

Forecasting and Time Series Econometrics

ECO374 Winter 2019

Tianyu Du

February 14, 2019

Contents

1	Introduction and Statistics Review	2
2	Statistics and Time Series	3
2.1	Stochastic Processes	3
2.2	Auto-correlations	3
2.3	Causality and Inevitability	5
3	Forecasting Tools	6
3.1	Information Set	6
3.2	Forecast Horizon	6
3.2.1	Forecasting Environments: <i>Recursive</i>	7
3.2.2	Forecasting Environments: <i>Rolling</i>	7
3.2.3	Forecasting Environments: <i>Fixed</i>	8
3.3	Loss Function	8
3.4	Optimal Forecast	9
4	Moving Average Process	10
4.1	Wold Decomposition Theorem	10
4.2	Moving Average Process	10
4.3	Forecasting with MA(1)	12
4.3.1	Forecasting with Horizon $h = 1$	12
4.3.2	Forecasting with Horizon $h = 2$	12
4.4	Properties of MA(2) Process	13
5	Auto-Regression Process and Seasonality	13
5.1	AR Process	13
5.1.1	Forecasting with AR(1) and $h = 1$	14
5.1.2	Forecasting with AR(1) and $h = s > 1$	14
5.1.3	Forecasting with AR(1) and $h \rightarrow \infty$	14
5.1.4	Forecasting with AR(2) process	15
5.1.5	AP(p) process	16
5.2	Seasonality	16
5.2.1	Deterministic Seasonality	16
5.2.2	Stochastic Seasonality	17

6 Model Assessment and Asymmetric Loss	17
6.1 Model Assessment	17
6.2 Asymmetric Loss	18

1 Introduction and Statistics Review

Definition 1.1. Given random variable X , the k^{th} **non-central moment** is defined as

$$\mathbb{E}[X^k] \quad (1.1)$$

Definition 1.2. Given random variable X , the k^{th} **central moment** is defined as

$$\mathbb{E}[(X - \mathbb{E}[X])^k] \quad (1.2)$$

Remark 1.1. Moments of order higher than a certain k may not exist for certain distribution.

Definition 1.3. Given the **joint density** $f(X, Y)$ of two *continuous* random variables, the **conditional density** of random Y conditioned on X is

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)} \quad (1.3)$$

Definition 1.4. Given discrete variables X and Y , the **conditional density** of Y conditioned on X is defined as

$$P(Y = y|X = x) = \frac{P(Y = y \wedge X = x)}{P(X = x)} \quad (1.4)$$

Assumption 1.1. Assumptions on linear regression on time series data:

(i) **Linearity**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \quad (1.5)$$

(ii) **Zero Conditional Mean**

$$\mathbb{E}[u|X_1, X_2, \dots, X_k] = 0 \quad (1.6)$$

(iii) **Homoscedasitcity**

$$\mathbb{V}[u|X_1, X_2, \dots, X_k] = \sigma_u^2 \quad (1.7)$$

(iv) **No Serial Correlation**

$$Cov(u_t, u_s) = 0 \quad \forall t \neq s \in \mathbb{Z} \quad (1.8)$$

(v) **No Perfect Collinearity**

(vi) **Sample Variation in Regressors**

$$\mathbb{V}[X_j] > 0 \quad \forall j \quad (1.9)$$

Theorem 1.1 (Gauss-Markov Theorem). Under assumptions 1.1, the OLS estimators $\hat{\beta}_j$ are *best linear unbiased estimators* of the unknown population regression coefficients β_j .

Remark 1.2. The *no serial correlation* assumption is typically not satisfied for time series data. And the *linearity* assumption is also too restrictive for time series featuring complex dynamics. Hence, for time series data we typically use other models than linear regression with OLS.

2 Statistics and Time Series

2.1 Stochastic Processes

Definition 2.1 (1.1). A **stochastic process** (or **time series process**) is a family (collection) random variables indexed by $t \in \mathcal{T}$ and defined on some given probability space (Ω, \mathcal{F}, P) .

$$\{Y_t\} = Y_1, \dots, Y_T \quad (2.1)$$

Definition 2.2 (1.2). The function $t \rightarrow y_t$ which assigns to each point in time $t \in \mathcal{T}$ the realization of the random variable Y_t , y_t is called a **realization** or a **trajectory** or an **outcome** of the stochastic process.

Definition 2.3. An *outcome* of a stochastic process

$$\{y_t\} = y_1, \dots, y_T \quad (2.2)$$

is a **time series**.

Definition 2.4 (1.3). A **time series model** or a **model** for the observations (data), $\{y_t\}$, is a specification of the *joint distribution* of $\{Y_t\}$ for which $\{y_t\}$ is a realization.

Assumption 2.1. The **ergodicity** assumption requires the observations cover in principle all possible events.

Definition 2.5. A stochastic process $\{Y_t\}$ is **first order strongly stationary** if all random variables $Y_t \in \{Y_t\}$ has the *same probability density function*.

Definition 2.6 (1.7). A stochastic process $\{Y_t\}$ is **strictly stationary** if for all $h, n \geq 1$, (X_1, \dots, X_n) and $(X_{1+h}, \dots, X_{n+h})$ have the same distribution.

