Forecasting and Time Series Econometrics
ECO374 Winter 2019

Tianyu Du

February 12, 2019

Contents

1 Introduction and Statistics Review

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

$$\mathbb{E}[X^k] \tag{1.1}$$

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

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

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

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

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

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

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

(i) **Linearity**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \tag{1.5}$$

(ii) **Zero Conditional Mean**

$$\mathbb{E}[u|X_1, X_2, \ldots, X_k] = 0 \tag{1.6}$$

(iii) **Homoscedasitcity**

$$\mathbb{V}[u|X_1, X_2, \ldots, X_k] = \sigma_u^2 \tag{1.7}$$

(iv) **No Serial Correlation**

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

(v) **No Perfect Collinearity**

(vi) **Sample Variation in Regressors**

$$\mathbb{V}[X_j] > 0 \ \forall j \tag{1.9}$$

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 $(\Omega, \mathcal{F}, P)$.

$$\{Y_t\} = Y_1, \ldots, Y_T \tag{2.1}$$

Definition 2.2 (1.2)**.** The function $t \to y_t$ 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

$$\{y_t\} = y_1, \ldots, y_T \tag{2.2}$$

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 $h, n \geq 1$, $(X_1, \ldots, X_n)$ and $(X_{1+h}, \ldots, X_{n+h}$ have the same distribution.

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

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

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

$$\gamma_Y(t, s) \equiv Cov(Y_t, Y_s) \tag{2.4}$$

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

$$= \mathbb{E}[Y_t Y_s] - \mathbb{E}[Y_t]\mathbb{E}[Y_s] \tag{2.6}$$

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

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

Proposition 2.1. By the symmetry of covariance,

$$\gamma_Y(h) = \gamma_Y(-h) \tag{2.8}$$

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

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

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 $\rho_Y : k \to \rho_{Y_t, Y_{t-k}}$, defined as

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

Proposition 2.2. Note that

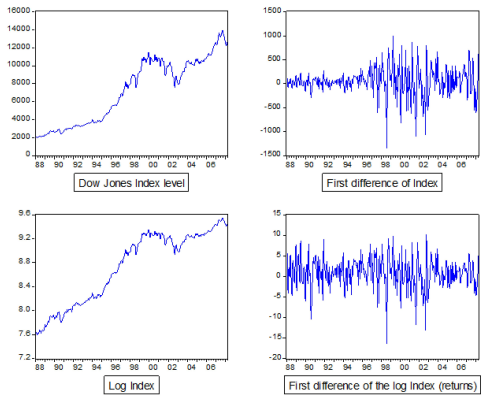
$$\rho_k = \rho_{-k} = \rho_{|k|} \tag{2.11}$$

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

$$\rho : k \to \rho_{|k|} \tag{2.12}$$

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

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

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

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

And Y_{T+1} can be expressed as

$$Y_{T+1} = \mathbb{P}_T Y_{T+1} + Z_{T+1} \tag{2.15}$$

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

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

$$\alpha(0) = 1 \tag{2.16}$$

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

$$\alpha(h) = corr\Big(Y_{h+1} - \mathbb{P}(Y_{h+1}|1, Y_2, \ldots, Y_h), X_1 - \mathbb{P}(Y_1|1, Y_2, \ldots, Y_h)\Big) \tag{2.18}$$

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 $(Y_{t+1}, \ldots, Y_{t+k-1})$.

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

$$H_0 : \rho_k = 0 \tag{2.19}$$

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

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0 \tag{2.20}$$

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

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

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,

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

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*.

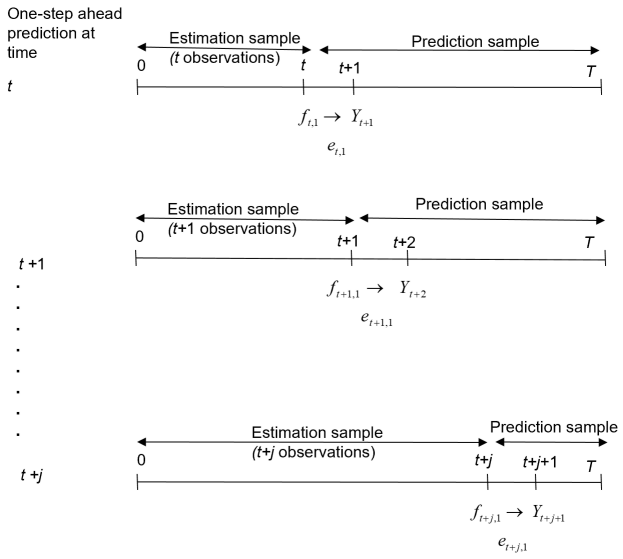


Figure 3.1: Recursive Forecasting Scheme

3.2.2 Forecasting Environments: *Rolling*

(i) *Updating* with *fixed-size* information set.

