + F_{11}^{(s)}(\boldsymbol{Y}_t - \boldsymbol{\mu}) + F_{12}^{(s)}(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}) + \ldots + F_{1p}^{(s)}(\boldsymbol{Y}_{t-p+1} - \boldsymbol{\mu}).\end{aligned}$$

- $\Psi_j = F_{11}^{(j)}$.
- $F_{1k}^{(j)}$ is comprised of rows 1 to n and columns $(k-1)n+1$ to kn of matrix F^j.
- Note that the matrices $(F \times F)[1:n, 1:n]$ and $F[1:n, 1:n] \times F[1:n, 1:n]$ are not the same.

Vector $MA(\infty)$ Representation

Suppose all eigenvalues of F are inside the unit circle.

- Then $F^s \to 0$ as $s \to \infty$.
- This means $F_{1k}^{(s)} \to 0$ as $s \to \infty$.
- In the limit

$$\begin{aligned}\boldsymbol{Y}_{t+s} &= \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t+s} + \Psi_1\boldsymbol{\varepsilon}_{t+s-1} + \Psi_2\boldsymbol{\varepsilon}_{t+s-2} + \ldots \\ &= \boldsymbol{\mu} + \Psi(L)\boldsymbol{\varepsilon}_{t+s}.\end{aligned}$$

Inverse of $MA(\infty)$ Lag Polynomial

In this case $\Psi(L) = \Phi(L)^{-1}$ or

$$[1 - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p][1 + \Psi_1 L + \Psi_2 L^2 + \ldots] = I_n.$$

Representation with Uncorrelated Noise

We can always write a stationary and causal $VAR(p)$ as a vector $MA(\infty)$ with a mutually uncorrelated white noise vector.

- Define $\boldsymbol{u}_t = H\boldsymbol{\varepsilon}_t$ such that $H\Omega H' = D$.
- Then

$$\begin{aligned}\boldsymbol{Y}_t &= \boldsymbol{\mu} + H^{-1}H\boldsymbol{\varepsilon}_t + \Psi_1(H^{-1}H)\boldsymbol{\varepsilon}_{t-1} + \dots \\ &= \boldsymbol{\mu} + J_0\boldsymbol{u}_t + J_1\boldsymbol{u}_{t-1} + J_2\boldsymbol{u}_{t-2} + \dots\end{aligned}$$

where $J_s = \Psi_s H^{-1}$.

Representation with Uncorrelated Noise

- In this case the leading matrix $J_0 \neq I_n$.
- The noise vector is uncorrelated:

$$\begin{aligned}E[\boldsymbol{u}_t\boldsymbol{u}_t^{'}] &= E[H\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^{'}H^{'}] \\ &= HE[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t]H^{'} \\ &= H\Omega H^{'} \\ &= D.\end{aligned}$$

Autocovariances of Vector Processes

Vector Autocovariance

Given an n -dimensional, weakly stationary vector process, $\boldsymbol{Y}_t$, the j th autocovariance matix is defined as:

$$\Gamma_{j,t} = E[(\boldsymbol{Y}_t - \boldsymbol{\mu})(\boldsymbol{Y}_{t-j} - \boldsymbol{\mu})^{'}].$$

Since $Y_{1,t}$ is different from $Y_{2,t}$, $\Gamma_j \neq \Gamma_{-j}$:

$$\Gamma_j(1,2) = Cov(Y_{1,t}, Y_{2,t-j}) \neq Cov(Y_{1,t}, Y_{2,t+j}) = \Gamma_{-j}(1,2).$$

Vector Autocovariance

It is true that $\Gamma_j = \Gamma_{-j}^{'}$:

$$\Gamma_j(1,2) = Cov(Y_{1,t}, Y_{2,t-j}) = Cov(Y_{2,t}, Y_{1,t+j}) = \Gamma_{-j}(2,1).$$

- Stationarity does impose $Cov(Y_{1,t}, Y_{2,t-j}) = Cov(Y_{1,t+j}, Y_{2,t})$.

Vector MA(q) Process

A vector moving average process of order q is

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \Theta_1\boldsymbol{\varepsilon}_{t-1} + \Theta_2\boldsymbol{\varepsilon}_{t-2} + \ldots + \Theta_q\boldsymbol{\varepsilon}_{t-q}, \quad \boldsymbol{\varepsilon}_t \overset{i.i.d}{\sim} WN(\mathbf{0}, \Omega),$$

where Θ_j is an $N \times N$ matrix of MA coefficients for $j = 1, \ldots, q$.

- We can define $\Theta_0 = I_n$.

- Clearly $E[\boldsymbol{Y}_t] = \boldsymbol{\mu} \ \ \forall \, t$.

Vector MA(q) Autocovariances

The jth autocovariance matrix is:

$$\begin{aligned}\Gamma_j &= E[(\boldsymbol{Y}_t - \boldsymbol{\mu})(\boldsymbol{Y}_{t-j} - \boldsymbol{\mu})^{'}] \\ &= E[(\Theta_0 \boldsymbol{\varepsilon}_t + \Theta_1 \boldsymbol{\varepsilon}_{t-1} + \ldots + \Theta_q \boldsymbol{\varepsilon}_{t-q}) \\ &\qquad \times (\Theta_0 \boldsymbol{\varepsilon}_{t-j} + \Theta_1 \boldsymbol{\varepsilon}_{t-j-1} + \ldots + \Theta_q \boldsymbol{\varepsilon}_{t-j-q})^{'}]\end{aligned}$$

Vector MA(q) Autocovariances

- For $|j| > q : \Gamma_j = \mathbf{0}_{N \times N}$.
- For $j = 0:$

$$\begin{aligned}\Gamma_0 &= \Theta_0 \Omega \Theta_0^{'} + \Theta_1 \Omega \Theta_1^{'} + \ldots + \Theta_q \Omega \Theta_q^{'} \\ &= \sum_{i=0}^{q} \Theta_i \Omega \Theta_i^{'}.\end{aligned}$$

- For $j = 1, \ldots, q$:

$$\begin{aligned}\Gamma_j &= \Theta_j \Omega \Theta_0^{'} + \Theta_{j+1} \Omega \Theta_1^{'} + \ldots + \Theta_q \Omega \Theta_{q-j}^{'} \\ &= \sum_{i=0}^{q-j} \Theta_{j+i} \Omega \Theta_i^{'}.\end{aligned}$$

Vector MA(q) Autocovariances

- For $j = -1, \ldots, -q$:

$$\begin{aligned}\Gamma_j &= \Theta_0 \Omega \Theta_{-j}^{'} + \Theta_1 \Omega \Theta_{-j+1}^{'} + \ldots + \Theta_{q+j} \Omega \Theta_q^{'} \\ &= \sum_{i=0}^{q+j} \Theta_i \Omega \Theta_{-j+i}^{'}.\end{aligned}$$

- $\Gamma_j^{'} = \Gamma_{-j}$.
- Because 1st and 2nd moments of $\boldsymbol{Y}_t$ are independent of time, the vector $MA(q)$ process is weakly stationary.

Vector $MA(\infty)$ Autocovariances

The vector $MA(\infty)$ is the limit of the vector $MA(q)$:

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \Theta_1 \boldsymbol{\varepsilon}_{t-1} + \Theta_2 \boldsymbol{\varepsilon}_{t-2} + \ldots$$

- The sequence of matrices $\{\Theta_s\}_{s=0}^{\infty}$ is absolutely summable if each component sequence is absolutely summable.

