Statistical Inference and Methods

David A. Stephens

Department of Mathematics
Imperial College London

d.stephens@imperial.ac.uk
http://stats.ma.ic.ac.uk/~das01/

10th January 2006

Part III

Session 3: Time Series Analysis

- Exploratory Analysis
- Time Domain Models
- Frequency Domain modelling
- Inference and Estimation
- Non stationarity/Unit Roots

Time series analysis is a branch of **applied stochastic processes**. We start with an indexed family of **random variables**

$$\{X_t : \quad t \in T\}$$

where t is the index, here taken to be time (but it could be space). T is called the index set. We have a state space of values of X.

In addition X could be **univariate** or **multivariate**. We shall concentrate on discrete time. Samples are taken at equal intervals. We wish to use time series analysis to characterize time series and understand structure.

State (possible values of X)	Time	Notation
Continuous	Continuous	$X(t)$
Continuous	Discrete	X_t
Discrete	Continuous	
Discrete	Discrete	

Exploratory Analysis

We consider **lag** k scatter plots by plotting x_t versus x_{t+k}, but they are unwieldy. Suppose we make the assumption that a **linear** relationship holds approximately between x_{t+k} and x_t for all k, *i.e.*,

$$x_{t+k} = \alpha_k + \beta_k x_t + \varepsilon_{t+k}$$

where ε_{t+k} is an random error term.

The association between two variables $\{y_t\}$ and $\{z_t\}$ is the **Pearson product moment correlation coefficient**

$$\hat{\rho} = \frac{\sum(y_t - \bar{y})(z_t - \bar{z})}{\sqrt{\sum(y_t - \bar{y})^2 \sum(z_t - \bar{z})^2}}$$

where $\bar{y}$ and $\bar{z}$ are the sample means.

Hence if $y_t = x_{t+k}$ and $z_t = x_t$ we are led to the lag k sample autocorrelation for a time series:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

with $\hat{\rho}_0 = 1$.

The sequence $\{\hat{\rho}_k\}$ is called the **sample autocorrelation sequence** (sample acfs) for the time series.

The series $x_1, \ldots, x_N$ can be regarded as a realization of the corresponding random variables $X_1, \ldots, X_N$, $\hat{\rho}_k$ is an estimate of a corresponding population quantity called the lag k theoretical autocorrelation, defined as

$$\rho_k = \frac{E\left[(X_t - \mu)(X_{t+k} - \mu)\right]}{\sigma^2}$$

where

$$\mu = E\left[X_t\right] \qquad \sigma^2 = E\left[(X_t - \mu)^2\right]$$

are the process mean and process variance.

Note that ρ_k, μ and σ^2 do not depend on t

Denote the process by $\{X_t\}$. For fixed t, X_t is a random variable (r.v.), and hence there is an associated cumulative distribution function (cdf):

$$F_t(a) = P(X_t \le a).$$

But we are interested in the relationships between the various r.v.s that form the process. For example, for any t_1 and $t_2 \in T$,

$$F_{t_1,t_2}(a_1,a_2) = P(X_{t_1} \le a_1, X_{t_2} \le a_2)$$

gives the bivariate cdf. More generally for any $t_1, t_2, \ldots, t_n \in T$,

$$F_{t_1,t_2,\ldots,t_n}(a_1,a_2,\ldots,a_n) = P(X_{t_1} \le a_1, \ldots, X_{t_n} \le a_n)$$

We consider the subclass of **stationary processes**.

Stationarity

Strong stationarity $\{X_t\}$ is said to be strongly (strictly, completely) stationary if, for all $n \ge 1$, for any

$$t_1, t_2, \ldots, t_n \in T$$

and for any τ such that

$$t_1+\tau, t_2+\tau, \ldots, t_n+\tau \in T$$

are also contained in the index set, the joint cdf of $\{X_{t_1}, \ldots, X_{t_n}\}$ is the same as that of $\{X_{t_1+\tau}, \ldots, X_{t_n+\tau}\}$ i.e.,

$$F_{t_1,t_2,\ldots,t_n}(a_1,a_2,\ldots,a_n) = F_{t_1+\tau,t_2+\tau,\ldots,t_n+\tau}(a_1,a_2,\ldots,a_n).$$

Second-order stationarity $\{X_t\}$ is said to be second-order (weakly) stationary if, for all $n \ge 1$, for any

$$t_1, t_2, \ldots, t_n \in T$$

and for any τ such that $t_1+\tau, t_2+\tau, \ldots, t_n+\tau \in T$ are also contained in the index set, all the joint moments of orders 1 and 2 of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ exist and are finite.

Most importantly, these moments are identical to the corresponding joint moments of $\{X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_n+\tau}\}$.

Hence,

$$E[X_t] \equiv \mu \qquad Var[X_t] \equiv \sigma^2 \qquad (= E\left[X_t^2\right] - \mu^2),$$

are constants independent of t. If we let $\tau = -t_1$,

$$E[X_{t_1}X_{t_2}] = E[X_{t_1+\tau}X_{t_2+\tau}] = E[X_0 X_{t_2-t_1}],$$

and with $\tau = -t_2$,

$$E[X_{t_1}X_{t_2}] = E[X_{t_1+\tau}X_{t_2+\tau}] = E[X_{t_1-t_2}X_0].$$

Hence, $E\{X_{t_1}X_{t_2}\}$ is a function of the absolute difference $|t_2 - t_1|$ only, similarly, for the **covariance** between X_{t_1} & X_{t_2}:

$$Cov\left[X_{t_1}, X_{t_2}\right] = E\left[(X_{t_1} - \mu)(X_{t_2} - \mu)\right] = E\left[X_{t_1}X_{t_2}\right] - \mu^2.$$

The **autocovariance sequence** (**acvs**), s_τ, is defined by

$$s_\tau \equiv Cov\left[X_t, X_{t+\tau}\right] = Cov\left[X_0, X_\tau\right].$$

- τ is called the lag.
- $s_0 = \sigma^2$ and $s_{-\tau} = s_\tau$, with $|s_\tau| \leq s_0$ for $\tau > 0$.
- The autocorrelation sequence (acfs) is given by

$$\rho_\tau = \frac{s_\tau}{s_0} = \frac{Cov\left[X_t, X_{t+\tau}\right]}{\sigma^2}.$$

The sequence $\{s_\tau\}$ is positive semidefinite, i.e., for all $n \geq 1$, for any $t_1, t_2, \ldots, t_n$ contained in the index set, and for any set of nonzero real numbers $a_1, a_2, \ldots, a_n$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} s_{t_j - t_k} a_j a_k \geq 0.$$

- Let $\mathbf{a} = (a_1, a_2, \ldots, a_n)^\mathsf{T}$, $\mathbf{V} = (X_{t_1}, X_{t_2}, \ldots, X_{t_n})^\mathsf{T}$, and let Σ be the variance-covariance matrix of $\mathbf{V}$. Its j, k-th element is given by

$$s_{t_j - t_k} = E\left[(X_{t_j} - \mu)(X_{t_k} - \mu)\right]$$

- Define the r.v.

$$w = \sum_{j=1}^{n} a_j X_{t_j} = \mathbf{a}^\mathsf{T}\mathbf{V}.$$

Then

$$\begin{aligned} 0 \leq Var\left[w\right] &= Var\left[\mathbf{a}^\mathsf{T}\mathbf{V}\right] = \mathbf{a}^\mathsf{T} Var\left[\mathbf{V}\right]\mathbf{a} = \mathbf{a}^\mathsf{T}\Sigma\mathbf{a} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} s_{t_j - t_k} a_j a_k. \end{aligned}$$

- The variance-covariance matrix of equispaced X's, $(X_1, X_2, \ldots, X_N)^\mathsf{T}$ has the form

$$\begin{bmatrix} s_0 & s_1 & \cdots & s_{N-2} & s_{N-1} \\ s_1 & s_0 & \cdots & s_{N-3} & s_{N-2} \\ \vdots & & \ddots & & \\ s_{N-2} & s_{N-3} & \cdots & s_0 & s_1 \\ s_{N-1} & s_{N-2} & \cdots & s_1 & s_0 \end{bmatrix}$$

which is known as a symmetric Toeplitz matrix – all elements on a diagonal are the same.
- Note the above matrix has only N unique elements, $s_0, s_1, \ldots, s_{N-1}$.