Definition 2.7. A stochastic process $\{Y_t\}$ is **first order weakly stationary** if

$$\forall t \in \mathcal{T}, \mu_{Y_t} \equiv \mathbb{E}[Y_t] = \bar{\mu} \quad (2.3)$$

Definition 2.8. A stochastic process $\{Y_t\}$ is **second order weakly stationary**, or **covariance stationary** if all random variables $\{Y_t\}$ have the same mean and variance. And the covariances do not depend on t . That's, for all $t \in \mathcal{T}$,

- (i) $\mathbb{E}[Y_t] = \mu \forall t$
- (ii) $\mathbb{V}[Y_t] = \sigma^2 < \infty \forall t$
- (iii) $Cov(Y_t, Y_s) = Cov(Y_{t+r}, Y_{s+r}) \forall t, s, r \in \mathbb{Z}$

2.2 Auto-correlations

Definition 2.9. Let $\{Y_t\}$ be a stochastic process with $\mathbb{V}[Y_t] < \infty \forall t \in \mathcal{T}$, the **auto-covariance function** is defined as

$$\gamma_Y(t, s) \equiv Cov(Y_t, Y_s) \quad (2.4)$$

$$= \mathbb{E}[(Y_t - \mathbb{E}[Y_t])(Y_s - \mathbb{E}[Y_s])] \quad (2.5)$$

$$= \mathbb{E}[Y_t Y_s] - \mathbb{E}[Y_t] \mathbb{E}[Y_s] \quad (2.6)$$

Lemma 2.1. If $\{Y_t\}$ is stationary, then the auto-covariance function does not depend on specific time point t . We can write the $h \in \mathbb{Z}$ degree auto-covariance as

$$\gamma_Y(h) \equiv \gamma_X(t, t+h) \quad \forall t \in \mathcal{T} \quad (2.7)$$

Proposition 2.1. By the symmetry of covariance,

$$\gamma_Y(h) = \gamma_Y(-h) \quad (2.8)$$

Definition 2.10. The **auto-correlation coefficient** of order k is given by

$$\rho_{Y_t, Y_{t-k}} = \frac{\text{Cov}(Y_t, Y_{t-k})}{\sqrt{\mathbb{V}[Y_t]} \sqrt{\mathbb{V}[Y_{t-k}]}} \quad (2.9)$$

Definition 2.11. Let $\{Y_t\}$ be a *stationary process* and the **auto-correlation function** (ACF) is a mapping from *order* of auto-correlation coefficient to the coefficient $\rho_Y : k \rightarrow \rho_{Y_t, Y_{t-k}}$, defined as

$$\rho_Y(k) \equiv \frac{\gamma(k)}{\gamma(0)} = \text{corr}(Y_{t+k}, Y_t) \quad (2.10)$$

Proposition 2.2. Note that

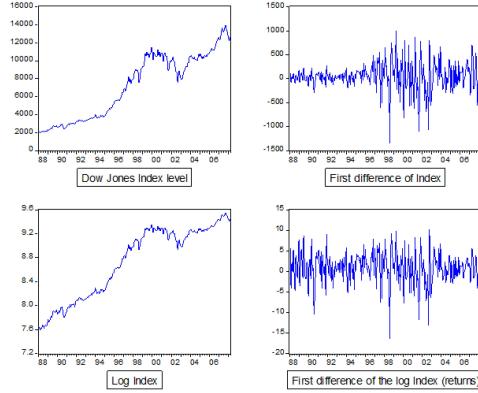
$$\rho_k = \rho_{-k} = \rho_{|k|} \quad (2.11)$$

so the ACF for stationary process can be simplified to a mapping

$$\rho : k \rightarrow \rho_{|k|} \quad (2.12)$$

Remark 2.1. Strong stationarity is difficult to test so we will focus on weak stationarity only.

Proposition 2.3. For a non-stationary stochastic process $\{Y_t\}$, $\{\Delta Y_t\}$ becomes *first order weakly stationary* and $\{\Delta \log(Y_t)\}$ becomes *second order weakly stationary*.



Definition 2.12 (1.8). A stochastic process $\{Y_t\}$ is called a **Gaussian process** if all *finite dimensional* distribution from the process are multivariate normally distributed. That's

$$\forall n \in \mathbb{Z}_{>0}, \quad \forall (t_1, \dots, t_n) \in \mathcal{T}^n, \quad (Y_{t_1}, \dots, Y_{t_n}) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad (2.13)$$

Notation 2.1. Consider the problem of forecasting Y_{T+1} from observations $\{Y_t\}_{t=1}^T$, the *best linear predictor* is denoted as

$$\mathbb{P}_T Y_{T+1} = \sum_{i=1}^T a_i L^i Y_{T+1} \quad (2.14)$$

And Y_{T+1} can be expressed as

$$Y_{T+1} = \mathbb{P}_T Y_{T+1} + Z_{T+1} \quad (2.15)$$

where Z_{T+1} denotes the forecast error which is *uncorrelated* with X_T, \dots, X_1 .

Definition 2.13 (3.3). The **partial auto-correlation function** (PACT) $\alpha(h)$ with $h \in \mathbb{Z}_{\geq 0}$ of a *stationary* process is defined as

$$\alpha(0) = 1 \quad (2.16)$$

$$\alpha(1) = \text{corr}(Y_2, Y_1) = \rho(1) \quad (2.17)$$

$$\alpha(h) = \text{corr}\left(Y_{h+1} - \mathbb{P}(Y_{h+1}|1, Y_2, \dots, Y_h), X_1 - \mathbb{P}(Y_1|1, Y_2, \dots, Y_h)\right) \quad (2.18)$$

Remark 2.2 (Interpretation of PACF). partial auto-correlation r_k only measures correlation between two variables Y_t and Y_{t+k} while controlling $(Y_{t+1}, \dots, Y_{t+k-1})$.

