Forecasting and Time Series Econometrics

Derek Li

Contents

1 Introduction

1.1 Basic Definition

Definition 1.1. A ***time series*** is a sequence of numerical values ordered by time.

Definition 1.2. A ***trend*** is a slow, smooth, long-run evolution of a time series over time.

Definition 1.3. A ***cycle*** is a periodic fluctuation of a time series, which may be seasonal or nonseasonal.

1.2 Notation

Description	Notation
Object to analyze - time series	$\{y_t\}$
Value at present time t - known value of the series	y_t
Future at time $t+h$ - random variable	Y_{t+h}
Value at future time $t+h$ - unknown value of the random variable	y_{t+h}
Collection of information - information set	$I_t = \{y_1, \cdots, y_t, x_1, \cdots, x_t\}$
Final objective - forecast h-step ahead	$f_{t,h}$
Uncertainty - forecast error	$e_{t,h} = y_{t+h} - f_{t,h}$

2 Statistics Review

2.1 Basic Definition

Definition 2.1. ***Population*** is the entire collection of elements about which information is desired.

Definition 2.2. ***Random process*** is the procedure involving a given population that can conceptually be repeated and leads to certain ***outcomes***. The outcome of a random process is a-priori uncertain.

Definition 2.3. ***Sample space*** is the set of all possible outcomes of the random process.

Definition 2.4. ***Random variable*** (r.v.) is the deterministic function from a sample space to the space of possible values of the variable ($\mathbb{R}$).

Definition 2.5. The ***cumulative distribution function*** (CDF) of r.v. X is

$$F_X(x) = P_X(X \leqslant x), \forall x \in \mathbb{R}.$$

Definition 2.6. The CDF of ***discrete r.v.*** is a step function, and its ***probability mass function*** (PMF) is

$$f_X(x_j) = P_X(X = x_j).$$

Definition 2.7. The CDF of ***continuous r.v.*** is a continuous function, and the Radon-Nikodym derivative of the CDF is the ***probability density function*** (PDF).

Note that $P_X(X = x) = 0$ if X is a continuous r.v.. The set $\{X = x\}$ is an example of a set of measure zero.

Definition 2.8. The ***expected value*** of a r.v. X is the weighted average of possible outcomes of X, where the weights are probabilities. In the discrete case,

$$\mathbb{E}[X] = \sum_x x f(x).$$

In the continuous case,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x.$$

Property 2.1.

$$\mathbb{E}\left[\sum_{i=1}^{K} a_i X_i\right] = \sum_{i=1}^{K} a_i \mathbb{E}[X_i],$$

where X is r.v., $a_i \in \mathbb{R}$.

Definition 2.9. The ***variance*** of a r.v. X is a measure of dispersion of X around its mean μ, denotes as $\mathrm{Var}[X]$ or σ_X^2, and defined as

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mu)^2].$$

In the discrete case,

$$\mathrm{Var}[X] = \sum_x (x - \mu)^2 f(x).$$

In the continuous cases,

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \mathrm{d}x.$$

Property 2.2. $\text{Var}[aX+bY] = a^2\text{Var}[X]+b^2\text{Var}[Y]+2ab\text{Cov}(X,Y)$, where X, Y are r.v.s., $a, b \in \mathbb{R}$.

Definition 2.10. The k^{th} ***moment*** (or ***non-central moment***) is

$$\kappa_k^* = \mathbb{E}[X^k].$$

Definition 2.11. The k^{th} ***central moment*** is

$$\kappa_k = \mathbb{E}[(X - \mathbb{E}[X])^k].$$

Note that the 0^{th} moment is 1, the 1^{st} non-central moment is mean, and the 2^{nd} moment is variance. We also use the 3^{rd} moment to define skewness, and the 4^{th} moment to define kurtosis. Moment of order $k > 2$ are typically called higher-order moment, and moment of order higher than a certain k may not exist for some distributions.

Definition 2.12. The ***standard deviation*** measures in the usual units:

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

Definition 2.13. ***Covariance*** measures the degree of joint variation of X and Y :

$$\text{Cov}(X,Y) = \sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Definition 2.14. ***Coefficient of correlation*** is the standardized covariance:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

where $-1 \leqslant \rho_{XY} \leqslant 1$.

Note that independence of X and Y implies $\rho_{XY} = 0$, but $\rho_{XY} = 0$ does not imply that $X \perp Y$.

Definition 2.15. The density of a r.v. Y conditional on the r.v. X taking on a specified value is called the ***conditional density*** of Y given X. In the discrete cases,

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}.$$

In the continuous cases,

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)}.$$

2.2 Linear Regression Model

Assumption 1 (Linearity)**.** $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$.

Assumption 2 (Zero Conditional Mean)**.** $\mathbb{E}[u|X_1, \cdots, X_k] = 0$.

Assumption 3 (Homoscedasticity)**.** $\text{Var}[u|X_1, \cdots, X_k] = \sigma_u^2$.