- A stochastic process $\{X_t\}$ is called Gaussian if, for all $n \geq 1$ and for any $t_1, t_2, \ldots, t_n$ contained in the index set, the joint cdf of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ is multivariate Gaussian.
- 2nd-order stationary Gaussian $\Rightarrow$ complete stationarity
- follows as the multivariate Normal distribution is completely characterized by 1st and 2nd moments
- not true in general.
- Complete stationarity $\not\Rightarrow$ 2nd-order stationary in general.

White noise process
Also known as a purely random process. Let $\{X_t\}$ be a sequence of uncorrelated r.v.s such that

$$E[X_t] = \mu \qquad Var[X_t] = \sigma^2 \qquad \forall t$$

and

$$s_\tau = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \quad \text{or} \quad \rho_\tau = \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

forms a basic building block in time series analysis. Very different realizations of white noise can be obtained for different distributions of $\{X_t\}$.

q-th order moving average process MA(q)
X_t can be expressed in the form

$$X_t = \mu - \theta_{0,q}\epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q} = \mu - \sum_{j=0}^{q} \theta_{j,q}\epsilon_{t-j},$$

where μ and $\theta_{j,q}$'s are constants ($\theta_{0,q} \equiv -1, \theta_{q,q} \neq 0$), and $\{\epsilon_t\}$ is a zero-mean white noise process with variance σ_ϵ^2.
We assume $E[X_t] = \mu = 0$. Then

$$Cov[X_t, X_{t+\tau}] = E\{X_t X_{t+\tau}\}$$

Recall: $Cov(X, Y) = E\{(X - E\{X\})(Y - E\{Y\})\}$. Since $E\{\epsilon_t \epsilon_{t+\tau}\} = 0 \ \forall \tau \neq 0$ we have for $\tau \geq 0$.

$$\begin{aligned} Cov[X_t, X_{t+\tau}] &= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_{j,q}\theta_{k,q} E\{\epsilon_{t-j}\epsilon_{t+\tau-k}\} \\ &= \sigma_\epsilon^2 \sum_{j=0}^{q-\tau} \theta_{j,q}\theta_{j+\tau,q} \qquad (k = j + \tau) \\ &\equiv s_\tau, \end{aligned}$$

which does not depend on t.

Since $s_\tau = s_{-\tau}$, $\{X_t\}$ is a stationary process with acvs given by

$$s_\tau = \begin{cases} \sigma_\epsilon^2 \sum_{j=0}^{q-|\tau|} \theta_{j,q}\theta_{j+|\tau|,q} & |\tau| \le q \\ 0 & |\tau| > q \end{cases}$$

No restrictions were placed on the $\theta_{j,q}$'s to ensure stationarity.

Example: $X_t = \epsilon_t - \theta_{1,1}\epsilon_{t-1}$ MA(1)

$\boxed{\text{acvs:}}$

$$s_\tau = \sigma_\epsilon^2 \sum_{j=0}^{1-|\tau|} \theta_{j,1}\theta_{j+|\tau|,1} \quad |\tau| \le 1,$$

so,

$$s_0 = \sigma_\epsilon^2(\theta_{0,1}\theta_{0,1} + \theta_{1,1}\theta_{1,1}) = \sigma_\epsilon^2(1 + \theta_{1,1}^2);$$

and,

$$s_1 = \sigma_\epsilon^2\theta_{0,1}\theta_{1,1} = -\sigma_\epsilon^2\theta_{1,1}.$$

$\boxed{\text{acfs:}}$

$$\rho_\tau = \frac{s_\tau}{s_0} : \rho_0 = 1.0 \qquad \rho_1 = \frac{-\theta_{1,1}}{1+\theta_{1,1}^2}$$

For $\theta_{1,1} = 1.0, \sigma_\epsilon^2 = 1.0$, we have,

$$s_0 = 2.0, s_1 = -1.0, s_2, s_3, \ldots = 0.0,$$

giving,

$$\rho_0 = 1.0, \rho_1 = -0.5, \rho_2, \rho_3, \ldots = 0.0.$$

For $\theta_{1,1} = -1.0, \sigma_\epsilon^2 = 1.0$, we have,

$$s_0 = 2.0, s_1 = 1.0, s_2, s_3, \ldots = 0.0,$$

giving,

$$\rho_0 = 1.0, \rho_1 = 0.5, \rho_2, \rho_3, \ldots = 0.0.$$

Note: if we replace $\theta_{1,1}$ by $\theta_{1,1}^{-1}$ the model becomes

$$X_t = \epsilon_t - \frac{1}{\theta_{1,1}}\epsilon_{t-1}$$

and the autocorrelation becomes

$$\rho_1 = \frac{-\dfrac{1}{\theta_{1,1}}}{1 + \left(\dfrac{1}{\theta_{1,1}}\right)^2} = \frac{-\theta_{1,1}}{\theta_{1,1}^2 + 1},$$

i.e., is unchanged. Thus we cannot identify the MA(1) process uniquely from the autocorrelation.

p-th order autoregressive process AR(p)
$\{X_t\}$ is expressed in the form

$$X_t = \phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \ldots + \phi_{p,p}X_{t-p} + \epsilon_t,$$

where $\phi_{1,p}, \phi_{2,p}, \ldots, \phi_{p,p}$ are constants ($\phi_{p,p} \neq 0$) and $\{\epsilon_t\}$ is a zero mean white noise process with variance σ_ϵ^2.

In contrast to the parameters of an MA(q) process, the $\{\phi_{k,p}\}$ **must satisfy certain conditions** for $\{X_t\}$ to be a stationary process – not all AR(p) processes are stationary.

Example

$$\begin{aligned} X_t &= \phi_{1,1}X_{t-1} + \epsilon_t \\ &= \phi_{1,1}\{\phi_{1,1}X_{t-2} + \epsilon_{t-1}\} + \epsilon_t \\ &= \phi_{1,1}^2 X_{t-2} + \phi_{1,1}\epsilon_{t-1} + \epsilon_t \\ &\vdots \\ &= \sum_{k=0}^{\infty} \phi_{1,1}^k \epsilon_{t-k} \qquad \text{(initial condition } X_{-N} = 0; \text{let } N \to \infty \end{aligned}$$

$$Var\,[X_t] = Var\left[\sum_{k=0}^{\infty} \phi_{1,1}^k \epsilon_{t-k}\right] = \sum_{k=0}^{\infty} Var\{\phi_{1,1}^k \epsilon_{t-k}\} = \sigma_\epsilon^2 \sum_{k=0}^{\infty} \phi_{1,1}^{2k}.$$

For $Var\,[X_t] < \infty$ we must have $|\phi_{1,1}| < 1$, in which case

$$Var\,[X_t] = \frac{\sigma_\epsilon^2}{1 - \phi_{1,1}^2}.$$

To find the form of the acvs, we notice that for $\tau > 0$, $X_{t-\tau}$ is a linear function of $\epsilon_{t-\tau}, \epsilon_{t-\tau-1}, \ldots$ and is therefore uncorrelated with ϵ_t. Hence

$$E\,[\epsilon_t X_{t-\tau}] = 0.$$

Assuming stationarity and multiplying the defining equation (1) by $X_{t-\tau}$:

$$\begin{aligned} X_t X_{t-\tau} &= \phi_{1,1} X_t X_{t-\tau} + \epsilon_t X_{t-\tau} \\ \Longrightarrow E\,[X_t X_{t-\tau}] &= \phi_{1,1} E\,[X_{t-1} X_{t-\tau}] \end{aligned}$$

so that

$$s_\tau = \phi_{1,1} s_{\tau-1} = \phi_{1,1}^2 s_{\tau-2} = \ldots = \phi_{1,1}^\tau s_0 \qquad \Rightarrow \rho_\tau = \frac{s_\tau}{s_0} = \phi_{1,1}^\tau$$

However ρ_τ is an even function of τ, so

$$\rho_\tau = \phi_{1,1}^{|\tau|} \qquad \tau = 0, \pm 1, \pm 2, \ldots.$$

giving exponential decay

(p, q)'th order autoregressive-moving average process ARMA(p, q)

Here $\{X_t\}$ is expressed as

$$X_t = \phi_{1,p}X_{t-1} + \ldots + \phi_{p,p}X_{t-p} + \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

where the $\phi_{j,p}$'s and the $\theta_{j,q}$'s are all constants ($\phi_{p,p} \neq 0; \theta_{q,q} \neq 0$) and again $\{\epsilon_t\}$ is a zero mean white noise process with variance σ_ϵ^2.