Vector $MA(\infty)$ Autocovariances

If $\{\Theta_s\}_{s=0}^{\infty}$ is absolutely summable:

- $E[\boldsymbol{Y}_t] = \boldsymbol{\mu}$.
- $\Gamma_j = \sum_{i=0}^{\infty} \Theta_{j+i} \Omega \Theta_i^{'}, \quad j = 0, 1, 2, \ldots$
- $\boldsymbol{Y}_t$ is ergodic for 1st and 2nd moments.
- $\boldsymbol{Y}_t$ is stationary.

Vector $VAR(p)$ Autocovariances

When a stationary $VAR(p)$ is expressed as a vector $MA(\infty)$, it satisfies the absolute summability condition.

- $\Theta_s = F^s = T\Lambda^s T^{-1}$.
- The component-wise sum of absolute values over $s = 0, 1, 2, \ldots$ will be a weighted average of absolute values of eigenvalues raised to powers.
- Because of stationarity, $|\lambda_i| < 1, i = 1, \ldots, np$, which means $\{F^s\}_{s=0}^{\infty}$ is absolutely summable.

Vector $VAR(p)$ Autocovariances

Recall that a $VAR(p)$ can be expressed as:

$$\boldsymbol{\xi}_t = F\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t$$

In this case

$$\Sigma = E[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^{'}] = \begin{bmatrix} \Gamma_0 & \Gamma_1 & \ldots & \Gamma_{p-1} \\ \Gamma_1^{'} & \Gamma_0 & \ldots & \Gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{p-1}^{'} & \Gamma_{p-2}^{'} & \ldots & \Gamma_0 \end{bmatrix}.$$

Vector $VAR(p)$ Autocovariances

By the definition of $\boldsymbol{\xi}_t$,

$$
\begin{aligned}
\Sigma &= E[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{'}] \\
&= E\left[(F\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t)(F\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t)^{'}\right] \\
&= F\underbrace{E[\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-1}^{'}]}_{\Sigma}F^{'} + \underbrace{E[\boldsymbol{v}_t\boldsymbol{v}_t^{'}]}_{Q} \\
&= F\Sigma F^{'} + Q.
\end{aligned}
$$

Using the Vec Operator

In general

$$Vec(ABC) = C^{'} \bigotimes A \cdot Vec(B).$$

Thus,

$$
\begin{aligned}
Vec(\Sigma) &= F \bigotimes F \cdot Vec(\Sigma) + Vec(Q) \\
\implies \quad Vec(\Sigma) &= [I - F \bigotimes F]^{-1} \cdot Vec(Q).
\end{aligned}
$$

- $F \otimes F$ is an $(np)^2 \times (np)^2$ matrix.
- Because all eigenvalues of F lie inside the unit circle, so do all eigenvalues of $F \otimes F$, which means $F \otimes F$ is invertible.

Vector $VAR(p)$ Autocovariances

$$
\begin{aligned}
\Sigma_j &= E[\boldsymbol{\xi}_t\boldsymbol{\xi}_{t-j}^{'}] \\
&= FE[\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-j}^{'}] \\
&= F\Sigma_{j-1}, j = 1, 2, 3, \ldots \\
\implies \quad \Sigma_j &= F^j\Sigma.
\end{aligned}
$$

State Space Models

State Space Representation

A state space model is a dynamic system of equations

$$
\begin{aligned}
\boldsymbol{\xi}_{t+1} &= F\boldsymbol{\xi}_t + \boldsymbol{v}_{t+1} \\
\boldsymbol{Y}_t &= A^{'}\boldsymbol{x}_t + H^{'}\boldsymbol{\xi}_t + \boldsymbol{w}_t \\
E[\boldsymbol{v}_t\boldsymbol{v}_\tau^{'}] &= \begin{cases} Q & t = \tau \\ \mathbf{0} & \text{o/w} \end{cases} \\
E[\boldsymbol{w}_t\boldsymbol{w}_\tau^{'}] &= \begin{cases} R & t = \tau \\ \mathbf{0} & \text{o/w} \end{cases} \\
E[\boldsymbol{v}_t\boldsymbol{w}_\tau^{'}] &= \mathbf{0} \ \forall \ t, \tau.
\end{aligned}
$$

State Space Representation

- $\boldsymbol{Y}_t$ is a vector of n variables observed at t.
- $\boldsymbol{\xi}_t$ is a vector of r unobserved variables at t.
- $\boldsymbol{x}_t$ is a vector of exogenous or predetermined variables at t.

Completely determined by the prior past values,
purely a function of past self

- The first equation of the system is the state equation.
- The second equation of the system is the observation equation.
- $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are vector WN processes and mutually uncorrelated at all lags.

Mutually Uncorrelated Errors

If we assume $E[\boldsymbol{v}_t\boldsymbol{\xi}_1^{'}] = E[\boldsymbol{w}_t\boldsymbol{\xi}_1^{'}] = \mathbf{0} \ \forall t > 1$:

$$
\begin{aligned}
E[\boldsymbol{v}_t\boldsymbol{\xi}_\tau^{'}] &= E[\boldsymbol{v}_t(\boldsymbol{v}_\tau^{'} + \boldsymbol{v}_{\tau-1}^{'}F^{'} + \ldots + \boldsymbol{\xi}_1^{'}F^{\tau-1^{'}})] = \mathbf{0} \ \forall \ \tau < t \\
E[\boldsymbol{v}_t\boldsymbol{Y}_\tau^{'}] &= E[\boldsymbol{v}_t(A^{'}\boldsymbol{x}_\tau + H^{'}\boldsymbol{\xi}_\tau + \boldsymbol{w}_\tau)^{'}] = \mathbf{0} \ \forall \ \tau < t.
\end{aligned}
$$

Similarly,

$$E[\boldsymbol{w}_t \boldsymbol{\xi}_\tau^{'}] = \mathbf{0} \;\forall\; \tau < t$$
$$E[\boldsymbol{w}_t \boldsymbol{Y}_\tau^{'}] = \mathbf{0} \;\forall\; \tau < t.$$

Example *AR*(*p*)

The standard form of an $AR(p)$:

$$Y_{t+1} - \mu = \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p+1} - \mu) + \varepsilon_{t+1}.$$

Define:

$$\boldsymbol{\xi}_t = \begin{bmatrix} Y_t - \mu \\ Y_{t-1} - \mu \\ \vdots \\ Y_{t-p+1} - \mu \end{bmatrix}, \quad F = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

Example *AR*(*p*)

$$\boldsymbol{Y}_t = Y_t, \quad A^{'} = \mu, \quad \boldsymbol{x}_t = 1, \quad H^{'} = [1\ 0\ \ldots\ 0], \quad \boldsymbol{w}_t = 0, \quad R = 0,$$

$$Q = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Example *ARMA*(*p*, *q*)

The standard form of an $ARMA(p, q)$:

$$\begin{aligned} Y_{t+1} - \mu = \phi_1(Y_t - \mu) + \ldots + \phi_r(Y_{t-r+1} - \mu) \\ + \varepsilon_{t+1} + \theta_1\varepsilon_t + \ldots + \theta_{r-1}\varepsilon_{t-r+2} \end{aligned}$$

where $r = \max\{p, q + 1\}$ and $\phi_j = 0$ for $j > p$ and $\theta_j = 0$ for $j > q$.