Remark 2.3. Properties of ACF and PACF

processes	ACF	PACF
$AR(p)$	Declines exponentially (monotonic or oscillating) to zero	$\alpha(h) = 0 \forall h > p$
$MA(q)$	$\rho(h) = 0 \forall h > q$	Declines exponentially (monotonic or oscillating) to zero

Test for Auto-correlation To test single auto-correlation with

$$H_0 : \rho_k = 0 \quad (2.19)$$

we can use usual t-statistic. While testing the joint hypothesis

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_k = 0 \quad (2.20)$$

we are using the **Ljung-Box Q-statistic**:

$$Q_k = T(T+1) \sum_{j=1}^k \frac{\hat{\rho}_j^2}{T-j} \sim \chi_k^2 \quad (2.21)$$

2.3 Causality and Inevitability

Definition 2.14 (Causality). An $ARMA(p, q)$ process $\{Y_t\}$ with

$$\Phi(L)Y_t = \Theta(L)\varepsilon_t \quad (2.22)$$

is called **causal with respect to** $\{\varepsilon_t\}$ if there exists a sequence $\{\psi_j\}$ with the property $\sum_j^\infty |\psi_j| < \infty$ such that

$$Y_t = \Psi(L)\varepsilon_t \quad \text{with } \psi_0 = 1 \quad (2.23)$$

where $\Psi(L) \equiv \sum_{j=0}^\infty \psi_j L^j$. The above equation is referred to as the **causal representation** of $\{Y_t\}$ with respect to $\{\varepsilon_t\}$.

Theorem 2.1. Let $\{Y_t\}$ be an $ARMA(p, q)$ process with

$$\Phi(L)Y_t = \Theta(L)\varepsilon_t \quad (2.24)$$

such that polynomials $\Phi(z)$ and $\Theta(z)$ have no common roots.

Then $\{Y_t\}$ is causal with respect to $\{\varepsilon_t\}$ if and only if all roots of the equation $\Phi(z) = 0$ are outside the unit circle. The coefficients Ψ are then uniquely defined by identity

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)} \quad (2.25)$$

Definition 2.15 (Invertibility). An $ARMA(p, q)$ process for $\{Y_t\}$ satisfying

$$\Phi(L)Y_t = \Theta(L)\varepsilon_t \quad (2.26)$$

is called **invertible** with respect to $\{\varepsilon\}$ if and only if there exists a sequence Π with the property $\sum_{j=0}^{\infty} |\pi_j| < \infty$ such that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j L^j Y_t \quad (2.27)$$

Theorem 2.2. Let $\{Y_t\}$ be an $ARMA(p, q)$ process with

$$\Phi(L)Y_t = \Theta(L)\varepsilon_t \quad (2.28)$$

Then $\{Y_t\}$ is invertible with respect to $\{\varepsilon_t\}$ if and only if all roots of $\Theta(z)$ are outside the unit circle. And the coefficients of Π are then uniquely determined by the relation

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\Phi(z)}{\Theta(z)} \quad (2.29)$$

3 Forecasting Tools

3.1 Information Set

Definition 3.1. For stochastic process $\{Y_t\}$, the **information set** I_t is the *known time series* up to time t .

Definition 3.2. A **forecast** $f_{t,h}$ the image of a *time series model* g under given information set I_t . Specifically,

$$f_{t,h} = g(I_t) \in \Omega \quad (3.1)$$

3.2 Forecast Horizon

Remark 3.1. Covariance stationary processes are **short-memory processes**. More recent observation contains information far more relevant for the future than older information.

Remark 3.2. Non-stationary processes are **long-memory processes** and older information is as relevant for the forecast as more recent information.

3.2.1 Forecasting Environments: *Recursive*

- (i) *Updating* with *flexible* information set.
- (ii) Advantageous if model is *stable over time*.
- (iii) Not robust to *structural break*.

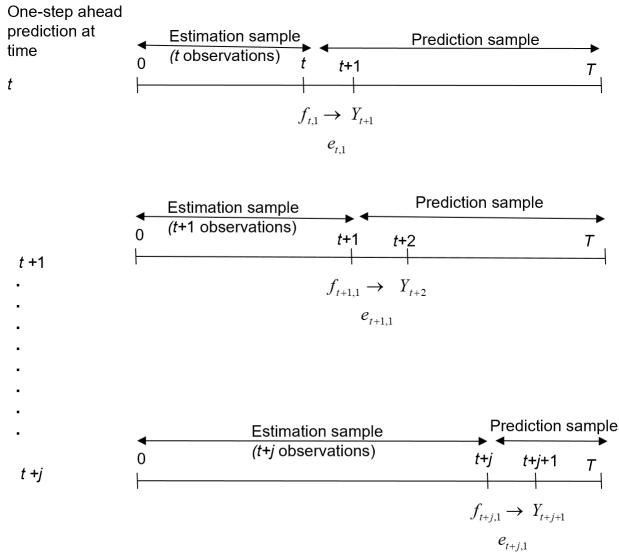


Figure 3.1: Recursive Forecasting Scheme

3.2.2 Forecasting Environments: *Rolling*

- (i) *Updating* with *fixed-size* information set.
- (ii) Robust against *structural breaks*.
- (iii) Not fully exploit information available.

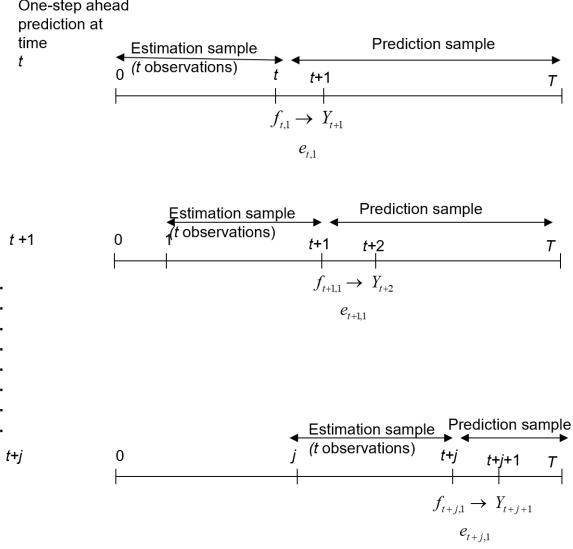


Figure 3.2: Rolling Forecasting Scheme

3.2.3 Forecasting Environments: *Fixed*

- (i) One estimation and forecast with *fixed-size but updated information set*.
- (ii) Computationally cheap.