The ARMA class is important as many data sets may be approximated in a more parsimonious way (meaning fewer parameters are needed) by a mixed ARMA model than by a pure AR or MA process.

The General Linear Process

Consider a process of the form

$$X_t = \sum_{k=-\infty}^{\infty} g_k \epsilon_{t-k},$$

where $\{\epsilon_t\}$ is a purely random process, with

$$\sum_{k=-\infty}^{\infty} g_k^2 < \infty.$$

This condition ensures that $\{X_t\}$ has finite variance. Now $|\rho_t| \leq 1$, so, also,

$$|s_\tau| = |Cov\,[X_t, X_{t-\tau}]| \leq \sigma_X^2 = \sigma_\epsilon^2 \sum_k g_k^2 < \infty.$$

If $g_{-1}, g_{-2}, \ldots = 0$, then we obtain what is called the *General Linear Process*

$$X_t = \sum_{k=0}^{\infty} g_k \epsilon_{t-k},$$

where X_t depends only on past and present values $\epsilon_t, \epsilon_{t-2}, \epsilon_{t-2}, \ldots$ of the purely random process. Consider the function

$$G(z) = \sum_{k=0}^{\infty} g_k z^k,$$

"z-polynomial" where $z = e^{-i\omega}$. Note $X_t = G(B)\epsilon_t$.

Then write

$$G(z) = \frac{G_1(z)}{G_2(z)}$$

Call the zeros of $G_2(z)$ (the "poles" of $G(z)$) in the **complex plane** $z_1, z_2, \ldots, z_p$, where the zeros are ordered so that $z_1, \ldots, z_k$ are inside and $z_{k+1}, \ldots, z_p$ are outside the unit circle $|z| = 1$.

Then, if all the roots of $G_2(z)$ are outside the unit circle (i.e. all the poles of $G(z)$ are outside the unit circle) only past and present values of $\{\epsilon_t\}$ are involved and the General Linear Process exists.

Another way of stating this is that

$$G(z) < \infty \qquad |z| \leq 1$$

i.e., $G(z)$ is analytic inside and on the unit circle. Thus

- all the poles of $G(z)$ lie outside the unit circle
- all the roots of $G^{-1}(z) = 0$ lie outside the unit circle

Consider the MA(q) model

$$X_t = \Theta(B)\epsilon_t \qquad \Longrightarrow \qquad \Theta^{-1}(B)X_t = \epsilon_t$$

and in general, the expansion of $\Theta^{-1}(B)$ is a polynomial of infinite order. Similarly, consider the AR(p) model

$$\Phi(B)X_t = \epsilon_t \qquad \Longrightarrow \qquad X_t = \Phi^{-1}(B)\epsilon_t.$$

Hence

$$\begin{aligned} \text{MA (finite order)} &\equiv \text{AR (infinite order)} \\ \text{AR (finite order)} &\equiv \text{MA (infinite order)} \end{aligned}$$

provided the infinite order expansions exist

Invertibility
Consider inverting the general linear process into autoregressive form

$$\begin{aligned} X_t &= \sum_{k=0}^{\infty} g_k \epsilon_{t-k} = \sum_{k=0}^{\infty} g_k B^k \epsilon_t \\ &= G(B)\epsilon_t \end{aligned}$$

so that

$$G^{-1}(B)X_t = \epsilon_t$$

The expansion of $G^{-1}(B)$ in powers of B gives the required autoregressive form **provided $G^{-1}(B)$ admits a power series expansion**

$$G^{-1}(z) = \sum_{k=0}^{\infty} h_k z^k$$

i.e. if $G^{-1}(z)$ is analytic, $|z| \leq 1$. Thus the model is invertible if all the poles of $G^{-1}(z)$ are outside the unit circle.

$$G^{-1}(z) < \infty, \quad |z| \leq 1.$$

For the MA(q) process, $G(z) = \Theta(z)$, and so the invertibility condition is that $\Theta(z)$ has no roots inside or on the unit circle; i.e. all the roots of $\Theta(z)$ lie outside the unit circle.

Stationarity of ARMA processes
For the AR(p) process

$$\Phi(B)X_t = \epsilon_t$$

so that

$$X_t = \Phi^{-1}(B)\epsilon_t = G(B)\epsilon_t,$$

so that $G(z) = \Phi^{-1}(z)$. Hence the requirement for stationarity is that all the roots of $G^{-1}(z) = \Phi(z)$ must lie outside the unit circle.

For the MA(q) process

$$X_t = \Theta(B)\epsilon_t = G(B)\epsilon_t$$

and since $G(B) = \Theta(B)$ is a polynomial of finite order $G(z) < \infty$, $|z| \leq 1$, automatically.

Example:

$$X_t = 1.3X_{t-1} - 0.4X_{t-2} + \epsilon_t - 1.5\epsilon_{t-1}.$$

Writing in B notation:

$$(1 - 1.3B + 0.4B^2)X_t = (1 - 1.5B)\epsilon_t$$

we have

$$\Phi(z) = 1 - 1.3z + 0.4z^2$$

with roots $z = 2$ and $5/4$, so the roots of $\Phi(z) = 0$ both lie outside the unit circle, and the model is stationary, and

$$\Theta(z) = 1 - 1.5z,$$

so the root of $\Theta(z) = 0$ is given by $z = 2/3$ which lies inside the unit circle and the model is not invertible.

Directionality and Reversibility
Consider again the general linear model

$$X_t = \sum_{k=0}^{\infty} g_k \epsilon_{t-k} = \sum_{k=0}^{\infty} g_k B^k \epsilon_t = G(B)\epsilon_t$$

The reversed form is clearly,

$$X_t = \sum_{k=0}^{\infty} g_k \epsilon_{t+k} = \sum_{k=0}^{\infty} g_k B^{-k} \epsilon_t = G\left(\frac{1}{B}\right)\epsilon_t,$$

with some stationarity condition.

Now consider the ARMA(p, q) model given by

$$\Phi(B)X_t = \Theta(B)\epsilon_t,$$

where,

$$\begin{aligned}
\Phi(B) &= 1 - \phi_{1,p}B - \phi_{2,p}B^2 - \ldots - \phi_{p,p}B^p \\
\Theta(B) &= 1 - \theta_{1,q}B - \theta_{2,q}B^2 - \ldots - \theta_{q,q}B^q
\end{aligned}$$

The reversed form of the ARMA(p, q) model is,

$$\Phi\left(\frac{1}{B}\right) X_t = \Theta\left(\frac{1}{B}\right)\epsilon_t \Longrightarrow \Phi^R(B)X_t = B^{p-q}\Theta^R\epsilon_t$$

where,

$$\begin{aligned}\Phi^R(B) &= B^p - \phi_{1,p}B^{p-1} - \phi_{2,p}B^{p-2} - \ldots - \phi_{p,p}\\ \Theta^R(B) &= B^q - \theta_{1,q}B^{q-1} - \theta_{2,q}B^{q-2} - \ldots - \theta_{q,q}\end{aligned}$$

For example, for the ARMA(1,1) model,

$$(1 - \phi_{1,1})X_t = (1 - \theta_{1,1})\epsilon_t,$$

reversed form is

$$(B - \phi_{1,1})X_t = (B - \theta_{1,1})\epsilon_t$$

Now $\Phi(z) = 1 - \phi_{1,1}z$, and a root is the solution of $1 - \phi_{1,1}z = 0$, i.e.,

$$|z| = \left|\frac{1}{\phi_{1,1}}\right| > 1 \Rightarrow |\phi_{1,1}| < 1.$$

But, $\Phi^R(z) = z - \phi_{1,1}$, and so a root is the solution of $z - \phi_{1,1} = 0$, i.e., $z = \phi_{1,1}$. But, since for stationarity $|\phi_{1,1}| < 1$ we have

$$|z| = |\phi_{1,1}| < 1,$$

so the root of $\Phi^R(z)$ is inside the unit circle.

Hence the standard assumption for stationarity (roots outside the unit circle) has within it an assumption of directionality. [N.B. only if the roots of $\Phi(z)$ are on the unit circle is model ALWAYS non-stationary].