Example *ARMA*(*p*, *q*)

Define:

$$\boldsymbol{\xi}_t = \begin{bmatrix} Y_t - \mu \\ \phi_2(Y_{t-1} - \mu) + \ldots + \phi_r(Y_{t-r+1} - \mu) + \theta_1\varepsilon_t + \ldots + \theta_{r-1}\varepsilon_{t-r+2} \\ \phi_3(Y_{t-1} - \mu) + \ldots + \phi_r(Y_{t-r+2} - \mu) + \theta_2\varepsilon_t + \ldots + \theta_{r-1}\varepsilon_{t-r+3} \\ \vdots \\ \phi_r(Y_{t-1} - \mu) + \theta_{r-1}\varepsilon_t \end{bmatrix},$$

Example $ARMA(p, q)$

$$F = \begin{bmatrix} \phi_1 & 1 & 0 & \ldots & 0 \\ \phi_2 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \phi_{r-1} & 0 & 0 & \ldots & 1 \\ \phi_r & 0 & 0 & \ldots & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} \varepsilon_t \\ \theta_1\varepsilon_t \\ \vdots \\ \theta_{r-2}\varepsilon_t \\ \theta_{r-1}\varepsilon_t \end{bmatrix},$$

$$\boldsymbol{Y}_t = Y_t, \quad A^{'} = \mu, \quad \boldsymbol{x}_t = 1, \quad H^{'} = [1\ 0\ \ldots\ 0], \quad \boldsymbol{w}_t = 0, \quad R = 0.$$

Example $ARMA(p, q)$

Alternatively, define $\boldsymbol{\xi}_t = (\xi_t, \xi_{t-1}, \ldots, \xi_{t-r+1})^{'}$ and

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{r-1} & \phi_r \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\boldsymbol{Y}_t = Y_t, \quad A^{'} = \mu, \quad \boldsymbol{x}_t = 1, \quad H^{'} = [1\ \theta_1\ \ldots\ \theta_{r-1}], \quad \boldsymbol{w}_t = 0, \quad R = 0,$$

$$Q = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Example $ARMA(p, q)$

Then, by the state equation

$$\begin{aligned} \boldsymbol{\xi}_t &= F\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t \\ \Longrightarrow \quad \phi_r(L)\xi_t &= \varepsilon_t, \end{aligned}$$

and the observation equation

$$\begin{aligned} \boldsymbol{Y}_t &= A^{'}\boldsymbol{x}_t + H^{'}\boldsymbol{\xi}_t + \boldsymbol{w}_t \\ \Longrightarrow \quad (Y_t - \mu) &= \theta_r(L)\xi_t. \end{aligned}$$

Example $ARMA(p, q)$

Combining these two equations,

$$\begin{aligned} \phi_r(L)(Y_t - \mu) &= \theta_r(L)\phi_r(L)\xi_t \\ &= \theta_r(L)\varepsilon_t, \end{aligned}$$

which is the equation for an $ARMA(p, q)$.

The Kalman Filter

State-Space Representation

Recall the basic state-space representation

$$\begin{aligned} \xi_{t+1}(r\times 1) &= F(r\times r)\xi_t(r\times 1) + \boldsymbol{v}_{t+1}(r\times 1) \\ \boldsymbol{Y}_t(n\times 1) &= A'(n\times k)\boldsymbol{x}_t(k\times 1) + H'(n\times r)\xi_t(r\times 1) + \boldsymbol{w}_t(n\times 1) \end{aligned}$$

$$E[\boldsymbol{v}_t\boldsymbol{v}_\tau'] = \begin{cases} Q(n\times n) & t=\tau \\ 0 & \text{o/w} \end{cases}$$

rxr not nxn

$$E[\boldsymbol{w}_t\boldsymbol{w}_\tau'] = \begin{cases} R(n\times n) & t=\tau \\ 0 & \text{o/w} \end{cases}$$

$$E[\boldsymbol{v}_t\boldsymbol{w}_\tau'] = 0 \;\; \forall \;\; t, \tau.$$

Kalman Filter Overview

Collect all known information at time t into a vector:

$$\mathrm{Y}_t((n+k)t\times 1) = (\boldsymbol{Y}_t', \boldsymbol{Y}_{t-1}', \ldots, \boldsymbol{Y}_1', \boldsymbol{x}_t', \boldsymbol{x}_{t-1}', \ldots, \boldsymbol{x}_1')'$$

The Kalman Filter computes:

$$\begin{aligned} \hat{\xi}_{t+1|t} &= \hat{E}[\xi_{t+1}|\mathrm{Y}_t] \\ P_{t+1|t}(r\times r) &= E[(\xi_{t+1} - \hat{\xi}_{t+1|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'], \end{aligned}$$

Want a forecast of the latent state and the mean squared error

where $P_{t+1|t}$ is the MSE matrix for $\hat{\xi}_{t+1|t}$.

Starting the Recursion

We begin the recursion with

Conditional expectation but conditioned on the null set

$$\hat{\xi}_{1|0} = E[\xi_1|\mathcal{Y}_0 = \emptyset] = E[\xi_1]$$

Loading [MathJax]/jax/output/HTML-CSS/jax.js

$${}_0 = E[(\xi_1 - E[\xi_1])(\xi_1 - E[\xi_1])'].$$

According to the state equation, the unconditional expectation of ξ_t is:

$$\begin{aligned} E[\xi_{t+1}] &= FE[\xi_t] \\ \implies E[\xi_t] &= FE[\xi_t] \\ \implies (I_r - F)E[\xi_t] &= 0 \\ \implies E[\xi_t] &= 0. \end{aligned}$$

Starting the Recursion

Further, the state equation also implies the unconditional variance of ξ_t is:

$$\begin{aligned} \underbrace{E[\xi_{t+1}\xi_{t+1}^{'}]}_{\Sigma} &= E[(F\xi_t + \boldsymbol{v}_{t+1})(F\xi_t + \boldsymbol{v}_{t+1})^{'}] \\ &= F\underbrace{E[\xi_t\xi_t^{'}]}_{\Sigma}F^{'} + F\underbrace{E[\xi_t\boldsymbol{v}_{t+1}^{'}]}_{0} + \underbrace{E[\boldsymbol{v}_{t+1}\xi_t^{'}]}_{0}F^{'} + \underbrace{E[\boldsymbol{v}_{t+1}\boldsymbol{v}_{t+1}^{'}]}_{Q} \\ \implies \Sigma &= F\Sigma F^{'} + Q \\ \implies Vec(P_{1|0}) &= Vec(\Sigma) = [I_{r^2} - (F \otimes F)]^{-1}Vec(Q). \end{aligned}$$

Forecasting Y_t

Our objective will be to obtain $\hat{\xi}_{t+1|t}$ and $P_{t+1|t}$, given values for $\hat{\xi}_{t|t-1}$ and $P_{t|t-1}$.

- $\boldsymbol{x}_t$ contains no information about ξ_t beyond what is contained in Y_{t-1}:

$$E[\xi_t|\boldsymbol{x}_t, \mathrm{Y}_{t-1}] = E[\xi_t|\mathrm{Y}_{t-1}] = \hat{\xi}_{t|t-1}.$$

According to the observation equation:

$$\begin{aligned} \hat{\boldsymbol{Y}}_{t|t-1} &= \hat{E}[\boldsymbol{Y}_t|\boldsymbol{x}_t, \mathrm{Y}_{t-1}] \\ &= A^{'}\boldsymbol{x}_t + H^{'}E[\xi_t|\boldsymbol{x}_t, \mathrm{Y}_{t-1}] + \underbrace{E[\boldsymbol{w}_t|\boldsymbol{x}_t, \mathrm{Y}_{t-1}]}_{0} \\ &= A^{'}\boldsymbol{x}_t + H^{'}\hat{\xi}_{t|t-1}. \end{aligned}$$