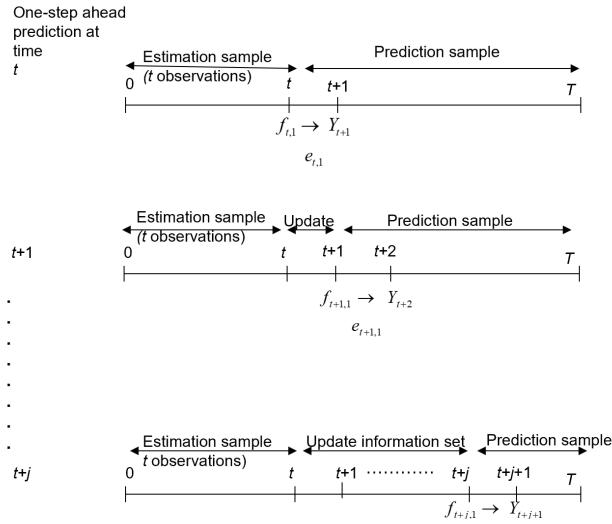


Figure 3.3: Fixed Forecasting Scheme

3.3 Loss Function

Definition 3.3. A loss function $L(e)$ is a real-valued function defined on the space of *forecast errors*, \mathcal{E} , and satisfies the following properties

- (i) $L(e) = 0 \iff \|e\| = 0$;

- (ii) $\forall e \in \mathcal{E}, L(e) \geq 0$ ¹;
- (iii) L is monotonically increasing in the norm of forecast error.

Example 3.1 (Symmetric Loss Functions with $\mathcal{E} = \mathbb{R}$).

$$L(e) = ae^2, \quad a > 0 \quad (3.2)$$

$$L(e) = a|e|, \quad a > 0 \quad (3.3)$$

Example 3.2 (Asymmetric Loss Functions with $\mathcal{E} = \mathbb{R}$).

$$L(e) = \exp(ae) - ae - 1, \quad a > 0 \quad \text{Linex Function} \quad (3.4)$$

$$L(e) = a|e| \mathbb{I}(e \geq 0) + b|e| \mathbb{I}(e < 0) \quad \text{Lin-lin Function} \quad (3.5)$$

3.4 Optimal Forecast

Definition 3.4. Based on information set I_t , the optimal forecast for future value y_{t+h} is the $f_{t,h}^*$ minimize the **expected loss function**

$$\mathbb{E}[L|I_t] = \int L(y_{t+h} - f_{t,h})f(y_{t+h}) dy_{t+h} \quad (3.6)$$

Assumption 3.1. Assuming the forecast $f(y_{t+h}|I_t)$ follows

$$f(y_{t+h}|I_t) \sim \mathcal{N}(\mathbb{E}[Y_{t+h}|I_t], \mathbb{V}[Y_{t+h}|I_t]) \quad (3.7)$$

Proposition 3.1. Given symmetric quadratic L , the optimal forecast $f_{t,h}^*$ is

$$\mu_{t+h|t} \equiv \mathbb{E}[Y_{t+h}|I_t] \quad (3.8)$$

Proof.

$$\min_{f_{t,h} \in \mathbb{R}} L \equiv \int (y_{t+h} - f_{t,h})^2 f(y_{t+h}|I_t) dy_{t+h} \quad (3.9)$$

$$\frac{\partial L}{\partial f_{t,h}} = -2 \int (y_{t+h} - f_{t,h}) f(y_{t+h}|I_t) dy_{t+h} = 0 \quad (3.10)$$

$$\implies \int (y_{t+h} - f_{t,h}) f(y_{t+h}|I_t) dy_{t+h} = 0 \quad (3.11)$$

$$\implies \int y_{t+h} f(y_{t+h}|I_t) dy_{t+h} = f_{t,h} \int f(y_{t+h}|I_t) dy_{t+h} \quad (3.12)$$

$$\implies f_{t,h} = \mathbb{E}[y_{t+h}|I_t] \equiv \mu_{t+h|t} \quad (3.13)$$

■

¹Since forecasting here can be considered as an optimization process, with L as the objective function. It's fine for L not satisfying the non-negativity condition. However, by convention, we assume L to be non-negative.

4 Moving Average Process

4.1 Wold Decomposition Theorem

Theorem 4.1 (Wold Decomposition Theorem). Every covariance stationary stochastic process $\{Y_t\}$ with mean zero and finite positive variance can be *uniquely* represented as

$$Y_t = V_t + \sum_{j=0}^{\infty} \psi_j L^j \varepsilon_t = V_t + \Psi(L) \varepsilon_t \quad (4.1)$$

where

- (i) $\{V_t\}$ is a *deterministic component* (e.g. trend or cycle);
- (ii) $\varepsilon_t \sim WN(0, \sigma^2)$ is the *stochastic component*;
- (iii) $\psi_0 = 1^2$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$;
- (iv) $\mathbb{E}[\varepsilon_t, V_s] = 0 \forall t, s \in \mathcal{T}$.

Definition 4.1. The stochastic component $\{\varepsilon_t\}$ in the decomposition is called **random shocks** or **innovations**.

Lemma 4.1. Given $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, then for all $\varepsilon > 0$ there exists a natural number J such that

$$\sum_{j=J}^{\infty} \psi_j^2 < \varepsilon \quad (4.2)$$

Corollary 4.1. By above lemma, assuming $V_t = 0$, we can approximate the decomposition by a linear combination of finite innovations.

$$Y_t \approx \hat{Y}_t = \sum_{j=0}^n \psi_j L^j \varepsilon_t \quad (4.3)$$

and the approximation is *accurate in Euclidean norm*, that's,

$$\mathbb{E}[Y_t - \sum_{j=0}^n \psi_j L^j \varepsilon_t]^2 \rightarrow 0 \text{ as } n \rightarrow 0 \quad (4.4)$$

4.2 Moving Average Process

Definition 4.2. The **Moving Average** of order q with deterministic trend, $MA(q)$, process is defined by the following stochastic difference equation

$$Y_t = \mu + \Theta(L) \varepsilon_t = \mu + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \text{ where } \theta_0 = 1, \theta_q \neq 0 \quad (4.5)$$

where $\{\varepsilon_t\}$ is innovations.