Assumption 4 (No Serial Correlation)**.** $\text{Cov}(u_t, u_{t-s}) = 0$ for $s = \pm 1, \pm 2, \cdots$.

Assumption 5 (No Perfect Collinearity)**.** There is no exact linear relationship among regressors.

Assumption 6 (Sample Variation in Regressors)**.** $\text{Var}[X_j] > 0$ for $j = 1, \cdots, k$.

Theorem 2.1 (Gauss-Markov Theorem)**.** Under A1 to A6, the ordinary least squares estimators are the best linear unbiased estimators (BLUE) of the unknown population regression coefficients.

Note that for time series data, A4 is typically not satisfied since A1 is too restrictive for time series featuring complex dynamics. Hence, for time series data we typically use other models than linear regression with OLS.

3 Statistics

3.1 Stochastic Process

Definition 3.1. A collection of r.v.s. $\{Y_t\} = Y_1, \cdots, Y_T$ indexed by time is called a ***stochastic process*** or a ***time-series process***.

Note that each r.v. Y_t in a stochastic process $\{Y_t\}$ is associated with a density function conditional on time.

Definition 3.2. An outcome of a stochastic process is a ***time series*** $\{y_t\} = y_1, \cdots, y_T$, where each $y_1, \cdots, y_T$ is an observed data point.

3.1.1 Stationarity

For each r.v. Y_t we only have one observation y_t and thus we cannot calculate estimated $\mathbb{E}[Y_t]$ or $\text{Var}[Y_t]$ from the observed data unless we impose further assumptions on $\{Y_t\}$, such as stationarity and ergodicity.

Definition 3.3. A stochastic process $\{Y_t\}$ is said to be ***first oder strongly stationary*** if all r.v.s. Y_t for all $t = 1, \cdots, T$ have the same probability density function.

Definition 3.4. A stochastic process $\{Y_t\}$ is said to be ***first order weakly stationary*** if all r.v.s. Y_t for all $t = 1, \cdots, T$ have the same mean, i.e.,

$$\mu_{Y_1} = \cdots = \mu_{Y_T}.$$

Definition 3.5. A stochastic process $\{Y_t\}$ is said to be ***second order weakly stationary*** or ***covariance stationary*** if all r.v.s. Y_t for all $t = 1, \cdots, T$ have the same mean and variance, and the covariances do not depend on t, i.e.,

$$\mu_{Y_1} = \cdots = \mu_{Y_T},\ \sigma^2_{Y_1} = \cdots = \sigma^2_{Y_T},\ \rho_{Y_t, Y_{t-k}} = \rho_{|k|}.$$

3.1.2 Transformation

We need stationarity so that averages can characterize the process moments.

Definition 3.6. A ***lag operator*** L applied to Y_t is defined by

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

Definition 3.7. The ***first difference operator*** Δ applied to Y_t is defined by

$$\Delta Y_t = Y_t - LY_t = Y_t - Y_{t-1}.$$

In many cases, for a non-stationary $\{Y_t\}$, $\{\Delta Y_t\}$ becomes first order weakly stationary and $\{\Delta \ln(Y_t)\}$ becomes second order weakly stationary.

3.2 Autocorrelation

3.2.1 Autocorrelation of a Covariance Stationary Process

Definition 3.8. The ***autocorrelation coefficient*** of order k is given by

$$\rho_{Y_t,Y_{t-k}} = \frac{\mathrm{Cov}(Y_t, Y_{t-k})}{\sqrt{\mathrm{Var}[Y_t]}\sqrt{\mathrm{Var}[Y_{t-k}]}}.$$

The ***autocorrelation function*** (ACF) is the mapping $\rho : k \to \rho_{Y_t,Y_{t-k}}$.

For a covariance stationary process, the ACF can be simplified:

$$\mathrm{Var}[Y_t] = \mathrm{Var}[Y_{t-k}] = \sigma^2$$

and

$$\mathrm{Cov}(Y_t, Y_{t-k}) = \sigma_k,$$

for all t, k. Thus, the autocorrelation coefficient of order k does not depend on t and can be expressed as

$$\rho_k = \frac{\sigma_k}{\sigma^2} := \frac{\gamma_k}{\gamma_0}.$$

Moreover, $\rho_k = \rho_{-k} = \rho_{|k|}$ and the ACF can be written as $\rho : k \to \rho_{|k|}$.

3.2.2 Partial Autocorrelation

$\mathrm{Cov}(Y_t, Y_{t+k})$ quantifies the covariance between the two end variables of the sequence, and thus the quantity $\mathrm{Cov}(Y_t, Y_{t+k})$ is influenced by the movement of any of the intermediary variables $Y_{t+1}, \cdots, Y_{t+k-1}$ that carry information from Y_t to T_{t+k}. The ***partial autocorrelation coefficient*** r_k only measures correlation between Y_t and Y_{t+k} while controlling for the influence of the intermediary variables.