(ii) Robust against *structural breaks*.

(iii) Not fully exploit information available.

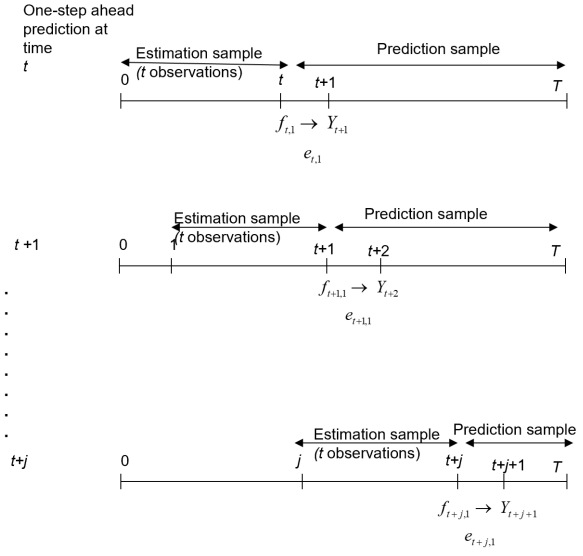


Figure 3.2: Rolling Forecasting Scheme

3.2.3 Forecasting Environments: *Fixed*

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

(ii) Computationally cheap.

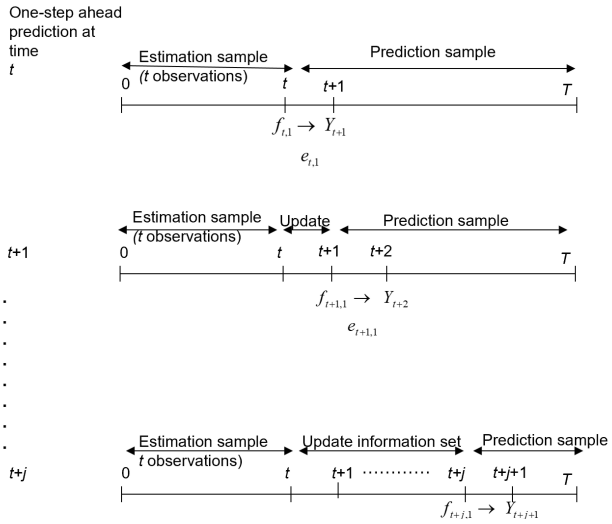


Figure 3.3: Fixed Forecasting Scheme

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 $\mathcal{E} = \mathbb{R}$)**.**

$$L(e) = ae^2, \ a > 0 \tag{3.2}$$

$$L(e) = a|e|, \ a > 0 \tag{3.3}$$

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

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

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

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

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

[1]Since 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.

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

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

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

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

Proof.

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

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

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

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

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

■

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

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

where

(i) $\{V_t\}$ is a *deterministic component* (e.g. trend or cycle);

(ii) $\varepsilon_t \sim \text{WN}(0, \sigma^2)$ 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

$$\sum_{j=J}^{\infty} \psi_j^2 < \varepsilon \tag{4.2}$$

[2]We can always normalize ψ_0 to 1.