Lemma 4.2. The infinite lag polynomial in $\Psi(L)$ can be approximated by

$$\Psi(L) \approx \frac{\Theta_q(L)}{\Phi_p(L)} \quad (4.6)$$

Definition 4.3. $MA(1)$ process takes the form of

$$Y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1} \quad (4.7)$$

²We can always normalize ψ_0 to 1.

Unconditional Moments of MA(1)

$$\mathbb{E}[Y_t] = \mathbb{E}[\mu + \varepsilon_t + \theta\varepsilon_{t-1}] = \mu \quad (4.8)$$

$$\mathbb{V}[Y_t] = \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})^2] \quad (4.9)$$

$$= \mathbb{V}[\varepsilon_t] + \theta^2\mathbb{V}[\varepsilon_{t-1}] \quad (4.10)$$

$$= (1 + \theta^2)\sigma_\varepsilon^2 \quad (4.11)$$

Auto-covariance

$$\gamma_0 = \mathbb{V}[Y_t] = (1 + \theta^2)\sigma_\varepsilon^2 \quad (4.12)$$

$$\gamma_1 = \mathbb{E}[(Y_t - \mu)(Y_{t-1} - \mu)] \quad (4.13)$$

$$= \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2})] \quad (4.14)$$

$$= \theta\sigma_\varepsilon^2 \quad (4.15)$$

$$\gamma_k = 0 \quad \forall k > 1 \quad (4.16)$$

Auto-correlation

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1 + \theta^2} \quad (4.17)$$

$$\rho_k = 0 \quad \forall k > 1 \quad (4.18)$$

Definition 4.4. A MA(1) process is **invertible** if $|\theta| < 1$, so that it can be written as an AR(∞) process.

Inverting. Let

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t+1} \quad (4.19)$$

where $|\theta| < 1$. Then,

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t+1} \quad (4.20)$$

$$\implies Y_t - \mu = (1 + \theta L)\varepsilon_t \quad (4.21)$$

$$\implies \frac{Y_t - \mu}{1 - (-\theta L)} = \varepsilon_t \quad (4.22)$$

$$\implies \varepsilon_t = (Y_t - \mu) \sum_{j=0}^{\infty} (-\theta L)^j \quad (4.23)$$

■

Equivalence note that for MA(1) process,

$$r_1 = \rho_1 = \frac{\theta}{1 + \theta^2} \quad (4.24)$$

and for any θ , $\frac{1}{\theta}$ will generate the same auto-correlation. We always choose the invertible MA representation with $|\theta| < 1$.

4.3 Forecasting with MA(1)

4.3.1 Forecasting with Horizon $h = 1$

Point estimate

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] \quad (4.25)$$

$$= \mathbb{E}[\mu + \theta\varepsilon_t + \varepsilon_{t+1}|I_t] \quad (4.26)$$

$$= \mu + \theta\varepsilon_t \quad (4.27)$$

Forecasting error

$$e_{t,1} = Y_{t+1} - f_{t,1} = \varepsilon_{t+1} \quad (4.28)$$

Forecasting uncertainty

$$\sigma_{t+1|t}^2 = \mathbb{V}[Y_{t+1}|I_t] \quad (4.29)$$

$$= \mathbb{E}[\varepsilon_{t+1}^2|I_t] \quad (4.30)$$

$$= \sigma_\varepsilon^2 \quad (4.31)$$

Density forecast assuming *normality* of ε

$$\mathcal{F} \equiv \mathcal{N}(\mu_{t+1|t}, \sigma_{t+1|t}^2) \quad (4.32)$$

$$\mu_{t+1|t} = \mu + \theta\varepsilon_t \quad (4.33)$$

$$\sigma_{t+1|t}^2 = \sigma_\varepsilon^2 \quad (4.34)$$

4.3.2 Forecasting with Horizon $h = 2$

Point estimate

$$f_{t,2} = \mathbb{E}[Y_{t+2}|I_t] \quad (4.35)$$

$$= \mathbb{E}[\mu + \theta\varepsilon_{t+1} + \varepsilon_{t+2}|I_t] \quad (4.36)$$

$$= \mu \quad (4.37)$$

Forecasting Error

$$Y_{t+2} - f_{t,2} = \theta\varepsilon_{t+1} + \varepsilon_{t+2} \quad (4.38)$$

Forecasting Uncertainty

$$\sigma_{t+2|t}^2 \equiv \mathbb{V}[Y_{t+2}|I_t] \quad (4.39)$$

$$= (1 + \theta^2)\sigma_\varepsilon^2 \quad (4.40)$$

Density forecast

$$\mathcal{F} = \mathcal{N}(\mu, (1 + \theta^2)\sigma_\varepsilon^2) \quad (4.41)$$

Remark 4.1. Since for any $h > 1$, MA(1) only generates the unconditional mean μ as the point estimate, we say MA(1) process is **short memory**.