Spectral Representations

Spectral analysis is a study of the frequency domain characteristics of a process, and describes the contribution of each frequency to the variance of the process. Let us define a complex "jump" process $\{Z(f)\}$ on the interval $[0, 1/2]$, such that

$$dZ(f) \equiv \begin{cases} Z(f + df) - Z(f), & 0 \le f < 1/2;\\ 0, & f = 1/2;\\ dZ^*(-f), & -1/2 \le f < 0,\end{cases}$$

where df is a small positive increment. If the intervals $[f, f + df]$ and $[f', f' + df']$ are non-intersecting subintervals of $[-1/2, 1/2]$, then the r.v.'s $dZ(f)$ and $dZ(f')$ are uncorrelated.

We say that the process has **orthogonal increments**, and the process itself is called an **orthogonal process** – this orthogonality results is very important.

Let $\{X_t\}$ be a real-valued discrete time stationary process, with zero mean, the **spectral representation theorem** states that there exists such an orthogonal process $\{Z(f)\}$, defined on $(-1/2, 1/2]$, such that

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f)$$

for all integers t.

The process $\{Z(f)\}$ has the following properties:

- $E\{dZ(f)\} = 0 \ \forall \ |f| \leq 1/2$.
- $E\{|dZ(f)|^2\} \equiv dS^{(I)}(f)$ say $\forall \ |f| \leq 1/2$, where $dS^{(I)}(f)$ is called the integrated spectrum of $\{X_t\}$, and
- for any two distinct frequencies f and $f' \in (-1/2, 1/2]$

$$Cov\{dZ(f'), dZ(f)\} = E\{dZ^*(f')dZ(f)\} = 0.$$

The spectral representation

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f) = \int_{-1/2}^{1/2} e^{i2\pi ft}\, |dZ(f)| e^{i \arg\{dZ(f)\}},$$

means that we can represent any discrete stationary process as an "infinite" sum of complex exponentials at frequencies f with associated random amplitudes $|dZ(f)|$ and random phases $\arg\{dZ(f)\}$.

The orthogonal increments property can be used to define the relationship between the autocovariance sequence $\{s_\tau\}$ and the integrated spectrum $S^I(f)$:

$$\begin{aligned} s_\tau = E\left[X_t X_{t+\tau}\right] &= E\left[X_t^* X_{t+\tau}\right] \\ &= E\left[\int_{-1/2}^{1/2} e^{-i2\pi f' t}\, dZ^*(f') \int_{-1/2}^{1/2} e^{i2\pi f(t+\tau)}\, dZ(f)\right] \\ &= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} e^{i2\pi(f-f')t} e^{i2\pi f\tau} E\{dZ^*(f')dZ(f)\}. \end{aligned}$$

Because of the orthogonal increments property,

$$E\{dZ^*(f')dZ(f)\} = dS^{(I)}(f) \qquad f = f'$$

and zero otherwise, so

$$s_\tau = \int_{-1/2}^{1/2} e^{i2\pi f\tau}\, dS^{(I)}(f),$$

which shows that the integrated spectrum determines the acvs for a stationary process. If $S^{(I)}(f)$ is differentiable with derivative $S(f)$ (the *spectral density function* (sdf)), we have

$$E\{|dZ(f)|^2\} = dS^{(I)}(f) = S(f)\, df.$$

Hence

$$s_\tau = \int_{-1/2}^{1/2} e^{i2\pi ft} S(f)\, df.$$

But a square summable deterministic sequence $\{g_t\}$ say has the **Fourier representation**

$$g_t = \int_{-1/2}^{1/2} G(f)e^{i2\pi ft}\, df \qquad \text{where} \qquad G(f) = \sum_{t=-\infty}^{\infty} g_t e^{-i2\pi ft},$$

If we assume that $S(f)$ is square integrable, then $S(f)$ is the **Fourier transform** of $\{s_\tau\}$,

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau}.$$

Hence,

$$\{s_\tau\} \longleftrightarrow S(f),$$

i.e., $\{s_\tau\}$ and $S(f)$ are a FT. pair.

$S(\cdot)$ has the following interpretation: $S(f)\, df$ is the average contribution (over all realizations) to the power from components with frequencies in a small interval about f. The power – or variance – is

$$\int_{-1/2}^{1/2} S(f)\, df.$$

Hence, $S(f)$ is often called the *power spectral density function* or just *power spectrum*.

Properties :

- $S^{(I)}(f) = \int_{-1/2}^{f} S(f')\, df'$.
- $0 \leq S^{(I)}(f) \leq \sigma^2$ where $\sigma^2 = Var\,[X_t]$; $\quad S(f) \geq 0$.
- $S^{(I)}(-1/2) = 0$; $\quad S^{(I)}(1/2) = \sigma^2$; $\quad \int_{-1/2}^{1/2} S(f)\, df = \sigma^2$.
- $f < f' \Rightarrow S^{(I)}(f) \leq S^{(I)}(f')$; $\quad S(-f) = S(f)$.

Except, basically, for the scaling factor σ^2, $S^{(I)}(f)$ has all the properties of a probability distribution function, and hence is sometimes called a **spectral distribution function**.

The integrated spectrum, $S^{(I)}(f)$ can be decomposed as

$$S^{(I)}(f) = S_1^{(I)}(f) + S_2^{(I)}(f)$$

where the $S_j^{(I)}(f)$'s are nonnegative, nondecreasing functions with $S_j^{(I)}(-1/2) = 0$ and are of the following types:

- $S_1^{(I)}(\cdot)$ has its derivative $S(\cdot)$ for all f, and

$$S^{(I)}(f) = \int_{-1/2}^{f} S(f')df'.$$

- $S_2^{(I)}(\cdot)$ is a step function with jumps of size $\{p_l\} : \ l = 1, 2, \ldots\}$ at the points $\{f_l : \ l = 1, 2, \ldots\}$.

(a) If $S_1^{(I)}(f) \geq 0; S_2^{(I)}(f) = 0$, $\{X_t\}$ has a *purely continuous* spectrum and $S(f)$ is absolutely integrable, with

$$\int_{-1/2}^{1/2} S(f)\cos(2\pi f\tau)\,df \quad \text{and} \quad \int_{-1/2}^{1/2} S(f)\sin(2\pi f\tau) \to 0,$$

as $\tau \to \infty$. But,

$$\begin{aligned} s_\tau &= \int_{-1/2}^{1/2} e^{i2\pi f\tau} S(f)\,df \\ &= \int_{-1/2}^{1/2} S(f)\cos(2\pi f\tau)\,df + i\int_{-1/2}^{1/2} S(f)\sin(2\pi f\tau)\,df \end{aligned}$$

so that $s_\tau \to 0$ as $|\tau| \to \infty$. In other words, the acvs diminishes to zero (called "mixing condition").

(b) If $S_1^{(I)}(f) = 0; S_2^{(I)}(f) \geq 0$, the integrated spectrum consists entirely of a step function, and the $\{X_t\}$ is said to have a *purely discrete spectrum* or a *line spectrum* .

The acvs for a process with a line spectrum never damps down to 0.

White noise spectrum
Recall that a white noise process $\{\epsilon_t\}$ has acvs:

$$s_\tau = \begin{cases} \sigma_\epsilon^2 & \tau = 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the spectrum of a white noise process is given by:

$$S_\epsilon(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau} = s_0 = \sigma_\epsilon^2.$$

i.e., white noise has a constant spectrum.

The sdf and acvs contain the same amount of information in that if we know one of them, we can calculate the other. However, they are often not equally informative.

- The sdf usually proves to be the more sensitive and interpretable diagnostic or exploratory tool.
- The sdf is able to distinguish between the processes while the acvs's are not noticeably different.
- dB $= 10\log_{10}(\text{power})$ scale often used.

Sampling and Aliasing

So far we have only looked at discrete time series $\{X_t\}$. However, such a process is usually obtained by sampling a continuous time process at equal intervals Δt, i.e., for a sampling interval $\Delta t > 0$ and an arbitrary time offset t_0, we can define a discrete time process through

$$X_t \equiv X(t_0 + t\Delta t), \qquad t = 0, \pm 1, \pm 2, \ldots.$$

If $\{X(t)\}$ is a stationary process with, say, sdf $S_{X(t)}(\cdot)$ and acvf $s(\tau)$, then $\{X_t\}$ is also a stationary process with, say, sdf $S_{X_t}(\cdot)$ and acvs $\{s_\tau\}$.

It can be shown that when $S^{(I)}_{X(t)}$ is differentiable:

$$S_{X_t}(f) = \sum_{k=-\infty}^{\infty} S_{X(t)}\left(f + \frac{k}{\Delta t}\right) \quad \text{for} \quad |f| \le \frac{1}{2\Delta t}.$$

Thus, the discrete time sdf at f is the sum of the continuous time sdf at frequencies $f \pm \frac{k}{\Delta t}$, $k = 0, 1, 2, \ldots$.