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

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

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

$$\mathbb{E}[Y_t - \sum_{j=0}^{n} \psi_j L^j \varepsilon_t]^2 \to 0 \text{ as } n \to 0 \tag{4.4}$$

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

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

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

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

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

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

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \tag{4.7}$$

Unconditional Moments of MA(1)

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

$$\begin{aligned} \mathbb{V}[Y_t] &= \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})^2] && (4.9)\\ &= \mathbb{V}[\varepsilon_t] + \theta^2\mathbb{V}[\varepsilon_{t-1}] && (4.10)\\ &= (1+\theta^2)\sigma_\varepsilon^2 && (4.11) \end{aligned}$$

Auto-covariance

$$\begin{aligned} \gamma_0 &= \mathbb{V}[Y_t] = (1+\theta^2)\sigma_\varepsilon^2 && (4.12)\\ \gamma_1 &= \mathbb{E}[(Y_t - \mu)(Y_{t-1} - \mu)] && (4.13)\\ &= \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2})] && (4.14)\\ &= \theta\sigma_\varepsilon^2 && (4.15)\\ \gamma_k &= 0 \ \forall k > 1 && (4.16) \end{aligned}$$

Auto-correlation

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2} \tag{4.17}$$

$$\rho_k = 0 \ \forall k > 1 \tag{4.18}$$

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

Inverting. Let

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t+1} \tag{4.19}$$

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

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t+1} \tag{4.20}$$

$$\implies Y_t - \mu = (1+\theta L)\varepsilon_t \tag{4.21}$$

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

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

■

Equivalence note that for MA(1) process,

$$r_1 = \rho_1 = \frac{\theta}{1+\theta^2} \tag{4.24}$$

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

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

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

$$= \mu + \theta\varepsilon_t \tag{4.27}$$

Forecasting error

$$e_{t,1} = Y_{t+1} - f_{t,1} = \varepsilon_{t+1} \tag{4.28}$$

Forecasting uncertainty

$$\sigma^2_{t+1|t} = \mathbb{V}[Y_{t+1}|I_t] \tag{4.29}$$

$$= \mathbb{E}[\varepsilon^2_{t+1}|I_t] \tag{4.30}$$

$$= \sigma^2_\varepsilon \tag{4.31}$$

Density forecast assuming *normality* of ε

$$\mathcal{F} \equiv \mathcal{N}(\mu_{t+1|t}, \sigma^2_{t+1|t}) \tag{4.32}$$
$$\mu_{t+1|t} = \mu + \theta\varepsilon_t \tag{4.33}$$
$$\sigma^2_{t+1|t} = \sigma^2_\varepsilon \tag{4.34}$$

4.3.2 Forecasting with Horizon $h = 2$

Point estimate

$$f_{t,2} = \mathbb{E}[Y_{t+2}|I_t] \tag{4.35}$$
$$= \mathbb{E}[\mu + \theta\varepsilon_{t+1} + \varepsilon_{t+2}|I_t] \tag{4.36}$$
$$= \mu \tag{4.37}$$

Forecasting Error

$$Y_{t+2} - ft, 2 = \theta\varepsilon_{t+1} + \varepsilon_{t+2} \tag{4.38}$$

Forecasting Uncertainty

$$\sigma^2_{t+2|t} \equiv \mathbb{V}[Y_{t+2}|I_t] \tag{4.39}$$
$$= (1 + \theta^2)\sigma^2_\varepsilon \tag{4.40}$$

Density forecast

$$\mathcal{F} = \mathcal{N}(\mu, (1 + \theta^2)\sigma^2_\varepsilon) \tag{4.41}$$

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

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

Unconditional Moments

$$\mathbb{E}[Y_t] = \mu \tag{4.43}$$
$$\mathbb{V}[Y_t] = (1 + \theta_1^2 + \theta_2^2)\sigma^2_\varepsilon \tag{4.44}$$

Auto-covariance

$$\gamma_0 = \mathbb{V}[Y_t] = (1 + \theta_1^2 + \theta_2^2)\sigma^2_\varepsilon \tag{4.45}$$
$$\gamma_1 = (\theta_1 + \theta_1\theta_2)\sigma^2_\varepsilon \tag{4.46}$$
$$\gamma_2 = \theta_2\sigma^2_\varepsilon \tag{4.47}$$

Auto-correlation

$$\rho_1 \equiv \frac{\gamma_1}{\gamma_0} = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \tag{4.48}$$
$$\rho_2 \equiv \frac{\gamma_2}{\gamma_0} = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \tag{4.49}$$

Optimal Forecasting

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

$$f_{t,2} = \mu + \theta_2 \varepsilon_t \tag{4.51}$$

$$f_{t,h} = \mu \ \forall h > 2 \tag{4.52}$$