Forecast Error

The forecast error is:

$$\begin{aligned} \boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1} &= A^{'}\boldsymbol{x}_t + H^{'}\xi_t + \boldsymbol{w}_t - A^{'}\boldsymbol{x}_t - H^{'}\hat{\xi}_{t|t-1} \\ &= H^{'}(\xi_t - \hat{\xi}_{t|t-1}) + \boldsymbol{w}_t, \end{aligned}$$

which has the MSE matrix:

$$
\begin{aligned}
E[(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})^{'}] \\
&= H^{'} \underbrace{E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})^{'}]}_{P_{t|t-1}} H + \underbrace{E[\boldsymbol{w}_t \boldsymbol{w}_t^{'}]}_{R} \\
&= H^{'} P_{t|t-1} H + R.
\end{aligned}
$$

Forecast MSE

We have used the fact that:

$$E[\boldsymbol{w}_t(\xi_t - \hat{\xi}_{t|t-1})] = 0$$

because $E[\boldsymbol{w}_t \boldsymbol{\xi}_t^{'}] = 0$ and because

$$E[\boldsymbol{w}_t \hat{\boldsymbol{\xi}}_{t|t-1}^{'}] = E[\boldsymbol{w}_t (F \xi_{t-1})^{'}] = E[\boldsymbol{w}_t \boldsymbol{\xi}_{t-1}^{'}] F^{'} = 0.$$

Update the forecast of ξ_t

After we observe $\boldsymbol{Y}_t$, we can obtain a new forecast of ξ_t:

$$\hat{\xi}_{t|t} = E[\boldsymbol{\xi}_t | \boldsymbol{Y}_t, \boldsymbol{x}_t, \Upsilon_{t-1}] = E[\boldsymbol{\xi}_t | \Upsilon_t].$$

Update the forecast of ξ_t

The formula for updating a linear projection in this fashion is:

$$
\begin{aligned}
\hat{\xi}_{t|t} &= \hat{\xi}_{t|t-1} + E[(\xi_t - \hat{\xi}_{t|t-1})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})^{'}] \\
&\qquad \times E[(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})^{'}]^{-1} (\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1}) \\
&= \hat{\xi}_{t|t-1} + E[(\xi_t - \hat{\xi}_{t|t-1})(\boldsymbol{w}_t^{'} + (\xi_t - \hat{\xi}_{t|t-1})^{'} H)] \\
&\qquad \times (H^{'} P_{t|t-1} H + R)^{-1} (\boldsymbol{Y}_t - A^{'} \boldsymbol{x}_t - H^{'} \hat{\xi}_{t|t-1}) \\
&= \hat{\xi}_{t|t-1} + \overbrace{P_{t|t-1} H (H^{'} P_{t|t-1} H + R)^{-1}}^{K_t} (\boldsymbol{Y}_t - A^{'} \boldsymbol{x}_t - H^{'} \hat{\xi}_{t|t-1}).
\end{aligned}
$$

Update the forecast of ξ_t

The associated MSE is:

$$
\begin{aligned}
P_{t|t} &= E[(\xi_t - \hat{\xi}_{t|t})(\xi_t - \hat{\xi}_{t|t})^{'}] \\
&= P_{t|t-1} - P_{t|t-1} H (H^{'} P_{t|t-1} H + R)^{-1} H^{'} P_{t|t-1}.
\end{aligned}
$$

Forecasting ξ_{t+1}

$$\begin{aligned}\hat{\boldsymbol{\xi}}_{t+1|t} &= \hat{E}[\boldsymbol{\xi}_{t+1} | \Upsilon_t] \\ &= F\hat{E}[\boldsymbol{\xi}_t | \Upsilon_t] + E[\boldsymbol{v}_{t+1} | \Upsilon_t] \\ &= F\hat{\boldsymbol{\xi}}_{t|t} \\ &= F\hat{\boldsymbol{\xi}}_{t|t-1} + FK_t(\boldsymbol{Y}_t - A^{'}\boldsymbol{x}_t + \overset{\text{Yesterdays forecast error}}{H^{'}\hat{\boldsymbol{\xi}}_{t|t-1}}).\end{aligned}$$

$$\begin{aligned}P_{t+1|t} &= E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})^{'}] \\ &= E[(F\boldsymbol{\xi}_t + \boldsymbol{v}_{t+1} - F\hat{\boldsymbol{\xi}}_{t|t})(F\boldsymbol{\xi}_t + \boldsymbol{v}_{t+1} - F\hat{\boldsymbol{\xi}}_{t|t})^{'}] \\ &= FP_{t|t}F^{'} + Q \\ &= F(P_{t|t-1} - P_{t|t-1}H(H^{'}P_{t|t-1}H + R)^{-1}H^{'}P_{t|t-1})F^{'} + Q.\end{aligned}$$

Forecasting Y_{t+1}

$$\hat{\boldsymbol{Y}}_{t+1|t} = E[\boldsymbol{Y}_{t+1} | \boldsymbol{x}_{t+1}, \Upsilon_t] = A^{'}\boldsymbol{x}_{t+1} + H^{'}\hat{\boldsymbol{\xi}}_{t+1|t},$$

which has associated MSE:

$$E[(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t})(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t})^{'}] = H^{'}P_{t+1|t}H + R.$$

Forecasting Y_{t+s}

Iterating on the state equation:

$$\begin{gathered}\boldsymbol{\xi}_{t+s} = F^s\boldsymbol{\xi}_t + F^{s-1}\boldsymbol{v}_{t+1} + F^{s-2}\boldsymbol{v}_{t+2} + \ldots + F\boldsymbol{v}_{t+s-1} + \boldsymbol{v}_{t+s} \\ \implies E[\boldsymbol{\xi}_{t+s} | \boldsymbol{\xi}_t, \Upsilon_t] = F^s\boldsymbol{\xi}_t.\end{gathered}$$

By the law of iterated expectations:

$$\hat{\boldsymbol{\xi}}_{t+s|t} = E[\boldsymbol{\xi}_{t+s} | \Upsilon_t] = E[E[\boldsymbol{\xi}_{t+s} | \boldsymbol{\xi}_t, \Upsilon_t] | \Upsilon_t] = E[F^s\boldsymbol{\xi}_t | \Upsilon_t] = F^s\hat{\boldsymbol{\xi}}_{t|t}.$$

Forecasting Y_{t+s}

The forecast error is:

$$\boldsymbol{\xi}_{t+s} - \hat{\boldsymbol{\xi}}_{t+s|t} = F^s(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t}) + F^{s-1}\boldsymbol{v}_{t+1} + \ldots + F\boldsymbol{v}_{t+s-1} + \boldsymbol{v}_{t+s},$$

with MSE:

$$P_{t+s|t} = F^sP_{t|t}(F^{'})^s + F^{s-1}Q(F^{'})^{s-1} + \ldots + FQF^{'} + Q.$$

Forecasting Y_{t+s}

$$\boldsymbol{Y}_{t+s} = A\ \boldsymbol{x}_{t+s} + H\ \boldsymbol{\xi}_{t+s} + \boldsymbol{w}_{t+s}.$$

Thus,

$$\hat{\boldsymbol{Y}}_{t+s|t} = E[\boldsymbol{Y}_{t+s}|\boldsymbol{x}_{t+s}, \Upsilon_t] = A^{'}\boldsymbol{x}_{t+s} + H^{'}\hat{\boldsymbol{\xi}}_{t+s|t}.$$