4.4 Properties of MA(2) Process

Model

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \quad (4.42)$$

Unconditional Moments

$$\mathbb{E}[Y_t] = \mu \quad (4.43)$$

$$\mathbb{V}[Y_t] = (1 + \theta_1^2 + \theta_2^2)\sigma_\varepsilon^2 \quad (4.44)$$

Auto-covariance

$$\gamma_0 = \mathbb{V}[Y_t] = (1 + \theta_1^2 + \theta_2^2)\sigma_\varepsilon^2 \quad (4.45)$$

$$\gamma_1 = (\theta_1 + \theta_1\theta_2)\sigma_\varepsilon^2 \quad (4.46)$$

$$\gamma_2 = \theta_2\sigma_\varepsilon^2 \quad (4.47)$$

Auto-correlation

$$\rho_1 \equiv \frac{\gamma_1}{\gamma_0} = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \quad (4.48)$$

$$\rho_2 \equiv \frac{\gamma_2}{\gamma_0} = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \quad (4.49)$$

Optimal Forecasting

$$f_{t,1} = \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} \quad (4.50)$$

$$f_{t,2} = \mu + \theta_2 \varepsilon_t \quad (4.51)$$

$$f_{t,h} = \mu \quad \forall h > 2 \quad (4.52)$$

5 Auto-Regression Process and Seasonality

5.1 AR Process

Definition 5.1. An **auto-regressive** model of order p is taken in the form of

$$Y_t = c + \sum_{j=1}^p \phi_j L^j Y_t + \varepsilon_t \quad (5.1)$$

or equivalently

$$\Phi_p(L)Y_t = c + \varepsilon_t \quad (5.2)$$

Definition 5.2. An **auto-regressive** process of order 1 takes the form of stochastic difference equation

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t \quad (5.3)$$

where ϕ is called the **persistence parameter**.

Proposition 5.1. AR(1) process is stationary if and only if $|\phi| < 1$.

ACF and PACF

$$\rho_1 = r_1 = \phi \quad (5.4)$$

$$r_k = 0 \quad \forall k > 1 \quad (5.5)$$

5.1.1 Forecasting with AR(1) and $h = 1$

³ Point Estimate

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] \quad (5.6)$$

$$= c + \phi Y_t \quad (5.7)$$

Forecast Variance(Uncertainty)

$$\mathbb{V}[Y_{t+1}|I_t] = \mathbb{V}[c + \phi Y_t + \varepsilon_{t+1}|I_t] = \sigma_\varepsilon^2 \quad (5.8)$$

Density Forecast

$$\mathcal{F} = \mathcal{N}(c + \phi Y_t, \sigma_\varepsilon^2) \quad (5.9)$$

5.1.2 Forecasting with AR(1) and $h = s > 1$

Point Estimate

$$f_{t,s} = \mathbb{E}[Y_{t+s}|I_t] \quad (5.10)$$

$$= c + \mathbb{E}[\phi Y_{t+s-1}|I_t] \quad (5.11)$$

$$= (1 + \phi + \dots + \phi^{s-1})c + \phi^s Y_t \quad (5.12)$$

Forecasting Uncertainty

$$\mathbb{V}[Y_{t+s}|I_t] = \mathbb{V}[\varepsilon_{t+s} + \phi \varepsilon_{t+s-1} + \dots + \phi^{s-1} \varepsilon_{t+1}|I_t] \quad (5.13)$$

$$= \sum_{j=0}^{s-1} \phi^{2j} \sigma_\varepsilon^2 \quad (5.14)$$

5.1.3 Forecasting with AR(1) and $h \rightarrow \infty$

Assumption 5.1. For this subsection, assuming the AR(1) process is **stationary**, that's, $|\phi| < 1$.

Point Estimate

$$\lim_{h \rightarrow \infty} f_{t,h} = \frac{c}{1 - \phi} \quad (5.15)$$

Forecasting Uncertainty

$$\lim_{h \rightarrow \infty} \mathbb{V}[Y_{t+h}|I_t] = \frac{\sigma_\varepsilon^2}{1 - \phi^2} \quad (5.16)$$

Remark 5.1. The convergences demonstrated above suggest auto-regressive process is still a **short memory** process.

³While examining the optimal forecast in this section, we are assuming the loss function is *symmetric*.

5.1.4 Forecasting with AR(2) process

Definition 5.3. AR(2) process

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \quad (5.17)$$

Unconditional Moments

$$\mathbb{E}[Y_t] = c + \phi_1 \mathbb{E}[Y_{t-1}] + \phi_2 \mathbb{E}[Y_{t-2}] \quad (5.18)$$

$$\implies \mu_Y = \frac{c}{1 - \phi_1 - \phi_2} \quad (5.19)$$

Auto-covariance and Auto-correlation

$$\rho_1 = r_1 \quad (5.20)$$

$$r_2 = \phi_2 + \text{sampling error} \quad (5.21)$$

Optimal Forecasts $h = 1$

$$f_{t,1} = \mathbb{E}[Y_{t+1}|I_t] \quad (5.22)$$

$$= \mathbb{E}[c + \phi_1 Y_t + \phi_2 Y_{t-1} + \varepsilon_{t+1}|I_t] \quad (5.23)$$

$$= c + \phi_1 Y_t + \phi_2 Y_{t-1} \quad (5.24)$$

$$e_{t,1} = \varepsilon_{t+1} \quad (5.25)$$

$$\sigma_{t+1|t}^2 = \mathbb{V}[Y_{t+1}|I_t] = \sigma_\varepsilon^2 \quad (5.26)$$

Optimal Forecasts $h = 2$

$$f_{t,2} = \mathbb{E}[Y_{t+2}|I_t] \quad (5.27)$$

$$= \mathbb{E}[c + \phi_1 Y_{t+1} + \phi_2 Y_t + \varepsilon_{t+2}|I_t] \quad (5.28)$$

$$= c + \phi_1 f_{t,1} + \phi_2 Y_t \quad (5.29)$$

$$e_{t,2} = Y_{t+2} - f_{t,2} \quad (5.30)$$

$$= \phi_1 (Y_{t+1} - f_{t,1}) + \varepsilon_{t+2} \quad (5.31)$$

$$= \phi_1 e_{t,1} + \varepsilon_{t+2} \quad (5.32)$$

$$\sigma_{t+2|t}^2 = \mathbb{V}[Y_{t+2}|I_t] \quad (5.33)$$

$$= \phi_1^2 \sigma_{t+1|t}^2 + \sigma_\varepsilon^2 \quad (5.34)$$

$$= (1 + \phi_1^2) \sigma_\varepsilon^2 \quad (5.35)$$

Optimal Forecasts $h = s > 2$

$$f_{t,s} = \mathbb{E}[Y_{t+s}|I_t] \quad (5.36)$$

$$= c + \phi_1 f_{t,s-1} + \phi_2 f_{t,s-2} \quad (5.37)$$

$$e_{t,s} = \phi_1 e_{t,s-1} + \phi_2 e_{t,s-2} + \varepsilon_{t+s} \quad (5.38)$$