The frequency $1/(2\Delta t)$ is called the *Nyquist frequency*; previously we have taken $\Delta t = 1$, so that the frequency range was $|f| \le \frac{1}{2}$.

If $S_{X(t)}$ is essentially zero for $|f| > 1/(2\Delta t)$ we can expect good correspondence between $S_{X_t}(f)$ and $S_{X(t)}(f)$ for $|f| \le 1/(2\Delta t)$ (since

$$S_{X(t)}(f \pm k/(2\Delta t)) \approx 0$$

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

If $S_{X(t)}$ is large for some $|f| > 1/(2\Delta t)$, the correspondence can be quite poor, and an estimate of S_{X_t} will not tell us much about $S_{X(t)}$.

Estimation and Forecasting

Ergodic Property Methods we shall look at for estimating quantities such as the autocovariance function will use observations from a single realization.

Such methods are based on the strategy of replacing ensemble averages by their corresponding time averages.

Sample mean:
Given a time series $X_1, X_2, \ldots, X_N$, let

$$\bar{X} = \frac{1}{N}\sum X_t. \quad \left(\text{assume} \sum_{\tau=-\infty}^{\infty} |s_\tau| < \infty\right).$$

Then,

$$E\{\bar{X}\} = \frac{1}{N}\sum_{t=1}^{n} E\left[X_t\right] = \frac{1}{N}.N\mu = \mu$$

so $\bar{X}$ is an unbiased estimator of μ. Hence, $\bar{X}$ converges to μ in mean square if

$$\lim_{N\to\infty} Var\{\bar{X}\} = 0.$$

$$\begin{aligned} Var\{\bar{X}\} &= E\{(\bar{X}-\mu)^2\} = E\left\{\left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\right)^2\right\} \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{u=1}^{N} E\{(X_t-\mu)(X_u-\mu)\} = \frac{1}{N^2}\sum_{t=1}^{N}\sum_{u=1}^{N} s_{u-t} \\ &= \frac{1}{N^2}\sum_{\tau=-(N-1)}^{N-1}\sum_{k=1}^{N-|\tau|} s_\tau \\ &= \frac{1}{N^2}\sum_{\tau=-(N-1)}^{N-1}(N-|\tau|)s_\tau = \frac{1}{N}\sum_{\tau=-(N-1)}^{N-1}\left(1-\frac{|\tau|}{N}\right)s_\tau \end{aligned}$$

If

$$\sum_{\tau=-(N-1)}^{N-1} s_\tau$$

converges to a limit as $N \to \infty$, then

$$\text{it must since} \quad \left|\sum_{\tau=-(N-1)}^{N-1} s_\tau\right| \le \sum_{\tau=-(N-1)}^{N-1} |s_\tau| < \infty \ \forall N,$$

then $\sum_{\tau=-(N-1)}^{N-1}\left(1-\frac{|\tau|}{N}\right)s_\tau$ converges to the same limit.

We can thus conclude that,

$$\lim_{N\to\infty} N Var\{\bar{X}\} = \lim_{N\to\infty} \sum_{\tau=-(N-1)}^{N-1} \left(1 - \frac{|\tau|}{N}\right) s_\tau$$
$$= \lim_{N\to\infty} \sum_{\tau=-(N-1)}^{N-1} s_\tau = \sum_{\tau=-\infty}^{\infty} s_\tau.$$

The assumption of absolute summability of $\{s_\tau\}$ implies that $\{X_t\}$ has a purely continuous spectrum with sdf

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau}, \qquad \text{so that } S(0) = \sum_{\tau=-\infty}^{\infty} s_\tau.$$

Thus

$$\lim_{N\to\infty} N Var\{\bar{X}\} = S(0) \qquad \therefore \qquad Var\{\bar{X}\} \approx \frac{S(0)}{N} \text{ for large } N.$$

and therefore, $Var\{\bar{X}\} \to 0$. Note that the convergence of $\bar{X}$ depends only on the spectrum at $S(0)$, i.e. at $f = 0$.

Autocovariance Sequence: Now,

$$s_\tau = E\{(X_t - \mu)(X_{t+\tau} - \mu)\}$$

so that a natural estimator for the acvs is

$$\hat{s}_\tau^{(u)} = \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} (X_t - \bar{X})(X_{t+|\tau|} - \bar{X}) \quad \tau = 0, \pm 1, \ldots, \pm(N-1).$$

Note $\hat{s}_{-\tau}^{(u)} = \hat{s}_\tau^{(u)}$ as it should.

If we replace $\bar{X}$ by μ:

$$E\{\hat{s}_\tau^{(u)}\} = \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} E\{(X_t - \mu)(X_{t+|\tau|} - \mu)\}$$
$$= \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} s_\tau = s_\tau, \quad \tau = 0, \pm 1, \ldots, \pm(N-1).$$

Thus, $\hat{s}_\tau^{(u)}$ is an unbiased estimator of s_τ when μ is known. (Hence the (u) – for unbiased). Most texts refer to $\hat{s}_\tau^{(u)}$ as unbiased – however, if μ is estimated by $\bar{X}$, $\hat{s}_\tau^{(u)}$ is typically a biased estimator of s_τ.

A second estimator of s_τ is typically preferred:

$$\hat{s}_\tau^{(p)} = \frac{1}{N}\sum_{t=1}^{N-|\tau|}(X_t - \bar{X})(X_{t+|\tau|} - \bar{X}) \quad \tau = 0, \pm 1, \ldots, \pm(N-1).$$

With $\bar{X}$ replaced by μ:

$$E\{\hat{s}_\tau^{(p)}\} = \frac{1}{N}\sum_{t=1}^{N-|\tau|} s_\tau = \left(1 - \frac{|\tau|}{N}\right) s_\tau,$$

so that $\hat{s}_\tau^{(p)}$ is a biased estimator, and the magnitude of its bias increases as $|\tau|$ increases. Most texts refer to $\hat{s}_\tau^{(p)}$ as biased.

Why should we prefer the "biased" estimator $\hat{s}_\tau^{(p)}$ to the "unbiased" estimator $\hat{s}_\tau^{(u)}$?

1 For many stationary processes of practical interest

$$\text{mse}\{\hat{s}_\tau^{(p)}\} < \text{mse}\{\hat{s}_\tau^{(u)}\},$$

where

$$\begin{aligned}
\text{mse}\{\hat{s}_\tau\} &= E\{(\hat{s}_\tau - s_\tau)^2\} \\
&= E\{\hat{s}_\tau^2\} - 2s_\tau E\{\hat{s}_\tau\} + s_\tau^2 \\
&= (E\{\hat{s}_\tau^2\} - E^2\{\hat{s}_\tau\}) + E^2\{\hat{s}_\tau\} - 2s_\tau E\{\hat{s}_\tau\} + s_\tau^2 \\
&= Var\{\hat{s}_\tau\} + (s_\tau - E\{\hat{s}_\tau\})^2 \\
&= \text{variance} + (\text{bias})^2
\end{aligned}$$

2 If $\{X_t\}$ has a purely continuous spectrum we know that $s_\tau \to 0$ as $|\tau| \to \infty$. It therefore makes sense to choose an estimator that decreases nicely as $|\tau| \to N-1$ (i.e. choose $\hat{s}_\tau^{(p)}$).

3 We know that the acvs must be positive semidefinite, the sequence $\{\hat{s}_\tau^{(p)}\}$ has this property, whereas the sequence $\{\hat{s}_\tau^{(u)}\}$ may not.

The Periodogram

Suppose

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau} \quad |f| \le \frac{1}{2},$$

is purely continuous. We can use the (biased) estimator of s_τ:

$$\hat{s}_\tau^{(p)} = \frac{1}{N}\sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|}$$

for $|\tau| \le N-1$, but not for $|\tau| \ge N$. Hence we could replace s_τ by $\hat{s}_\tau^{(p)}$ for $|\tau| \le N-1$ and assume $s_\tau = 0$ for $|\tau| \ge N$.