Forecasting $\boldsymbol{Y}_{t+s}$

The forecast error is:

$$\begin{aligned}
\boldsymbol{Y}_{t+s} - \hat{\boldsymbol{Y}}_{t+s|t} &= A^{'}\boldsymbol{x}_{t+s} + H^{'}\boldsymbol{\xi}_{t+s} + \boldsymbol{w}_{t+s} - A^{'}\boldsymbol{x}_{t+s} - H^{'}\hat{\boldsymbol{\xi}}_{t+s|t} \\
&= H^{'}(\boldsymbol{\xi}_{t+s} - \hat{\boldsymbol{\xi}}_{t+s|t}) + \boldsymbol{w}_{t+s},
\end{aligned}$$

with MSE:

$$E[(\boldsymbol{Y}_{t+s} - \hat{\boldsymbol{Y}}_{t+s|t})(\boldsymbol{Y}_{t+s} - \hat{\boldsymbol{Y}}_{t+s|t})^{'}] = H^{'}P_{t+s|t}H + R.$$

Summary of Kalman Filter Steps

1. Start with forecast $\hat{\xi}_{1|0}$ and associated MSE matrix $P_{1|0}$

2. Given forecast $\hat{\xi}_{t|t-1}$ and MSE $P_{t|t-1}$, compute

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_{t|t} &= \hat{\boldsymbol{\xi}}_{t|t-1} + \overbrace{P_{t|t-1}H(H^{'}P_{t|t-1}H + R)^{-1}}^{K_t}(\boldsymbol{Y}_t - A^{'}\boldsymbol{x}_t + H^{'}\hat{\boldsymbol{\xi}}_{t|t-1}) \\
P_{t|t} &= P_{t|t-1} - P_{t|t-1}H(H^{'}P_{t|t-1}H + R)^{-1}H^{'}P_{t|t-1}.
\end{aligned}$$

Summary of Kalman Filter Steps

3. Given $\hat{\xi}_{t|t}$ and MSE $P_{t|t}$, compute

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_{t+1|t} &= F\hat{\boldsymbol{\xi}}_{t|t-1} + FK_t(\boldsymbol{Y}_t - A^{'}\boldsymbol{x}_t + H^{'}\hat{\boldsymbol{\xi}}_{t|t-1}) \\
P_{t+1|t} &= F(P_{t|t-1} - P_{t|t-1}H(H^{'}P_{t|t-1}H + R)^{-1}H^{'}P_{t|t-1})F^{'} + Q.
\end{aligned}$$

4. Given $\hat{\xi}_{t+1|t}$ and MSE $P_{t+1|t}$, compute

$$\begin{aligned}
\hat{\boldsymbol{Y}}_{t+1|t} &= A^{'}\boldsymbol{x}_{t+1} + H^{'}\hat{\boldsymbol{\xi}}_{t+1|t} \\
E[(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t})(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t})^{'}] &= H^{'}P_{t+1|t}H + R.
\end{aligned}$$

Example: Long-Run Risks

$$x_{t+1} = \rho x_t + \varphi_e \sigma e_{t+1}$$

$$g_{t+1} = \log\left(C_{t+1}/C_t\right) = \mu + x_t + \sigma \eta_{t+1}$$

$$g_{d,t+1} = \log\left(D_{t+1}/D_t\right) = \mu_d + \phi x_t + \varphi_d \sigma u_{t+1}$$

$$\varphi_{t+1}, u_{t+1}, \eta_{t+1} \overset{i.i.d.}{\sim} N(0, 1)$$

where C_t and D_t are aggregate consumption and aggregate dividends.

ML Estimation of State-Space Models

Kalman Filter Forecasts

The Kalman filter forecasts $\hat{\xi}_{t|t-1}$ and $\hat{\boldsymbol{Y}}_{t|t-1}$ are linear projections of ξ_t and $\boldsymbol{Y}_t$ on $(\boldsymbol{x}_t, \mathrm{Y}_{t-1})$.

- They are optimal among all forecasts that are linear functions of $(\boldsymbol{x}_t, \mathrm{Y}_{t-1})$.
- If ξ_1 and $\{\boldsymbol{w}_t, \boldsymbol{v}_t\}_{t=1}^{T}$ are multivariate Gaussian, $\hat{\xi}_{t|t-1}$ and $\hat{\boldsymbol{Y}}_{t|t-1}$ are optimal among *all* forecasts that are functions of $(\boldsymbol{x}_t, \mathrm{Y}_{t-1})$ (linear and non-linear).

Conditional Distribution of $\boldsymbol{Y}_t$

The distribution of $\boldsymbol{Y}_t | \boldsymbol{x}_t, \mathrm{Y}_{t-1}$ is also multivariate Gaussian, of the form:

$$\boldsymbol{Y}_t | \boldsymbol{x}_t, \mathrm{Y}_{t-1} \sim MVN(A^{'}\boldsymbol{x}_t + H^{'}\hat{\xi}_{t|t-1}, H^{'}P_{t|t-1}H + R)$$

Conditional Distribution of $\boldsymbol{Y}_t$

Thus, the density function is

$$\begin{aligned} & f_{\boldsymbol{Y}_t|\boldsymbol{x}_t, \mathrm{Y}_{t-1}}(\boldsymbol{Y}_t | \boldsymbol{x}_t, \mathrm{Y}_{t-1}, \boldsymbol{\theta}) \\ & \qquad = (2\pi)^{-n/2} |H^{'}P_{t|t-1}H + R|^{-1/2} \\ & \qquad\qquad \times \exp\left\{ -\frac{1}{2}\left(\boldsymbol{Y}_t - A^{'}\boldsymbol{x}_t - H^{'}\hat{\xi}_{t|t-1}\right)\right. \\ & \qquad\qquad\qquad \times \left(H^{'}P_{t|t-1}H + R\right)^{-1} \\ & \qquad\qquad\qquad\qquad \left. \times \left(\boldsymbol{Y}_t - A^{'}\boldsymbol{x}_t - H^{'}\hat{\xi}_{t|t-1}\right)^{'} \right\} \end{aligned}$$

where $\boldsymbol{\theta}$ aggregates all known parameters in $F, A, H, Q,$ and R.

Log-likelihood

The log-likelihood is the joint density

$$\ell(\boldsymbol{\theta}) = \sum^{T} \log\left(f_{\boldsymbol{Y}_t|\boldsymbol{X}_t, \mathrm{Y}_{t-1}}(\boldsymbol{Y}_t | \boldsymbol{x}_t, \mathrm{Y}_{t-1}, \boldsymbol{\theta})\right)$$

- The log-likelihood can be maximized numerically with respect to $F(\boldsymbol{\theta})$, $A(\boldsymbol{\theta})$, $H(\boldsymbol{\theta})$, $Q(\boldsymbol{\theta})$, and $R(\boldsymbol{\theta})$.

- This is an exact log likelihood and yields exact MLEs.

- Maximum likelihood estimation for MA and $ARMA$ can be performed in this manner.

Basic Prescription

1. Guess $\boldsymbol{\theta}^{(0)}$

2. Given $\boldsymbol{\theta}^{(s)}$, compute $F(\boldsymbol{\theta}^{(s)}), A(\boldsymbol{\theta}^{(s)}), H(\boldsymbol{\theta}^{(s)}), Q(\boldsymbol{\theta}^{(s)}),$ and $R(\boldsymbol{\theta}^{(s)})$.

3. Use the Kalman Filter to iteratively compute $\hat{\boldsymbol{\xi}}_{t|t-1}$ and $P_{t|t-1}, t = 1, \ldots, T$.

4. Compute the log-likelihood using $H(\boldsymbol{\theta}^{(s)}), A(\boldsymbol{\theta}^{(s)}), R(\boldsymbol{\theta}^{(s)})$, and $\{\hat{\boldsymbol{\xi}}_{t|t-1}, P_{t|t-1}\}_{t=1}^{T}$.

5. Use a numerical method to update $\boldsymbol{\theta}^{(s+1)}$.

6. If $||\boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)}|| < \tau$, stop. Otherwise, set $i = i + 1$ and return to step 2.