$$\sigma_{t+s|t}^2 = \mathbb{V}[Y_{t+s}|I_t] \quad (5.39)$$

$$= \mathbb{V}[e_{t,s}|I_t] \quad (5.40)$$

$$= \phi_1 \sigma_{t+s-1|t}^2 + \phi_2 \sigma_{t+s-2|t}^2 + \sigma_\varepsilon^2 \quad (5.41)$$

Remark 5.2. AR(2) is still classified as *short memory processes* as

$$\lim_{s \rightarrow \infty} f_{t,s} = \mu \quad (5.42)$$

$$\lim_{s \rightarrow \infty} \sigma_{t+2|t}^2 = \sigma_Y^2 \quad (5.43)$$

5.1.5 AP(p) process

Definition 5.4. Let $\Phi(L)Y_t = c + \varepsilon_t$ be an AR(p) process with trend c , then $\Phi(\cdot)$ is called the **characteristic polynomial** of this stochastic process.

Theorem 5.1. An autoregressive process is stationary if and only if all roots of its characteristic polynomial are **outside** the unit circle on \mathbb{C} .

Forecasting with AR(p) Process we apply a **recursive** scheme, or **chain rule of forecasting**, in which the we use the *forecasted* values to make prediction on even further values.

Example 5.1 (AP(p) chain rule of forecasting).

$$\textcolor{red}{f}_{t,1} = c + \sum_{j=1}^p \phi_j L^j Y_{t+1} \quad (5.44)$$

$$\textcolor{brown}{f}_{t,2} = c + \phi_1 \textcolor{red}{f}_{t,1} + \sum_{j=2}^p \phi_j L^j Y_{t+2} \quad (5.45)$$

$$\textcolor{blue}{f}_{t,3} = c + \phi_1 \textcolor{brown}{f}_{t,2} + \phi_2 \textcolor{red}{f}_{t,1} + \sum_{j=3}^p \phi_j L^j Y_{t+3} \quad (5.46)$$

$$f_{t,s} = c + \sum_{j=1}^p \phi_j f_{t,s-j} \quad \forall s > p \quad (5.47)$$

5.2 Seasonality

5.2.1 Deterministic Seasonality

Definition 5.5. The seasonality is **deterministic** if the regressors indicating seasons are always exactly predictable.

5.2.2 Stochastic Seasonality

Definition 5.6. The seasonality is **stochastic** if the seasonal component is driven by random variables.

Definition 5.7. A seasonal AP(p) model, S-AR(p), is defined by

$$Y_t = c + \phi_s Y_{t-s} + \phi_{2s} Y_{t-2s} + \cdots + \phi_{ps} Y_{t-ps} + \varepsilon_t \quad (5.48)$$

$$\Phi_p(L^s)Y_t = c + \varepsilon_t \quad (5.49)$$

where s refers to the **data frequency**. Such model seeks to explain the **dynamics across seasons**.

Definition 5.8. Characteristics of realizations from S-AR(p) process:

- (i) ACF decays slowly with spikes at multiples of s .
- (ii) PACF **only** spikes at multiples of s .

Definition 5.9. A seasonal MA(q) model, S-MA(q), is given by

$$Y_t = \mu + \Theta_q(L^s)\varepsilon_t \quad (5.50)$$

Remark 5.3. Characteristics of realizations from S-MA(q) process:

- (i) ACF **only** spikes at multiples of s .
- (ii) PACF decays slowly with spikes at multiples of s .

Proposition 5.2 (Combining ARMA and S-ARMA). Given ARMA

$$\Phi_p(L)Y_t = c + \Theta_q(L)\varepsilon_t \quad (5.51)$$

and S-ARMA

$$\Phi'_p(L^{s_1})Y_t = c + \Theta'_q(L^{s_2})\varepsilon_t \quad (5.52)$$

The combined model is given by **multiplying the lag polynomials**

$$\Phi_p(L)\Phi'_p(L^{s_1})Y_t = c + \Theta_q(L)\Theta'_q(L^{s_2})\varepsilon_t \quad (5.53)$$

6 Model Assessment and Asymmetric Loss

6.1 Model Assessment

Definition 6.1. Akaike information criterion(AIC) of a model with k parameters is defined as

$$AIC \equiv -2 \ln(\mathcal{L}) + 2k \quad (6.1)$$

Definition 6.2. Bayes information criterion(BIC)/Schwarz information criterion(SIC) of a model with k parameters and fitted on the sample with size N is defined as

$$BIC \equiv -2 \ln(\mathcal{L}) + 2 \ln(N)k \quad (6.2)$$

Definition 6.3. Given time series data sample with size T , and use the *recursive scheme* starting from $t < T$, with forecasting horizon h , given sequence of *ground truth*

$$\mathcal{Y} = (y_j)_{j=t+h}^T \quad (6.3)$$

we can construct a sequence of forecast

$$\mathcal{F} = (f_{t,h}, f_{t+1,h}, f_{t+2,h}, \dots, f_{T-h,h}) \quad (6.4)$$

and a sequence of forecasting errors

$$\mathcal{E} = (e_{j,t})_{j=t+h}^T \quad (6.5)$$

then,

$$\text{MSE} \equiv \frac{1}{|\mathcal{F}|} \sum_{e \in \mathcal{E}} e^2 \quad (6.6)$$

$$\text{MAE} \equiv \frac{1}{|\mathcal{F}|} \sum_{e \in \mathcal{E}} ||e|| \quad (6.7)$$

$$\text{MAPE} \equiv \frac{1}{|\mathcal{F}|} \sum_{(y,e) \in (\mathcal{Y}, \mathcal{E})} \left| \frac{e}{y} \right| \quad (6.8)$$