The concept of r_k corresponds to a multiple regression coefficient:

k	Regression	r_k
1	$Y_{t+k} = \beta_0 + \beta_1 Y_{t+k-1} + \varepsilon_{t+k}$	β_1
2	$Y_{t+k} = \beta_0 + \beta_1 Y_{t+k-1} + \beta_2 Y_{t+k-2} + \varepsilon_{t+k}$	β_2
$\vdots$	$\vdots$	$\vdots$
k	$Y_{t+k} = \beta_0 + \beta_1 Y_{t+k-1} + \beta_2 Y_{t+k-2} + \cdots + \beta_k Y_t + \varepsilon_{t+k}$	β_k

3.2.3 Test for Autocorrelation

We test $H_0 : \rho_k = 0$ by the usual t statistic, and we test $H_0 : \rho_1 = \cdots = \rho_k = 0$ by the Ljung-Box Q statistic:

$$Q_k = T(T+2)\sum_{j=1}^{k} \frac{\widehat{\rho_j}^2}{T-j} \sim \chi^2_{(k)}.$$

4 Tools of the Forecaster

Before forecast, we need to determine the information set I_t, the forecast horizon h, and the loss function $L(e_{t,h})$.

4.1 Information Set

Definition 4.1. For a stochastic process $\{Y_t\}$, the ***information set*** I_t is the known historical time series of the process up to time t.

Any forecast $f_{t,h}$ is constructed as a function of the information set

$$f_{t,h} = g(I_t).$$

where $g(\cdot)$ is a time series model.

4.2 Forecast Horizon

- ***Covariance stationary processes*** or ***shor memory processes*** are useful for short-term forecasting but have limited usefulness for forecasting in the long term: more recent observation contains information far more relevant for the future than older information.
- ***Non-Stationary processes*** or ***long memory processes*** are useful for long-term forecasts: older information is as relevant for the forecast as more recent information.

4.2.1 Forecasting Environments

- Recursive scheme will be advantageous if the model is stable over time, but not if the data have structural breaks.
- Rolling scheme is robust against structural breaks in the data, but uses less information.
- Fixed scheme requires only one estimation of the model, but does not allow for parameter updating.

4.2.2 Loss Function

Definition 4.2. A ***loss function*** $L(e_{t,h})$ quantifies the costs associated with the forecast errors $e_{t,h}$, satisfying the following properties:

1. if the forecast error is zero, the loss is zero: $e_{t,h} = 0 \Rightarrow L(e_{t,h}) = 0$;
2. the loss function is non-negative;
3. for $e_{t,h} > 0, L(e_{t,h})$ is monotonically increasing and for $e_{t,h} < 0, L(e_{t,h})$ is monotonically decreasing.

- Symmetric loss functions:
  - Quadratic loss function: $L(e) = ae^2, a > 0$.
  - Absolute value loss function: $L(e) = a|e|, a > 0$.
- Asymmetric loss functions:
  - LINEX function: $L(e) = \exp(ae) - ae - 1, a \neq 0$.
  - Lin-Lin function: $L(e) = \begin{cases} a|e|, & e > 0 \\ b|e|, & e \leqslant 0 \end{cases}$.

4.3 Optimal Forecast

We have

$$L(e_{t,h}) = L(y_{t+h} - f_{t,h}).$$

Ideally, the forecaster would minimize $L(e_{t,h})$ as a criterion for making the forecast $f_{t,h}$, setting $f_{t,h} = y_{t+h}$, but y_{t+h} is unknown at time t.

Note that y_{t+h} is an outcome of the r.v. Y_{t+h} with a density function $f(y_{t+h}|I_t)$. Hence in practice, the forecaster will minimize the expected loss function

$$\mathbb{E}[L(y_{t+h} - f_{t,h})] = \int_{-\infty}^{\infty} L(y_{t+h} - f_{t,h}) f(y_{t+h}|I_t) \mathrm{d}y_{t+h}.$$

Definition 4.3. ***Optimal forecast*** $f^*_{t,h}$ minimizes the expected loss

$$\mathbb{E}[L(y_{t+h} - f_{t,h})] = \int_{-\infty}^{\infty} L(y_{t+h} - f_{t,h}) f(y_{t+h}|I_t) \mathrm{d}y_{t+h}.$$

We can assume that $f(y_{t+h}|I_t)$ is Gaussian with $\mu_{t+h|t} = \mathbb{E}[Y_{t+h}|I_t]$ and $\sigma^2_{t+h|t} = \mathrm{Var}[Y_{t+h}|I_t]$, and further specify a symmetric quadratic loss function as

$$L(e_{t,h}) = aet, h^2, a > 0.$$

Then,

$$f^*_{t,h} = \mu_{t+h|t}.$$

5 MA Process

5.1 Moving Average

We will construct time series models out of building blocks.

Definition 5.1. The ***white noise process***, $\{\varepsilon_t\}$ where each ε_t is defined as a random shock with the property that

$$\rho_k = 0, r_k = 0, k \geqslant 1,$$

i.e., both ACF and PACF are zero for all lags.

Note that $\{\varepsilon_t\}$ is a covariance stationary process.