Hence,

$$\begin{aligned}
\hat{S}^{(p)}(f) &= \sum_{\tau=-(N-1)}^{(N-1)} \hat{s}_\tau^{(p)} e^{-i2\pi f\tau} \\
&= \frac{1}{N} \sum_{\tau=-(N-1)}^{(N-1)} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|} e^{-i2\pi f\tau} \\
&= \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} X_j X_k e^{-i2\pi f(k-j)} = \frac{1}{N} \left| \sum_{t=1}^{N} X_t e^{-i2\pi ft} \right|^2,
\end{aligned}$$

$\hat{S}^{(p)}(f)$ defined above is known as the *periodogram*, and is defined over $[-1/2, 1/2]$.

Note that $\{s_\tau^{(p)}\}$ and $\hat{S}^{(p)}(f)$,

$$\{s_\tau^{(p)}\} \longleftrightarrow \hat{S}^{(p)}(f)$$

just like the process quantities

$$\{s_\tau\} \longleftrightarrow S(f).$$

Hence, $\{s_\tau^{(p)}\}$ can be written as

$$s_\tau^{(p)} = \int_{-1/2}^{1/2} \hat{S}^{(p)}(f) e^{i2\pi f\tau}\, df \quad |\tau| \leq N-1.$$

If $\hat{S}^{(p)}(f)$ were an ideal estimator of $S(f)$ we would have

- i $E\{\hat{S}^{(p)}(f)\} \approx S(f) \quad \forall f.$
- ii $Var\{\hat{S}^{(p)}(f)\} \to 0$ as $N \to \infty$ and,
- iii $Cov\{\hat{S}^{(p)}(f), \hat{S}^{(p)}(f')\} \approx 0$ for $f \neq f'$.

We find that

- i is a good approximation for some processes,
- ii is patently false,
- iii holds if f and f' are certain distinct frequencies, namely, the Fourier frequencies $f_k = k/N \quad (\Delta t = 1)$.

We firstly look at the expectation in i. (assuming $\mu = 0$).

$$\begin{aligned}
E\{\hat{S}^{(p)}(f)\} &= \sum_{\tau=-(N-1)}^{(N-1)} E\{s_\tau^{(p)}\} e^{-i2\pi f\tau} \\
&= \sum_{\tau=-(N-1)}^{(N-1)} \left(1 - \frac{|\tau|}{N}\right) s_\tau e^{-i2\pi f\tau}.
\end{aligned}$$

Hence, if we know the acvs $\{s_\tau\}$ we can work out from this what $E\{\hat{S}^{(p)}(f)\}$ will be.

We can obtain much more insight by considering:

$$E\{|J(f)|^2\} \quad \text{where} \quad J(f) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} X_t e^{-i2\pi ft}, \quad |f| \le \frac{1}{2}.$$

as $\hat{S}^{(p)}(f) = |J(f)|^2$.

We know from the spectral representation theorem that,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi f' t}\, dZ(f'),$$

so that,

$$\begin{aligned} J(f) &= \sum_{t=1}^{N} \left(\int_{-1/2}^{1/2} \frac{1}{\sqrt{N}} e^{i2\pi f' t}\, dZ(f') \right) e^{-i2\pi ft} \\ &= \int_{-1/2}^{1/2} \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi (f-f')t}\, dZ(f') \end{aligned}$$

We find that,

$$\begin{aligned} E\{\hat{S}^{(p)}(f)\} &= E\{|J(f)|^2\} = E\{J^*(f)J(f)\} \\ &= \int_{-1/2}^{1/2} \mathcal{F}(f-f')S(f')\, df', \end{aligned}$$

where $\mathcal{F}$ is Féjer's kernel defined by

$$\mathcal{F}(f) = \left| \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi ft} \right|^2 = \frac{\sin^2(N\pi f)}{N \sin^2(\pi f)}.$$

This result tells us that the expected value of $\hat{S}^{(p)}(f)$ is the true spectrum convolved with Féjer's kernel.

Properties of Féjer's kernel:

(a) For all integers $N \ge 1, \mathcal{F}(f) \to N$ as $f \to 0$.

(b) For $N \ge 1$, $f \in [-1/2, 1/2]$ and $f \ne 0$, $\mathcal{F}(f) < \mathcal{F}(0)$.

(c) For $f \in [-1/2, 1/2]$, $f \ne 0$, $\mathcal{F}(f) \to 0$ as $N \to \infty$.

(d) For any integer $k \ne 0$ such that $f_k = k/N \in [-1/2, 1/2]$, $\mathcal{F}(f_k) = 0$.

(e) $\int_{-1/2}^{1/2} \mathcal{F}(f)\, df = 1$.

$\mathcal{F}(f)$ is symmetric about the origin and consists of a broad central peak ("lobe") and $N-2$ sidelobes which decrease as f increases. From (a), (c) and (e) it follows that as $N \to \infty$, $\mathcal{F}(f)$ acts as a *Dirac δ function*, with an infinite spike at $f = 0$.

For a process with large dynamic range, defined as

$$10 \log_{10} \left(\frac{\max_f S(f)}{\min_f S(f)} \right)$$

as the expected value of the periodogram is a convolution of Féjer's kernel and the true spectrum, power from parts of the spectrum where $S(f)$ is large can "leak" via the sidelobes to other frequencies where $S(f)$ is small.

Bias reduction – Tapering
To reduce the bias in the periodogram we can use a technique called tapering.

Let $X_1, X_2, \ldots, X_N$ be a portion of length N of a zero mean stationary process with sdf $S(f)$. We form the product $\{h_t X_t\}$ where $\{h_t\}$ is a sequence of real-valued constants called a data taper . Define

$$J(f) = \sum_{t=1}^{N} h_t X_t e^{-i2\pi ft} \qquad |f| \leq 1/2.$$

By the spectral representation theorem,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi f' t} \, dZ(f'),$$

so that,

$$\begin{aligned} J(f) &= \sum_{t=1}^{N} h_t \left(\int_{-1/2}^{1/2} e^{i2\pi f' t} \, dZ(f') \right) e^{-i2\pi ft} \\ &= \int_{-1/2}^{1/2} \sum_{t=1}^{N} h_t e^{-i2\pi (f-f')t} \, dZ(f') = \int_{-1/2}^{1/2} H(f-f') \, dZ(f'), \end{aligned}$$

where,

$$H(f) = \sum_{t=1}^{N} h_t e^{-i2\pi ft} \quad \text{i.e.,} \quad \{h_t\} \longleftrightarrow H(f).$$

Let,

$$\hat{S}^{(d)}(f) = |J(f)|^2 = \left| \sum_{t=1}^{N} h_t X_t e^{-i2\pi ft} \right|^2 .$$

Then,

$$\begin{aligned} |J(f)|^2 &= J^*(f) J(f) \\ &= \int_{-1/2}^{1/2} H^*(f-f') \, dZ^*(f') \int_{-1/2}^{1/2} H(f-f'') \, dZ(f''). \end{aligned}$$

Hence

$$E\{\hat{S}^{(d)}(f)\} = E\{|J(f)|^2\} = \int_{-1/2}^{1/2} |H(f-f')|^2 S(f')\,df'$$
$$= \int_{-1/2}^{1/2} \mathcal{H}(f-f') S(f')\,df',$$

where $\mathcal{H}(f-f') = |H(f-f')|^2$, i.e.,

$$\mathcal{H}(f) = \left| \sum_{t=1}^{N} h_t e^{-i2\pi f t} \right|^2 .$$

We take,

$$\sum_{t=1}^{N} h_t^2 = 1.$$

A spectral estimator of the form of $\hat{S}^{(d)}(f)$ is called a *direct spectral estimator* (hence the (d)).

Note, if $h_t = \frac{1}{\sqrt{N}}$ for $1 \le t \le N$, then

$$\hat{S}^{(d)}(f) = \hat{S}^{(p)}(f) \quad \text{and} \quad \mathcal{H}(f) = \mathcal{F}(f),$$

i.e., $\hat{S}^{(d)}(f)$ is the same as the periodogram, and $\mathcal{H}(f)$ is the same as Féjer's kernel.

The key idea behind tapering is to select $\{h_t\}$ so that $\mathcal{H}(f)$ has much lower sidelobes that $\mathcal{F}(f)$. Recall that $\mathcal{F}(f)$ corresponds to a rectangular taper

$$h_t = \begin{cases} \frac{1}{\sqrt{N}} & \text{for } 1 \le t \le N, \\ 0 & \text{otherwise.} \end{cases}$$

There is thus a sharp discontinuity between where the taper is "ON" ($1 \le t \le N$) and where it is "OFF". Tapering effectively creates a smooth transition at the ends of the data.