Basic Prescription

Updating $\boldsymbol{\theta}^{(i)} \rightarrow \boldsymbol{\theta}^{(i+1)}$ may involve numerical or analytical derivatives.

- Analytical derivatives of the log likelihood with respect to each θ_i will involve
$$\frac{\partial \hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta})}{\partial \theta_i} \text{ and } \frac{\partial P_{t|t-1}}{\partial \theta_i}.$$
- These derivatives can be updated recursively similar to $\hat{\boldsymbol{\xi}}_{t|t-1}$ and $P_{t|t-1}$.

Method of Moments

Method of Moments

Suppose $\{y_t\}_{t=1}^T$ is an i.i.d. sample of random variable Y from density $f_Y(y\,|\,\boldsymbol{\theta})$.

- $\boldsymbol{\theta}$ is a $(k \times 1)$ dimensional vector of parameters.

Suppose k population moments can be written as functions of $\boldsymbol{\theta}$:

$$E[Y_t^i] = \mu_i(\boldsymbol{\theta}), \quad i = i_1, i_2, \ldots, i_k.$$

Method of Moments

The method of moments estimator, $\hat{\boldsymbol{\theta}}_{mm}$, of $\boldsymbol{\theta}$ is the value:

$$\mu_i(\hat{\theta}_{mm}) = \frac{1}{T}\sum_{t=1}^{T} y_t^i, \quad i = i_1, i_2, \ldots, i_k.$$

- Note that if you need to estimate k parameters, you must specify exactly k moments.

Example: Normal

- $\boldsymbol{\theta} = (\mu, \sigma^2)'$

- $k = 2$

- $E[Y^1] = \mu$

- $E[Y^2] = Var(Y) + E[Y]^2 = \sigma^2 + \mu^2$.

Example: Beta Distribution

Suppose $Y \sim \text{Beta}(\alpha, \beta)$:

$$f_Y(y\,|\,\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}.$$

In this case, $\boldsymbol{\theta} = (\alpha, \beta)'$ and:

$$\mu_1 = E[Y^1] = \frac{\alpha}{\alpha + \beta}$$

$$\sigma^2 = Var(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Example: Beta Distribution

$$\begin{aligned} \implies \mu_2 &= E[Y^2] \\ &= Var(Y) + E[Y^1]^2 \\ &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \frac{\alpha^2}{(\alpha + \beta)^2} \\ &= \frac{\alpha\beta + \alpha^2(\alpha + \beta + 1)}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned}$$

Example: Beta Distribution

Solve for β using μ_1:

$$\begin{aligned} \alpha &= \mu_1(\alpha + \beta) \\ \implies \alpha &= \mu_1\alpha + \mu_1\beta \\ \implies \alpha(1 - \mu_1) &= \mu_1\beta \\ \implies \beta &= \frac{\alpha(1 - \mu_1)}{\mu_1}. \end{aligned}$$

Example: Beta Distribution

From the relationship $\alpha = \mu_1(\alpha + \beta)$ we have:

$$(\alpha + \beta) = \frac{\alpha}{\mu_1}$$

$$(\alpha + \beta + 1) = \frac{\alpha}{\mu_1} + 1 = \frac{\alpha + \mu_1}{\mu_1}.$$

Example: Beta Distribution

Now substitute for β, $(\alpha + \beta)$ and $(\alpha + \beta + 1)$ in μ_2:

$$\begin{aligned}
\mu_2 &= \frac{\alpha^2\left(\frac{1-\mu_1}{\mu_1}\right) + \alpha^2\left(\frac{\alpha+\mu_1}{\mu_1}\right)}{\frac{\alpha^2}{\mu_1^2}\cdot\frac{\alpha+\mu_1}{\mu_1}} \\
&= \frac{1-\mu_1+\alpha+\mu_1}{\frac{\alpha+\mu_1}{\mu_1^2}} \\
&= \frac{(1+\alpha)\mu_1^2}{\alpha+\mu_1}.
\end{aligned}$$

Example: Beta Distribution

$$\begin{aligned}
&\implies \alpha\mu_2 + \mu_1\mu_2 = \mu_1^2 + \alpha\mu_1^2 \\
&\implies \alpha(\mu_2 - \mu_1^2) = \mu_1^2 - \mu_1\mu_2 \\
&\implies \alpha = \frac{\mu_1^2 - \mu_1\mu_2}{\underbrace{\mu_2 - \mu_1^2}_{\sigma^2}} = \frac{\mu_1^2 - \mu_1\mu_2 + \overbrace{\mu_1^3 - \mu_1^3}^{\sigma^2}}{\sigma^2} \\
&\qquad = \frac{\mu_1^2(1-\mu_1) - \mu_1(\mu_2 - \mu_1^2)}{\sigma^2} \\
&\qquad = \frac{\mu_1^2(1-\mu_1)}{\sigma^2} - \mu_1.
\end{aligned}$$

Example: Beta Distribution

Thus,

$$\beta = \frac{\alpha(1-\mu_1)}{\mu_1} = \frac{\mu_1(1-\mu_1)^2}{\sigma^2} - (1-\mu_1).$$

The result is,

$$\begin{aligned}
\hat{\alpha}_{mm} &= \frac{\hat{\mu}_1^2(1-\hat{\mu}_1)}{\hat{\sigma^2}} - \hat{\mu}_1 \\
\hat{\beta}_{mm} &= \frac{\hat{\mu}_1(1-\hat{\mu}_1)^2}{\hat{\sigma^2}} - (1-\hat{\mu}_1).
\end{aligned}$$

Example: Beta Distribution

Where,

$$
\begin{aligned}
\hat{\mu}_1 &= \frac{1}{T}\sum_{t=1}^{T} y_t \\
\hat{\mu}_2 &= \frac{1}{T}\sum_{t=1}^{T} y_t^2 \\
\hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2.
\end{aligned}
$$

Example: t distribution

Suppose $Y \sim t(v)$:

$$
f_Y(y \mid v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{(\pi v)^{1/2}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{y^2}{v}\right)^{-\frac{v+1}{2}}.
$$

In this case $\boldsymbol{\theta} = v$.

Example: t distribution

If $v > 2$,

$$
\begin{aligned}
\mu_2 &= \frac{v}{v-2} \\
\implies \quad v &= v\mu_2 - 2\mu_2 \\
\implies \quad v &= \frac{2\mu_2}{\mu_2 - 1} \\
\implies \quad \hat{v}_{mm} &= \frac{2\hat{\mu}_2}{\hat{\mu}_2 - 1},
\end{aligned}
$$

where

$$
\hat{\mu}_2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2.
$$

Generalized Method of Moments

Setup

Let $\boldsymbol{Y}_t$ be an $(n \times 1)$ vector of random variables and $\boldsymbol{\theta}$ a $(k \times 1)$ vector of parameters governing the process $\{\boldsymbol{Y}_t\}$.

- Denote the true parameter vector as $\boldsymbol{\theta}_0$.

Moment Conditions

Suppose we can specify an $(r \times 1)$ vector valued function $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t) : (\mathrm{R}^k \times \mathrm{R}^n) \rightarrow \mathrm{R}^r$ such that:

$$E[\boldsymbol{h}(\boldsymbol{\theta}_0, \boldsymbol{Y}_t)] = 0, \quad \text{where} \;\; r \geq k.$$

Define $\mathcal{Y}_t = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$ and

$$\boldsymbol{g}_T(\boldsymbol{\theta} \,|\, \mathcal{Y}_T) = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t).$$

Note that $\boldsymbol{g}_T(\boldsymbol{\theta} \,|\, \mathcal{Y}_T) : \mathrm{R}^k \rightarrow \mathrm{R}^r$.