Theorem 5.1 (Wold Decomposition Theorem)**.** If $\{Y_t\}$ is a covariance stationary process and $\{\varepsilon_t\}$ is a white noise zero-mean process, then there exists a unique linear representation

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

where V_t is a deterministic component and $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ is the stochastic component with $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

5.1.1 Model Components

The sequence $\{\varepsilon_t\}$ is called random shocks or innovations. Since $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, there must be a j from which all subsequent $\psi_{j+1}, \cdots$ are getting smaller s.t. the corresponding innovations $\varepsilon_{t-(j+1)}, \cdots$ have a negligible effect of Y_t. The deterministic component V_t can include a trend or cycle.

5.1.2 Lag Operator Representation

Now suppose $V_t = 0$, then

$$Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots = \varepsilon_t + \psi_1 L \varepsilon_t + \psi_2 L^2 \varepsilon_t + \cdots = (1 + \psi_1 L + \psi_2 L^2 + \cdots)\varepsilon_t := \Psi(L)\varepsilon_t,$$

where Ψ is a composite lag operator.

5.1.3 Model Approximation

We can approximate Y_t with

$$Y_t = \sum_{j=0}^{q} \psi_j \varepsilon_{t-j},$$

called the moving average of order q, denoted by MA(q). In practice, we will seek to have a good approximation of the dynamics in Y_t with few parameters, for a small q.

Property 5.1. The model approximation is accurate in the mean-square distance, i.e.,

$$\mathbb{E}\left[Y_t - \sum_{j=0}^{q} \psi_j \varepsilon_{t-j} \right]^2 \to 0$$

as $q \to \infty$.

Thus, the Wold decomposition guarantees that there always exists a linear model that can approximate the dynamics of a covariance stationary process.

5.2 Properties of MA(1)

Suppose an MA(q) process with a deterministic trend component μ :

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

5.2.1 Unconditional Moments

The MA(1) process is

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

The unconditional mean of Y_t is

$$\mathbb{E}[Y_t] = \mathbb{E}[\mu] + \mathbb{E}[\theta \varepsilon_{t-1}] + \mathbb{E}[\varepsilon_t] = \mu + 0 + 0 = \mu.$$

The unconditional variance of Y_t is

$$\begin{aligned}\mathrm{Var}[Y_t] &= \mathbb{E}[(Y_t - \mu)^2] = \mathbb{E}[(\theta \varepsilon_{t-1} + \varepsilon_t)^2] = \theta^2 \mathbb{E}[\varepsilon_{t-1}^2] + 2\theta \mathbb{E}[\varepsilon_{t-1}\varepsilon_t] + \mathbb{E}[\varepsilon_t^2] \\ &= \theta^2 \mathrm{Var}[\varepsilon_{t-1}] + 2\theta \mathrm{Cov}(\varepsilon_{t-1}, \varepsilon_t) + \mathrm{Var}[\varepsilon_t] = (1 + \theta^2)\sigma_\varepsilon^2.\end{aligned}$$

Since $\mathrm{Var}[Y_t]$ is time-invariant, we can denote it by

$$\mathrm{Var}[Y_t] = \gamma_0.$$

Since the first two moments of MA(1) do not depend on time, the MA(1) process is covariance stationary.

5.2.2 ACF

The autocovariance of MA(1) is

$$\begin{aligned}\gamma_1 &= \mathbb{E}[(Y_t - \mu)(Y_{t-1} - \mu)] = \mathbb{E}[(\theta \varepsilon_{t-1} + \varepsilon_t)(\theta \varepsilon_{t-2} + \varepsilon_{t-1})] = \mathbb{E}[\theta^2 \varepsilon_{t-1}\varepsilon_{t-2} + \theta \varepsilon_t \varepsilon_{t-2} + \theta \varepsilon_{t-1}^2 + \varepsilon_t \varepsilon_{t-1}] \\ &= \theta^2 \mathbb{E}[\varepsilon_{t-1}\varepsilon_{t-1}] + \theta \mathbb{E}[\varepsilon_t \varepsilon_{t-2}] + \theta \mathbb{E}[\varepsilon_{t-1}^2] + \mathbb{E}[\varepsilon_t \varepsilon_{t-1}] \\ &= \theta^2 \mathrm{Cov}(\varepsilon_{t-1}, \varepsilon_{t-2}) + \theta \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-2}) + \theta \mathrm{Var}[\varepsilon_{t-1}] + \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) \\ &= \theta \sigma_\varepsilon^2\end{aligned}$$

and for $k > 1$,

$$\begin{aligned}\gamma_k &= \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] = \mathbb{E}[(\theta \varepsilon_{t-1} + \varepsilon_t)(\theta \varepsilon_{t-k-1} + \varepsilon_{t-k})] \\ &= \theta^2 \mathbb{E}[\varepsilon_{t-1}\varepsilon_{t-k-1}] + \theta \mathbb{E}[\varepsilon_t \varepsilon_{t-k-1}] + \theta \mathbb{E}[\varepsilon_{t-1}\varepsilon_{t-k}] + \mathbb{E}[\varepsilon_t \varepsilon_{t-k}] = 0.\end{aligned}$$

The autocorrelation of MA(1) is then

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta \sigma_\varepsilon^2}{(1 + \theta^2)\sigma_\varepsilon^2} = \frac{\theta}{1 + \theta^2}$$

and for $k > 1$,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = 0.$$

5.2.3 PACF

PAC decreases towards zero with alternating signs.