Parametric model fitting

We focus on $AR(p)$ models, for which the sdf is

$$S(f) = \frac{\sigma^2}{|1 - \phi_{1,p} e^{-i2\pi f} - \ldots - \phi_{p,p} e^{-i2\pi f p}|^2}.$$

This class of models is appealing for several reasons.

(i) Any time series with a purely continuous sdf can be approximated well by an AR(p) model if p is large enough.

(ii) There exist efficient algorithms for fitting AR(p) models to time series.

(iii) Quite a few physical phenomena are reverberant and hence an AR model is naturally appropriate.

The Yule-Walker Method

We start by multiplying the defining equation by X_{t-k}:

$$X_t X_{t-k} = \sum_{j=1}^{p} \phi_{j,p} X_{t-j} X_{t-k} + \epsilon_t X_{t-k}.$$

Taking expectations, for $k > 0$:

$$s_k = \sum_{j=1}^{p} \phi_{j,p} s_{k-j}.$$

Let $k = 1, 2, \ldots, p$ and recall that $s_{-\tau} = s_\tau$ to obtain

$$\begin{aligned} s_1 &= \phi_{1,p} s_0 + \phi_{2,p} s_1 + \ldots + \phi_{p,p} s_{p-1} \\ s_2 &= \phi_{1,p} s_1 + \phi_{2,p} s_0 + \ldots + \phi_{p,p} s_{p-2} \\ \vdots & \quad \vdots \\ s_p &= \phi_{1,p} s_{p-1} + \phi_{2,p} s_{p-2} + \ldots + \phi_{p,p} s_0 \end{aligned}$$

or in matrix notation, $\boldsymbol{\gamma}_p = \Gamma_p \boldsymbol{\phi}_p$, where $\boldsymbol{\gamma}_p = [s_1, s_2, \ldots, s_p]^T$, $\boldsymbol{\phi}_p = [\phi_{1,p}, \phi_{2,p}, \ldots, \phi_{p,p}]^T$ and

$$\Gamma_p = \begin{bmatrix} s_0 & s_1 & \ldots & s_{p-1} \\ s_1 & s_0 & \ldots & s_{p-2} \\ \vdots & \vdots & & \vdots \\ s_{p-1} & s_{p-2} & \ldots & s_0 \end{bmatrix}$$

Suppose we don't know the $\{s_\tau\}$, but the mean is zero, then take

$$\hat{s}_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|},$$

and substitute these for the s_τ's in γ and Γ_p to obtain $\hat{\gamma}_p, \hat{\Gamma}_p$, from which we estimate $\boldsymbol{\phi}_p$ as $\hat{\boldsymbol{\phi}}_p$:

$$\hat{\boldsymbol{\phi}}_p = \Gamma^{-1} \hat{\boldsymbol{\gamma}}_p.$$

Finally, we need to estimate σ_ϵ^2. To do so, we multiply the defining equation by X_t and take expectations to obtain

$$s_0 = \sum_{j=1}^{p} \phi_{j,p} s_j + E\{\epsilon_t X_t\} = \sum_{j=1}^{p} \phi_{j,p} s_j + \sigma_\epsilon^2,$$

so that as an estimator for σ_ϵ^2 we take

$$\hat{\sigma}_\epsilon^2 = \hat{s}_o - \sum_{j=1}^{p} \hat{\phi}_{j,p} \hat{s}_j.$$

The estimators $\hat{\boldsymbol{\phi}}_p$ and $\hat{\sigma}_\epsilon^2$ are called the Yule-Walker estimators of the AR(p) process.

The estimate of the sdf resulting is

$$\hat{S}(f) = \frac{\hat{\sigma}_\epsilon^2}{\left|1 - \sum_{j=1}^{p} \hat{\phi}_{j,p} e^{-i2\pi fj}\right|^2}.$$

There are important modifications which we can make to this approach: we could use for $\{\hat{s}_\tau\}$ a modified autocovariance incorporating tapering:

$$\hat{s}_\tau = \sum_{t=1}^{N-|\tau|} h_t X_t h_{t+|\tau|} X_{t+|\tau|}.$$

Levinson-Durbin
To invert $\hat{\Gamma}_p$ by brute force matrix inversion requires $O(p^3)$ operations.

Fortunately, there is an algorithm due to Levinson and Durbin which takes advantage of the highly structured nature of the Toeplitz matrix, and carries out the estimation in $O(p^2)$ or fewer operations.

Least squares estimation of the $\{\phi_{j,p}\}$
Let $\{X_t\}$ be a zero-mean AR(p) process, i.e.,

$$X_t = \phi_{1,p} X_{t-1} + \phi_{2,p} X_{t-2} + \ldots + \phi_{p,p} X_{t-p} + \epsilon_t.$$

We can formulate an appropriate least squares model in terms of data $X_1, X_2, \ldots, X_N$ as follows:

$$\mathbf{X}_F = F\boldsymbol{\phi} + \boldsymbol{\epsilon}_F,$$

where,

$$F = \begin{bmatrix} X_p & X_{p-1} & \ldots & X_1 \\ X_{p+1} & X_p & \ldots & X_2 \\ \vdots & & & \vdots \\ X_{N-1} & X_{N-2} & \ldots & X_{N-p} \end{bmatrix}$$

and,

$$\mathbf{X}_F = \begin{bmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_N \end{bmatrix}; \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_{1,p} \\ \phi_{2,p} \\ \vdots \\ \phi_{p,p} \end{bmatrix}; \quad \boldsymbol{\epsilon}_F = \begin{bmatrix} \epsilon_{p+1} \\ \epsilon_{p+2} \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

We can thus estimate ϕ by finding that ϕ such that

$$\begin{aligned} \mathrm{SS}_F(\phi) &= \sum_{t=p+1}^{N}\left(X_t - \sum_{k=1}^{p}\phi_{k,p}X_{t-k}\right)^2 \quad \left[= \sum_{t=p+1}^{N}\epsilon_t^2\right] \\ &= (\mathbf{X}_F - F\phi)^T(\mathbf{X}_F - F\phi) \end{aligned}$$

is minimized. If we denote the vector that minimizes the above as $\hat{\phi}_F$, standard least squares theory tells us that it is given by

$$\hat{\phi}_F = (F^T F)^{-1}F^T\mathbf{X}_F.$$

We can estimate the innovations variance σ_F^2 by the usual estimator of the residual variation, namely

$$\hat{\sigma}_F^2 = \frac{(\mathbf{X}_F - F\hat{\phi}_F)^T(\mathbf{X}_F - F\hat{\phi}_F)}{(N-2p)}.$$

(Note: there are $N-p$ effective observations, and p parameters are estimated).

The estimator $\hat{\phi}_F$ is known as the *forward* least squares estimator of ϕ.

Using a *time reversed* formulation;

$$\mathbf{X}_B = B\phi + \boldsymbol{\epsilon}_B,$$

where,

$$B = \begin{bmatrix} X_2 & X_3 & \dots & X_{p+1} \\ X_3 & X_4 & \dots & X_{p+2} \\ \vdots & & & \vdots \\ X_{N-p+1} & X_{N-p+2} & \dots & X_N \end{bmatrix}$$

and,

$$\mathbf{X}_B = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{N-p} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_B = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{N-p} \end{bmatrix}.$$

The function of ϕ to be minimized is now

$$\begin{aligned} \mathrm{SS}_B(\phi) &= \sum_{t=1}^{N-p}\left(X_t - \sum_{k=1}^{p}\phi_{k,p}X_{t+k}\right)^2 \\ &= (\mathbf{X}_B - B\phi)^T(\mathbf{X}_B - B\phi) \end{aligned}$$

The *backward* least squares estimator of ϕ is then given by

$$\hat{\phi}_B = (B^T B)^{-1}B^T\mathbf{X}_B.$$

The corresponding estimator of the innovations variance σ_B^2 is

$$\hat{\sigma}_B^2 = \frac{(\mathbf{X}_B - B\phi)^T(\mathbf{X}_B - B\phi)}{(N-2p)}.$$

The vector $\hat{\phi}_{FB}$ that minimizes

$$SS_F(\phi) + SS_B(\phi)$$

is called the *forward/backward* least squares estimator, and Monte-Carlo studies indicate that it performs better than forward or backward least squares.