GMM Estimator

We want to choose $\hat{\boldsymbol{\theta}}_{gmm}$ such that the sample moments $\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}_{gmm} \,|\, \mathcal{Y}_T)$ are close to zero.

- If $r = k$, we can choose $\hat{\boldsymbol{\theta}}_{gmm}$ such that $\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}_{gmm} \,|\, \mathcal{Y}_T) = 0$ because we have k equations and k unknowns.

- If $r > k$ we have more equations than unknowns; in general there is no $\hat{\boldsymbol{\theta}}_{gmm}$ such that $\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}_{gmm} \,|\, \mathcal{Y}_T) = 0$.

GMM Estimator

If $r > k$, we minimize a quadratic form:

$$Q_T(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)_{1 \times 1} = \boldsymbol{g}_T(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)'_{(1 \times r)} W_{T(r \times r)} \boldsymbol{g}_T(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)_{(r \times 1)}.$$

- The matrix W_T places more weight on some moment conditions and less on others.

- We might have to use numerical optimization to minimze $Q_T(\boldsymbol{\theta} | \mathrm{Y}_T)$.

Example: t-distribution

The method of moments estimator of the t-distribution is a special case of the GMM estimator.

- $\boldsymbol{Y}_t = Y_t$.
- $\boldsymbol{\theta} = v$
- $W_T = 1$
- $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t) = Y_t^2 - \frac{v}{v-2}$.

Note that

$$E[Y_t^2] = \frac{v}{v-2}.$$

Example: t-distribution

$$E[\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t)] = E\left[Y_t^2 - \frac{v}{v-2}\right] = 0$$

$$\boldsymbol{g}_T(\boldsymbol{\theta} | \mathrm{Y}_T) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t^2 - \frac{v}{v-2}\right).$$

In this case, $r = k = 1$, and

$$Q_T(\boldsymbol{\theta} | \mathrm{Y}_T) = \left[\frac{1}{T}\sum_{t=1}^{T}\left(y_t^2 - \frac{v}{v-2}\right)\right]^2.$$

Since $r = k = 1$, $\hat{v}_{gmm}$ can be chosen such that $Q_T(\boldsymbol{\theta} | \mathrm{Y}_T) = 0$.

Example: t-distribution with $r = 2$

Suppose we add a moment condition for the t-distribution.

- If $v > 4$, then

$$\mu_4 = E[Y_t^4] = \frac{3v^2}{(v-2)(v-4)}.$$

- In this case, $r = 2 > 1 = k$.
- We now have more moment conditions than parameters.

Example: t-distribution with $r = 2$

We map this problem into GMM form in the following way:

- $\boldsymbol{Y}_t = Y_t$
- $\boldsymbol{\theta} = v$

$$W_T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t) = \begin{bmatrix} Y_t^2 - \frac{v}{v-2} \\ Y_t^4 - \frac{3v^2}{(v-2)(v-4)} \end{bmatrix}$$

$$\boldsymbol{g}_T(\boldsymbol{\theta} | \mathrm{Y}_T) = \frac{1}{T}\sum_{t=1}^{T} h(\boldsymbol{\theta}, \boldsymbol{y}_t).$$

Example: t-distribution with $r = 2$

The weighting matrix $W_T = I_2$ places equal weight on the two moment conditions.

- We could alter this matrix to emphasize one condition more than another.

GMM Consistency

If $\boldsymbol{Y}_t$ is strictly stationary and $\boldsymbol{h}$ continuous, a law of large numbers will hold:

$$\boldsymbol{g}_T(\boldsymbol{\theta}) \overset{p}{\rightarrow} E[\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t)].$$

Under certain regularity conditions, it can be shown that

$$\boldsymbol{\theta}_{gmm} \overset{p}{\rightarrow} \boldsymbol{\theta}_0.$$

Docs » Generalized Method of Moments » GMM Optimal Weighting Matrix

GMM Optimal Weighting Matrix

Moment Conditions' Covariance

Suppose $\{\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{Y}_t)\}_{t=1}^{T}$ is strictly stationary and define

$$\Gamma_v = E[\boldsymbol{h}(\boldsymbol{\theta}_{\boldsymbol{0}}, \boldsymbol{Y}_t)\boldsymbol{h}(\boldsymbol{\theta}_{\boldsymbol{0}}, \boldsymbol{Y}_{t-v})^{'}]$$

and

$$S = \sum_{v=-\infty}^{\infty}\Gamma_v = \Gamma_0 + \sum_{v=1}^{\infty}(\Gamma_v + \Gamma_v^{'}).$$

Convergence in Distribution

Asymptotic theory dictates

$$\sqrt{T}(\boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_{\boldsymbol{T}}) - E[\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t)]) \xrightarrow{d} N(0, S)$$

where

$$\sum_{t=1}^{T}\boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_{\boldsymbol{T}})\boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_{\boldsymbol{T}})^{'} \xrightarrow{p} S.$$

Optimal Weighting Matrix

Another way to say this (intuitively):

$$\boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_T) \overset{approx}{\sim} N\left(E[\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t)], \frac{S}{T}\right)$$

The optimal GMM weighting matrix is S^{-1}:

$$Q_T(\boldsymbol{\theta}) = \boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_T)^{'} S^{-1}\boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_T).$$

Optimal Matrix Estimation

If $\{\boldsymbol{h}(\boldsymbol{\theta}_{\boldsymbol{0}}, \boldsymbol{y}_t)\}_{t=-\infty}^{\infty}$ is serially uncorrelated, S is consistently estimated by

$$S_T^{*} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}(\boldsymbol{\theta}_{\boldsymbol{0}}, \boldsymbol{y}_t)\boldsymbol{h}(\boldsymbol{\theta}_{\boldsymbol{0}}, \boldsymbol{y}_t)^{'}.$$

If it is serially correlated,

$$S_T^* = \Gamma_{0,T}^* + \sum_{v=1}\left(1 - \frac{v}{q+1}\right)\left(\Gamma_{v,T}^* + \Gamma_{v,T}^*\right),$$

where

$$\Gamma_{v,T}^* = \frac{1}{T}\sum_{t=v+1}^{T} \boldsymbol{h}(\boldsymbol{\theta}_0, \boldsymbol{y}_t)\boldsymbol{h}(\boldsymbol{\theta}_0, \boldsymbol{y}_{t-v})'.$$

Optimal Matrix Estimation

Notice that S^* depends on $\boldsymbol{\theta}_0$, which is unknown.

- We substitute an estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ in S^* and denote the estimated value as $\hat{S}$.
- $\hat{S}$ may make use of appropriate definitions of $\hat{\Gamma}_{v,T}$ if there is serial correlation.

Under certain regularity conditions

$$\hat{S} \xrightarrow{p} S.$$

Two Stage Estimation

Note that we want to use $\hat{S}^{-1}$ as the optimal weighting matrix to compute $\hat{\boldsymbol{\theta}}$, but that $\hat{S}^{-1}$ depends on $\hat{\boldsymbol{\theta}}$.

- To compute optimal $\hat{\boldsymbol{\theta}}_{gmm}$, first estimate $\hat{\boldsymbol{\theta}}_{gmm}$ with $W_T = I_r$.
- Use the initial $\hat{\boldsymbol{\theta}}_{gmm}$ to compute $\hat{S}_T(\hat{\boldsymbol{\theta}}_{gmm})$ and set $W_T = \hat{S}_T(\hat{\boldsymbol{\theta}}_{gmm})^{-1}$.
- Compute $\hat{\boldsymbol{\theta}}_{gmm}$ again.