5.3 Invertibility

Definition 5.2. We say that an MA(1) process is ***invertible*** if $|\theta| < 1$. If $|\theta| \geqslant 1$, the process is ***noninvertible***.

Invertibility of an MA process implies that we can always find an equivalent AR process.

Since

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

then

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

Recall that

$$\sum_{j=0}^{\infty} x^j = \frac{1}{1 - x}, |x| < 1,$$

and we let $x = (-\theta L), |\theta| < 1$:

$$\begin{aligned}\varepsilon_t &= (Y_t - \mu)\sum_{j=0}^{\infty}(-\theta L)^j = (Y_t - \mu)(1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \cdots) \\ &= (Y_t - \mu) - \theta(Y_{t-1} - \mu) + \theta^2(Y_{t-2} - \mu) - \theta^3(Y_{t-3} - \mu) + \cdots .\end{aligned}$$

Solving for Y_t :

$$Y_t - \mu = \theta(Y_{t-1} - \mu) - \theta^2(Y_{t-2} - \mu) + \cdots ,$$

which is equivalent to an AR(∞) process.

But for $|\theta| \geqslant 1$, we are not able to obtain the AR representation. Note that

$$r_1 = \rho_1 = \frac{\theta}{1 + \theta^2}.$$

Hence for each pair of MA(1) processes with $\theta_{(1)}$ and $\theta_2(2)$ s.t. $\theta_{(1)} = \frac{1}{\theta_{(2)}}$, which yield identical ACF and PACF and hence dynamics for Y_t, we choose the invertible MA representation with $|\theta| < 1$.

5.4 Forecasting

For constructing the optimal forecast with an MA(1) model, we will choose a quadratic loss function, then the optimal forecast is equal to the conditional expectation

$$f_{t,h} = \mu_{t+h|t} = \mathbb{E}[Y_{t+h}|I_t].$$

Note that at time t the value of ε_t is known and hence $\varepsilon_t \in I_t$, but any future $\varepsilon_{t+1}, \cdots$ is unknown.

5.4.1 Forecast for $h = 1$

For $h = 1$,

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] = \mathbb{E}[\mu + \theta\varepsilon_t + \varepsilon_{t+1}|I_t] = \mathbb{E}[\mu|I_t] + \theta\mathbb{E}[\varepsilon_t|I_t] + \mathbb{E}[\varepsilon_{t+1}|I_t] = \mu + \theta\varepsilon_t.$$

The value of ε_t is retrieved from data $\{y_t\} \subset I_t$ using

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

which is guaranteed to exist due to invertibility of the MA process.

The forecast error is

$$e_{t,1} = Y_{t+1} - f_{t,1} = \varepsilon_{t+1}.$$

The forecast uncertainty is

$$\sigma^2_{t+1|t} = \mathrm{Var}[Y_{t+1}|I_t] = \mathbb{E}[(Y_{t+1} - f_{t,1})^2|I_t] = \mathbb{E}[\varepsilon^2_{t+1}] = \sigma^2_\varepsilon.$$

If we assume $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ then the density forecast is

$$\mathcal{N}(\mu_{t+1|t}, \sigma^2_{t+1|t}) = \mathcal{N}(\mu + \theta\varepsilon_t, \sigma^2_\varepsilon),$$

from which we can construct the 95% CI:

$$[\mu + \theta\varepsilon_t - 1.96\sigma_\varepsilon, \mu + \theta\varepsilon_t + 1.96\sigma_\varepsilon].$$

5.4.2 Forecast for $h > 1$

For $h = 2$,

$$f(t, 2) = \mathbb{E}[Y_{t+2}|I_t] = \mathbb{E}[\mu + \theta\varepsilon_{t+1} + \varepsilon_{t+2}|I_t] = \mu.$$

In general, for $h > 1$, the MA(1) process does not predict any time dependence beyond the unconditional mean, consistent with $\rho_k = 0$ for $k > 1$. Hence, MA(1) is a process with a very short memory.

Besides,

$$e_{t,2} = Y_{t+2} - f_{t,2} = \theta\varepsilon_{t+1} + \varepsilon_{t+2}$$

and

$$\sigma^2_{t+2|t} = \mathbb{E}[(\theta\varepsilon_{t+1} + \varepsilon_{t+2})^2|I_t] = (1 + \theta^2)\sigma^2_\varepsilon = \mathrm{Var}[Y_t],$$

which are consistent with the short memory of MA(1).

Assuming normality of ε_t, the density forecast is $\mathcal{N}(\mu, \sigma^2_Y)$, which is the unconditional density of Y_t. MA(1) will yield the same prediction as in $h = 2$ for $h > 2$.