6.2 Asymmetric Loss

Definition 6.4 (Log-Normal Distribution). Let X be a Gaussian random variable with mean μ and variance σ^2 . Define $Y \equiv \exp(X)$, then Y follows **log-normal distribution**, with

$$\mathbb{E}[Y] = \exp(\mu + \frac{\sigma^2}{2}) \quad (6.9)$$

Example 6.1. Consider the Lin-ex loss function

$$L(e) = \exp(ae) - ae - 1 \quad (6.10)$$

then the expected loss for h step forecasting made at t is

$$\mathbb{E}[L(e_{t,h})|I_t] \quad (6.11)$$

$$= \mathbb{E}[\exp(a(y_{t+h} - f_{t,h})) - a(y_{t+h} - f_{t,h}) - 1 | I_t] \quad (6.12)$$

$$= \mathbb{E}[\exp(ay_{t+h}) \exp(-af_{t,h}) | I_t] - \mathbb{E}[ay_{t+h} | I_t] + af_{t,h} - 1 \quad (6.13)$$

$$= \exp(-af_{t,h}) \mathbb{E}[\exp(ay_{t+h}) | I_t] - a\mathbb{E}[y_{t+h} | I_t] + af_{t,h} - 1 \quad (6.14)$$

$$(6.15)$$

To find the optimal forecasting, take the FOC

$$\frac{\partial \mathbb{E}[L(e_{t,h})|I_t]}{\partial f_{t,h}} = 0 \quad (6.16)$$

$$\implies -a \exp(-af_{t,h}) \mathbb{E}[\exp(ay_{t+h}) | I_t] + a = 0 \quad (6.17)$$

$$\implies \exp(-af_{t,h}) \mathbb{E}[\exp(ay_{t+h}) | I_t] = 1 \quad (6.18)$$

$$\implies -af_{t,h} + \log(\mathbb{E}[\exp(ay_{t+h}) | I_t]) = 0 \quad (6.19)$$

$$\implies f_{t,h} = \frac{1}{a} \log(\mathbb{E}[\exp(ay_{t+h}) | I_t]) \quad (6.20)$$

Assuming $y_{t+h} \sim \mathcal{N}(\mu_{t+h|t}, \sigma_{t+h|t}^2)$,

$$\implies f_{t,h} = \frac{1}{a} \log(\exp(a\mathbb{E}[y_{t+h}|I_t] + \frac{a^2\sigma_{t+h|t}^2}{2})) \quad (6.21)$$

$$\implies f_{t,h} = \mathbb{E}[y_{t+h}|I_t] + \frac{a\sigma_{t+h|t}^2}{2} \quad (6.22)$$

So if $a < 0$, the penalty on making negative error is higher than positive error, and the optimal forecast would be *pushed down* to less than the conditional mean.

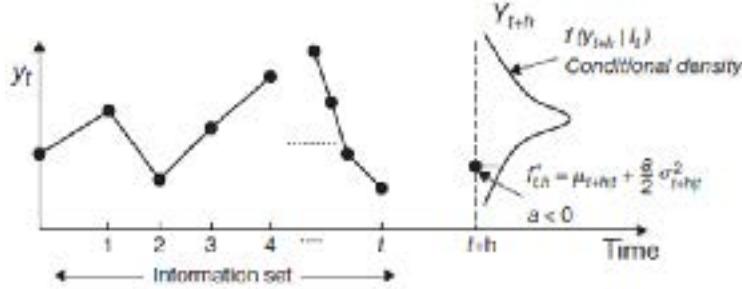


Figure 6.1: Illustration of the Optimal Forecast with Lin-Ex loss and $a < 0$

Forecasting Assuming AR(1) process.

Error with $h = 1$

$$e_{t+1,t} = c + \phi Y_t + \varepsilon_{t+1} - f_{t,1} \quad (6.23)$$

$$= \varepsilon_{t+1} - \frac{a\sigma_{t+1|t}^2}{2} \quad (6.24)$$

$$\mathbb{E}[e_{t+1,t}] = -\frac{a\sigma_{t+1|t}^2}{2} \quad (6.25)$$

$$\mathbb{V}[e_{t+1,t}|I_t] = \sigma_\varepsilon^2 \quad (6.26)$$

Error with $h = 2$

$$e_{t+2,t} = Y_{t+2} - f_{t,2} \quad (6.27)$$

$$= c + \phi(Y_{t+1}) + \varepsilon_{t+2} - f_{t,2} \quad (6.28)$$

$$= c + \phi c + \phi \varepsilon_{t+1} + \phi Y_t + \varepsilon_{t+2} - f_{t,2} \quad (6.29)$$

$$= c + \phi c + \phi \varepsilon_{t+1} + \phi Y_t + \varepsilon_{t+2} - \mathbb{E}[Y_{t+2}|I_t] - \frac{a\sigma_{t+2|t}^2}{2} \quad (6.30)$$

$$= \phi \varepsilon_{t+1} + \varepsilon_{t+2} - \frac{a\sigma_{t+2|t}^2}{2} \quad (6.31)$$

Note that for every $h > 0$, the term $\sigma_{t+h|t}$ is constant conditioned on I_t .

$$\mathbb{E}[e_{t+2,t}|I_t] = -\frac{a\sigma_{t+2|t}^2}{2} \quad (6.32)$$

$$\mathbb{V}[e_{t+2,t}|I_t] = (1 + \phi^2)\sigma_\varepsilon^2 \quad (6.33)$$