Two Stage Estimation

How is the two-stage procedure better?

- That is, why is S^{-1} optimal?
- Using S^{-1} or a consistent estimate, $\hat{S}^{-1}$, results in $\hat{\boldsymbol{\theta}}_{gmm}$ with less estimation error.

Asymptotic Distribution of GMM Estimator

A central limit theorem exists for $\hat{\boldsymbol{\theta}}_{gmm}$:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{gmm} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, V),$$

where

$$V = (DS^{-1}D^{'})^{-1}$$
$$\frac{\partial \boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_{\boldsymbol{T}})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \overset{p}{\longrightarrow} D.$$

Asymptotic Distribution of GMM Estimator

That is, for large T,

$$\hat{\boldsymbol{\theta}}_{gmm} \overset{approx}{\sim} N(\boldsymbol{\theta}_0, \frac{\hat{V}_T}{T}),$$

where,

$$\hat{V}_T = (\hat{D}_T \hat{S}_T^{-1} \hat{D}_T^{'})^{-1}$$
$$\hat{D}_T = \frac{\partial \boldsymbol{g}_T(\boldsymbol{\theta} \mid \mathrm{Y}_{\boldsymbol{T}})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

GMM Over-Identifying Restrictions

Criterion Function Limiting Distribution

Since

$$\begin{aligned} \sqrt{T}\boldsymbol{g}_T(\boldsymbol{\theta}_0) &\xrightarrow{d} N(0, S) \\ \implies (\sqrt{T}\boldsymbol{g}_T(\boldsymbol{\theta}_0)')S^{-1}(\sqrt{T}\boldsymbol{g}_T(\boldsymbol{\theta}_0)) = T\boldsymbol{g}_T(\boldsymbol{\theta}_0)'S^{-1}\boldsymbol{g}_T(\boldsymbol{\theta}_0) &\xrightarrow{d} \chi^2(r) \end{aligned}$$

where $r > k$ is the number of moment conditions.

Estimated Criterion Function

It turns out that

$$T\boldsymbol{g}_T(\hat{\boldsymbol{\theta}})'\hat{S}^{-1}\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}) \not\xrightarrow{d} \chi^2(r).$$

- This is because k moment conditions will be set to zero exactly.

Exact Identification

Consider $r = k$. In this case

$$\begin{aligned} \boldsymbol{g}_T(\hat{\boldsymbol{\theta}}) &= 0 \\ T\boldsymbol{g}_T(\hat{\boldsymbol{\theta}})'\hat{S}^{-1}\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}) &= 0. \end{aligned}$$

- $r - k$ of the moment conditions will be non-zero.

Over Identification

In general,

$$J_T(\hat{\boldsymbol{\theta}}) = T\boldsymbol{g}_T(\hat{\boldsymbol{\theta}})'\hat{S}^{-1}\boldsymbol{g}_T(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(r - k).$$

- To test if our moment conditions are close to zero, we compute $J_T(\hat{\boldsymbol{\theta}})$ and compare with a $\chi^2(r - k)$ distribution.

- If $J_T(\hat{\theta})$ is far in the tail of the $\chi^2(r-k)$ distribution, we might conclude that the model is misspecified.

Asset Pricing with GMM

Suppose an agent derives utility from consumption, c_t, and seeks to maximze the discounted sum of expected utility:

$$\sum_{\tau=0}^{\infty} \beta^\tau E[u(c_{t+\tau}) \,|\, \Omega_t],$$

where $u(c_t)$ is the period utility function and satisfies:

$$\frac{\partial u(c_t)}{\partial c_t} > 0 \;\text{ and }\; \frac{\partial^2 u(c_t)}{\partial c_t^2} < 0.$$

Equilibrium Conditions

Suppose that the agent can purchase m assets paying gross returns $(1 + r_{i,t+1})$ between periods t and $t+1$, for $i = 1, \ldots, m$.

- The agent's portfolio must satisfy

$$u^{'}(c_t) = \beta E[(1 + r_{i,t+1}) u^{'}(c_{t+1}) \,|\, \Omega_t] \;\text{ for }\; i = 1, \ldots, m.$$

Equilibrium Conditions

The equilibrium conditions say that marginal utility of consuming an extra unit today should be equivalent to the expected marginal consumption gained by purchasing a unit of any asset.

- If these conditions didn't hold, the agent wouldn't be at an optimum.

The portfolio conditions can be rewritten as:

$$E\left[\left(\beta \frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1 + r_{i,t+1}) - 1\right) \middle| \Omega_t\right] = 0 \;\text{ for} i = 1, \ldots, m.$$

Equilibrium Conditions

Given a vector $\boldsymbol{x}_t \in \Omega_t$, by the law of iterated expectations

$$E\left[\underbrace{\left(\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1+r_{i,t+1})-1\right)\boldsymbol{x}_t}_{\boldsymbol{h}(\boldsymbol{\theta},\boldsymbol{y}_t)}\right] = E\left[E\left[\left(\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1+r_{i,t+1})-1\right)\boldsymbol{x}_t \middle| \Omega_t\right]\right]$$

$$E\left[\underbrace{E\left[\left(\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1+r_{i,t+1})-1\right) \middle| \Omega_t\right]}_{0}\boldsymbol{x}_t\right] = 0,$$

for $i = 1, \ldots, m$.

Stochastic Discount Factor

Economic theory says that all returns discounted by $\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}$ should be identical:

$$E\left[\underbrace{\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}}_{m_{t,t+1}}(1+r_{i,t+1})\right] = 1$$

$$\implies E[m_{t,t+1}(1+r_{i,t+1})] = 1.$$

- $\beta\frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1+r_{i,t+1})-1$ is a forecast error and should be uncorrelated with any variable $\boldsymbol{x}_t \in \Omega_t$

Casting as GMM

This problem maps easily into GMM where

$$\boldsymbol{y}_t = (r_{1,+1}, \ldots, r_{m,t+1}, c_t, c_{t+1}, \boldsymbol{x}_t^{'})^{'}$$

$$\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t) = \begin{bmatrix} \left(1 - \beta \frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1 + r_{i,t+1})\right)\boldsymbol{x}_t \\ \vdots \\ \left(1 - \beta \frac{u^{'}(c_{t+1})}{u^{'}(c_t)}(1 + r_{m,t+1})\right)\boldsymbol{x}_t \end{bmatrix}$$

$$\boldsymbol{g}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=0}^{T} \boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t).$$

Weighting Matrix for Asset Problem

Since the forecast errors in $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t)$ are unpredictable, they exhibit no serial correlation.

- Thus, $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{y}_t)$ exhibits no serial correlation.

This means S can be simply be estimated by

$$\hat{S}_T = \frac{1}{T}\sum_{t=0}^{T} \boldsymbol{h}(\hat{\boldsymbol{\theta}}, \boldsymbol{y}_t)\boldsymbol{h}(\hat{\boldsymbol{\theta}}, \boldsymbol{y}_t)^{'}.$$

Hansen and Singleton (1982)

Hansen and Singleton (1982) used GMM to estimate parameters of a model where

$$u(c_t) = \begin{cases} \frac{c_t^{1-\gamma}}{1-\gamma} & \text{for } \gamma > 0 \text{ and } \gamma \neq 1 \\ log(c_t) & \text{for } \gamma = 1 \end{cases}.$$

- In this case, $\boldsymbol{\theta} = (\beta, \gamma)^{'}$.

- Since forecast errors are uncorrelated with past returns and consumption, the lagged values of asset returns and aggregate consumption in $\boldsymbol{x}_t$.