5.5 Properties of MA(q)

5.5.1 Unconditional Moments

The MA(2) process is

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

The unconditional mean and variance are

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

and

$$\mathrm{Var}[Y_t] = (1 + \theta^2_1 + \theta^2_2)\sigma^2_\varepsilon.$$

5.5.2 ACF

The autocovariance γ_1 is

$$\gamma_1 = \mathbb{E}[(Y_t - \mu)(Y_{t-1} - \mu)] = (\theta_1 + \theta_1\theta_2)\sigma_\varepsilon^2,$$

and the autocorrelation ρ_1 is

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}.$$

The autocovariance γ_2 is

$$\gamma_2 = \mathbb{E}[(Y_t - \mu)(Y_{t-2} - \mu)] = \theta_2\sigma_\varepsilon^2,$$

and the autocorrelation ρ_2 is

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}.$$

For $k > 2, \rho_k = 0$. For a general MA(q) process, $\rho_k = 0$ if $k > q$. We conclude that MA(q) process is covariance stationary.

5.5.3 PACF and Invertibility

The PACF r_k decays towards zero, but there is no cut-off point after which r_k would always equal to zero as in the ACF cases.

We need to choose between two MA representations that generate the same ACF and PACF and we choose the invertible representation.

5.5.4 Forecasting

Assuming quadratic loss function, the optimal forecast is

$$\begin{aligned} f_{t,1} &= \mu + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1}, \\ f_{t,2} &= \mu + \theta_2\varepsilon_t, \\ f_{t,h} &= \mu, h > 2, \end{aligned}$$

which is consistent with the short memory property of MA process.

6 AR Process

6.1 AR Model

We would like to fit on the data a time series model that can closely approximate the dynamic pattern in the data. Hence, we need a model with the given ACF pattern. The autoregressive (AR) model can generate such ACF and hence will be a suitable fit to the data.

Definition 6.1. An ***autoregressive model*** of order p, denoted AR(p) is given by

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

where $\{\varepsilon_t\}$ is the white noise process.

6.2 AR(1)

The model is

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

where the parameter ϕ is called the persistence parameter since it influences the persistence of the series.

Example 6.1. The series with $\phi = 0.95$ stays longer above or below the unconditional mean than the series with $\phi = 0.4$. The series with $\phi = 1$ is extremely persistent and non-stationary.

Property 6.1. AR(1) is stationary only for $|\phi| < 1$.

6.2.1 ACF and PACF

Note that

$$\rho_1 = r_1 = \phi.$$

The ACF decreases exponentially towards zero, with faster decay for smaller ϕ. $r_1 \neq 0$ but $r_k = 0$ for $k > 1$. The same features hold for negative ϕ.

6.2.2 Forecast for $h = 1$

For constructing the optimal forecast with an AR(1) model, we will choose a quadratic loss function, then the optimal forecast is equal to the conditional expectation:

$$f_{t,h} = \mu_{t+h|t} = \mathbb{E}[Y_{t+h}|I_t].$$

For the forecasting horizon $h = 1$,

$$Y_{t+1} = c + \phi Y_t + \varepsilon_{t+1}.$$

Since $Y_t \in I_t$,

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] = \mathbb{E}[c + \phi Y_t + \varepsilon_{t+1}|I_t] = \mathbb{E}[c] + \mathbb{E}[\phi Y_t] + \mathbb{E}[\varepsilon_{t+1}|I_t] = c + \phi Y_t.$$

The one-period ahead forecast error is

$$e_{t,1} = Y_{t+1} - f_{t,1} = \varepsilon_{t+1}.$$

The forecast variance is

$$\sigma^2_{t+1|t} = \mathrm{Var}[Y_{t+1}|I_t] = \mathrm{Var}[\varepsilon_{t+1}] = \sigma^2_\varepsilon.$$

Assume $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon)$, the density forecast is

$$f(Y_{t+1}|I_t) = \mathcal{N}(\mu_{t+1|t}, \sigma^2_{t+1|t}) = \mathcal{N}(c + \phi Y_t, \sigma^2_\varepsilon).$$

The 95% CI is then

$$[\mu_{t+1|t} - 1.96\sigma_{t+1|t}, \mu_{t+1|t} + 1.96\sigma_{t+1|t}].$$

6.2.3 Forecasts for $h > 1$

The optimal forecast for $h = s > 1$ is

$$f_{t,2} = \mathbb{E}[Y_{t+2}|I_t] = \mathbb{E}[c + \phi Y_{t+1} + \varepsilon_{t+2}|I_t] = c(1+\phi) + \phi^2 Y_t$$

and

$$f_{t,h} = \mathbb{E}[Y_{t+s}|I_t] = \mathbb{E}[c + \phi Y_{t+s-1} + \varepsilon_{t+s}|I_t] = c(1 + \phi + \cdots + \phi^{s-1}) + \phi^s Y_t.$$

Thus,

$$\text{Var}[Y_{t+s}|I_t] = \sigma_\varepsilon^2(1 + \phi^2 + \phi^4 + \cdots + \phi^{2(s-1)}).$$

6.2.4 Forecasts for $h \to \infty$

As $h \to \infty$, the forecast converges to

$$f_{t,\infty} = c(1 + \phi + \phi^2 + \cdots) = \frac{c}{1-\phi},$$

which is the unconditional mean of $\{Y_t\}$, and

$$\sigma^2_{t+\infty|t} = \sigma_\varepsilon^2(1 + \phi^2 + \phi^4 + \cdots) = \frac{\sigma_\varepsilon^2}{1-\phi^2},$$

which is the unconditional variance of $\{Y_t\}$.

Hence, AR(1) model is suitable for forecasts in the short to medium term: convergence of its forecasts to unconditional moments still indicates short memory of the process, albeit relatively longer than for MA(1). Note that these results hold only for stationary AR(1) with $|\phi| < 1$.

6.3 AR(2)

The model is

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

where the persistence measure is $\phi_1 + \phi_2$.

6.3.1 AR(2) Stationarity

For AR(2) to be covariance stationary, the parameter vector (ϕ_1, ϕ_2) must satisfy:
(1) The necessary conditions: $-1 < \phi_2 < 1$ and $-2 < \phi_1 < 2$;
(2) The sufficient conditions: $\phi_1 + \phi_2 < 1$ and $\phi_2 - \phi_1 < 1$.
Note that we first check the necessary and then the sufficient conditions.

6.3.2 Unconditional Moments of AR(2)

The unconditional mean of AR(2) is

$$\mathbb{E}[Y_t] = c + \phi_1 \mathbb{E}[Y_{t-1}] + \phi_2 \mathbb{E}[Y_{t-2}] + \mathbb{E}[\varepsilon_t].$$

Since the process is stationary, then $\mu = c + \phi_1\mu + \phi_2\mu$ and thus

$$\mu = \frac{c}{1 - \phi_1 - \phi_2}.$$

6.3.3 ACF and PACF

Note that

$$\rho_1 = r_1, r_2 = \phi_2 + \text{Sampling error.}$$

ACF decays to zero relatively slowly and PACF has two significant spikes $r_1 \neq 0$ and $r_2 \neq 0$, and then $r_k = 0$ for $k > 2$.

6.3.4 Forecasting for $h = 1$

Using the quadratic loss function for $h = 1$,

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] = \mathbb{E}[c + \phi_1 Y_t + \phi_2 Y_{t-1} + \varepsilon_{t+1}|I_t] = c + \phi_1 Y_t + \phi_2 Y_{t-1}.$$

Furthermore,

$$e_{t,1} = Y_{t+1} - f_{t,1} = \varepsilon_{t+1}$$

and

$$\sigma^2_{t+1|t} = \text{Var}[Y_{t+1}|I_t] = \mathbb{E}[\varepsilon^2_{t+1}] = \sigma^2_\varepsilon.$$

6.3.5 Forecasting for $h = 2$

For $h = 2$,

$$f_{t,2} = \mathbb{E}[Y_{t+2}|I_t] = \mathbb{E}[c + \phi_1 Y_{t+1} + \phi_2 Y_t + \varepsilon_{t+2}|I_t] = c + \phi_1 f_{t,1} + \phi_2 Y_t.$$

Furthermore,

$$\begin{aligned} e_{t,2} &= Y_{t+2} - f_{t,2} = c + \phi_1 Y_{t+1} + \phi_2 Y_t + \varepsilon_{t+2} - (c + \phi_1 f_{t,1} + \phi_2 Y_t) \\ &= \phi_1(Y_{t+1} - f_{t,1}) + \varepsilon_{t+2} = \phi_1 e_{t,1} + \varepsilon_{t+2}, \end{aligned}$$

and

$$\sigma_{t+2|t} = \text{Var}[Y_{t+2}|I_t] = \phi_1^2 \text{Var}[Y_{t+1}|I_t] + \sigma^2_\varepsilon = \sigma^2_\varepsilon(1 + \phi_1^2) = \mathbb{E}[e^2_{t,2}].$$

6.3.6 Forecasting for $h = s$

For $h = s$ with $s > 2$,

$$\begin{aligned} f_{t,s} &= \mathbb{E}[Y_{t+s}|I_t] = c + \phi_1 f_{t,s-1} + \phi_2 f_{t,s-2}, \\ e_{t,s} &= Y_{t+s} - f_{t,s} = \varepsilon_{t+s} + \phi_1 e_{t,s-1} + \phi_2 e_{t,s-2}, \end{aligned}$$

and

$$\sigma^2_{t+s|t} = \text{Var}[Y_{t+s}|I_t] = \mathbb{E}[e^2_{t,s}].$$

AR(2) is classified as short memory process: as $s \to \infty$, $f_{t,s} \to \mu$ and $\sigma^2_{t+s|t} \to \sigma^2_Y$.

6.4 AR(p)

The model is

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

The chain rule of forecasting calculates multi-step forecasts $f_{t,s}$ using previous forecasts:

$$\begin{aligned} f_{t,1} &= c + \phi_1 Y_t + \phi_2 Y_{t-1} + \cdots + \phi_p Y_{t+1-p} \\ f_{t,2} &= c + \phi_1 f_{t,1} + \phi_2 Y_t + \cdots + \phi_p Y_{t+2-p} \\ &\vdots \\ f_{t,s} &= c + \phi_1 f_{t,s-1} + \phi_2 f_{t,s-2} + \cdots + \phi_p f_{t,s-p}, s > p. \end{aligned}$$

6.5 ARMA

6.5.1 Lag Operator Representation

Recall the Wold decomposition of a covariance stationary process $\{Y_t\}$ by

$$Y_t = \Psi(L)\varepsilon_t.$$

The infinite lag polynomial in $\Psi(L)$ can be well approximated by

$$\Psi(L) \approx \frac{\Theta_q(L)}{\Phi_p(L)}$$

with small p and q where

$$\begin{aligned}\Theta_q(L) &= (1 + \theta_1 L + \cdots + \theta_q L^q),\\ \Phi_p(L) &= (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p).\end{aligned}$$

6.5.2 ARMA Model

We write

$$Y_t = \Psi(L)\varepsilon_t \approx \frac{\Theta_q(L)}{\Phi_p(L)}\varepsilon_t,$$

or equivalently,

$$\Phi_p(L)Y_t = \Theta_q(L)\varepsilon_t \Rightarrow Y_t = \underbrace{\phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}}_{\text{AR}(p)} + \underbrace{\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}}_{\text{MA}(q)},$$

where the AR(p) part is the autoregressive component and the MA(q) part is the moving average component of the resulting ARMA(p, q) model of Y_t. The number of $p + q$ is small, relative to ∞ in teh full Wold representation.

6.6 Seasonal Cycles

6.6.1 Deterministic Seasonality

Suppose we have quarterly data on Y_t, and Y_t differ seasonally by quarter, we can specify the regression model

$$Y_t = \beta_1 Q1_t + \beta_2 Q2_t + \beta_3 Q3_t + \beta_4 Q4_t + \varepsilon_t,$$

where $Q1_t, Q2_t, Q3_t$ and $Q4_t$ are quarterly dummy variables. This model expresses deterministic seasonality, whereby the regressors are always exactly predictable.

6.6.2 Stochastic Seasonality

In model with stochastic seasonality, the seasonal component is driven by r.v.s..

A seasonal AR(1) model is given by

$$Y_t = c + \phi_s Y_{t-s} + \varepsilon_t,$$

where the explanatory variable is random, the model seeks to explain the dynamics across seasons, the subscript s refers to data frequency ($s = 4$ for quarterly data, $s = 12$ for monthly data, etc) and the coefficient ϕ_s is the same across all seasons.

A seasonal AR(p) model, S – AR(p), is given by

$$Y_t = c + \phi_s Y_{t-s} + \phi_{2s} Y_{t-2s} + \cdots + \phi_{ps} Y_{t-ps} + \varepsilon_t,$$

which can be expressed as

$$(1 - \phi_s L^s + \phi_{2s} L^{2s} - \cdots - \phi_{ps} L^{ps}) Y_t = c + \varepsilon_t.$$

Example 6.2 (S – AR(2) for Quarterly Data)**.** $Y_t = c + \phi_4 Y_{t-4} + \phi_8 Y_{t-8} + \varepsilon_t$.

A seasonal MA(q) model, S – MA(q), is given by

$$Y_t = \mu + \theta_s + \varepsilon_{t-s} + \theta_{2s} \varepsilon_{t-2s} + \cdots + \theta_{qs} + \varepsilon_{t-qs} + \varepsilon_t.$$

Example 6.3 (S – MA(2) for Quarterly Data)**.** $Y_t = \mu + \theta_4 \varepsilon_{t-4} + \theta_8 \varepsilon_{t-8} + \varepsilon_t$.

Example 6.4 (S – ARMA$(1,2)$)**.** $Y_t = c + \phi_4 Y_{t-4} + \varepsilon_t + \theta_4 \varepsilon_{t-4} + \theta_8 \varepsilon_{t-8}$ or $(1 - \phi_4 L^4) Y_t = c + (1 + \theta_4 L^4 + \theta_8 L^8) \varepsilon_t$.

Example 6.5 (Combining S – ARMA$(1,2)$ and ARMA$(2,1)$)**.** For ARMA$(2,1)$, the model is $Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$ or $(1 - \phi_1 L - \phi_2 L^2) Y_t = c + (1 + \theta_1 L) \varepsilon_t$. Then the combining model is

$$(1 - \phi_4 L^4)(1 - \phi_1 L - \phi_2 L^2) Y_t = c + (1 + \theta_4 L^4 + \theta_8 L^8)(1 + \theta_1 L) \varepsilon_t.$$