Notes:

- $\hat{\phi}_{FB}, \hat{\phi}_B$ and $\hat{\phi}_F$ produce estimated models which need not be stationary. This may be a concern for prediction, however, for spectral estimation, the parameter values will still produce a valid sdf (i.e., nonnegative everywhere, symmetric about the origin and integrates to a finite number).
- The Yule-Walker estimates can be formulated as a least squares problem; consider adding zeros to our observations $X_1, X_2, \ldots, X_N$, both at the beginning and end of the data, to give:

$$\mathbf{X}_{YW} = W\phi + \boldsymbol{\epsilon}_{YW},$$

$$W = \begin{bmatrix} 0 & 0 & 0 & \ldots & \ldots & 0 \\ X_1 & 0 & 0 & \ldots & \ldots & 0 \\ X_2 & X_1 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & & & & \vdots \\ X_{p-1} & \vdots & & & & 0 \\ X_p & X_{p-1} & \ldots & \ldots & \ldots & X_1 \\ \vdots & \vdots & & & & \vdots \\ X_N & X_{N-1} & \ldots & \ldots & \ldots & X_{N-p+1} \\ 0 & X_N & & & & X_{N-p+2} \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & & & & X_N \end{bmatrix}$$

Therefore

$$\mathbf{X}_{YW} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_{YW} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\frac{1}{N}W^TW = \begin{bmatrix} \hat{s}_0^{(p)} & \hat{s}_1^{(p)} & \dots & \hat{s}_{p-1}^{(p)} \\ \hat{s}_1^{(p)} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \hat{s}_{p-1}^{(p)} & \dots & \dots & \hat{s}_0^{(p)} \end{bmatrix} = \hat{\Gamma}_p$$

and

$$\frac{1}{N}W^T\mathbf{X}_{YW} = \begin{bmatrix} \hat{s}_1^{(p)} \\ \vdots \\ \hat{s}_p^{(p)} \end{bmatrix} = \hat{\gamma}_p,$$

so that

$$(W^TW)^{-1}W^T\mathbf{X}_{YW} = (\hat{\Gamma}_p)^{-1}\hat{\gamma}_p.$$

which is identical to the Yule-Walker estimate.

Forecasting

Suppose we wish to predict the value of X_{t+l} of a process, given $X_t, X_{t-1}, X_{t-2}, \ldots$. Let the appropriate model for $\{X_t\}$ be an ARMA(p,q) process:

$$\Phi(B)X_t = \Theta(B)\epsilon_t.$$

Consider a forecast $X_t(l)$ of X_{t+1} (an l-step ahead forecast) which is a linear combination of $X_t, X_{t-1}, X_{t-2}, \ldots$:

$$X_t(l) = \sum_{k=0}^{\infty} \pi_k X_{t-k}.$$

Note: this assumes a semi-infinite realization of $\{X_t\}$. Let us now assume that $\{X_t\}$ can be written as a one-sided linear process, so that

$$X_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k} = \Psi(B)\epsilon_t,$$

and

$$X_{t+l} = \sum_{k=0}^{\infty} \psi_k \epsilon_{t+l-k} = \Psi(B)\epsilon_{t+l}.$$

Hence,

$$X_t(l) = \sum_{k=0}^{\infty} \pi_k X_{t-k} \quad = \quad \sum_{k=0}^{\infty} \pi_k \Psi(B)\epsilon_{t-k} = \Pi(B)\Psi(B)\epsilon_t.$$

Let $\delta(B) = \Pi(B)\Psi(B)$ so that,

$$X_t(l) = \delta(B)\epsilon_t = \sum_{k=0}^{\infty} \delta_k \epsilon_{t-k}.$$

Now,

$$X_{t+l} = \sum_{k=0}^{\infty} \psi_k \epsilon_{t+l-k} = \sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k} + \sum_{k=l}^{\infty} \psi_k \epsilon_{t+l-k} = (A) + (B)$$

- (A) Involves future ϵ_ts, represents the "unpredictable" part of X_{t+l}.
- (B) Depends only on past and present values of ϵ_t, represents the "predictable" part of X_{t+l}.

Hence we would expect

$$X_t(l) = \sum_{k=l}^{\infty} \psi_k \epsilon_{t+l-k} = \sum_{j=0}^{\infty} \psi_{j+l} \epsilon_{t-j}.$$

so that $\delta_k \equiv \psi_{k+l}$. This can be readily proved. For linear least squares, we want to minimize,

$$\begin{aligned} E\{(X_{t+l} - X_t(l))^2\} &= E\left\{\left(\sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k} + \sum_{k=0}^{\infty} [\psi_{k+l} - \delta_k]\epsilon_{t-k}\right)^2\right\} \\ &= \sigma_\epsilon^2 \left\{\left(\sum_{k=0}^{l-1} \psi_k^2\right) + \sum_{k=0}^{\infty} (\psi_{k+l} - \delta_k)^2\right\}. \end{aligned}$$

The first term is independent of the choice of the $\{\delta_k\}$ and the second term is clearly minimized by choosing $\delta_k = \psi_{k+l}, k = 0, 1, 2, \ldots$ as expected. With this choice of $\{\delta_k\}$ the second term vanishes, and we have,

$$\begin{aligned} \sigma^2(l) &= E\{(X_{t+l} - X_t(l))^2\} \\ &= \sigma_\epsilon^2 \sum_{k=0}^{l-1} \psi_k^2, \end{aligned}$$

which is known as the l-step prediction variance.

When $l = 1$, $\delta_k = \psi_{k+1}$,

$$\begin{aligned} X_t(1) &= \delta_0 \epsilon_t + \delta_1 \epsilon_{t-1} + \delta_2 \epsilon_{t-2} + \ldots \\ &= \psi_1 \epsilon_t + \psi_2 \epsilon_{t-1} + \psi_3 \epsilon_{t-2} + \ldots \\ X_{t+1} &= \psi_0 \epsilon_{t+1} + \psi_1 \epsilon_t + \psi_2 \epsilon_{t-1} + \ldots \end{aligned}$$

so that,

$$X_{t+1} - X_t(1) = \psi_0 \epsilon_{t+1} = \epsilon_{t+1} \quad \text{since} \quad \psi_0 = 1.$$

Hence ϵ_{t+1} can be thought of as the "one step prediction error". Also of course,

$$X_{t+1} = X_t(1) + \epsilon_{t+1}$$

so that ϵ_{t+1} is the essentially "new" part of X_{t+1} which is not linearly dependent on past observations. The sequence $\{\epsilon_t\}$ is often called the innovations process of $\{X_t\}$, and σ_ϵ^2 is often called the innovations variance.

If we wish to write $X_t(l)$ explicitly as a function of $X_t, X_{t-1}, \ldots$ then we note first that,

$$X_t(l) = \sum_{k=0}^{\infty} \delta_k \epsilon_{t-k} = \sum_{k=0}^{\infty} \psi_{k+l} \epsilon_{t-k},$$

so that,

$$X_t(l) = \Psi^{(l)}(B)\epsilon_t, \quad \text{say}$$

where,

$$\Psi^{(l)}(z) = \sum_{k=0}^{\infty} \psi_{k+l} z^k.$$

Assuming that $\Psi(z)$ is analytic in and on the unit circle (stationary and invertible) then we can write

$$X_t = \Psi(B)\epsilon_t \quad \text{and} \quad \epsilon_t = \Psi^{-1}(B)X_t,$$

and thus

$$\begin{aligned} X_t(l) = \Psi^{(l)}(B)\epsilon_t &= \Psi^{(l)}(B)\Psi^{-1}(B)X_t \\ &= G^{(l)}(B)X_t, \quad \text{say} \end{aligned}$$

with,

$$G^{(l)}(z) = \Psi^{(l)}(z)\Psi^{-1}(z).$$

If we consider the sequence of predictors $X_t(l)$ for different values of t (with l fixed) then this forms a new process, which since

$$X_t(l) = G^{(l)}(B)X_t,$$

may be regarded as the output of a linear filter acting on the $\{X_t\}$. Since,

$$X_t(l) = \left(\sum_u g_u^{(l)} B^u\right) X_t = \sum_u g_u^{(l)} X_{t-u},$$

we know that the transfer function is

$$G^{(l)}(f) = \sum_u g_u^{(l)} e^{-i2\pi f u}.$$

Example: AR(1)

$$X_t - \phi_{1,1}X_{t-1} = \epsilon_t \qquad |\phi_{1,1}| < 1.$$

Then

$$X_t = (1-\phi_{1,1}B)^{-1}\epsilon_t.$$

So,

$$\begin{aligned}\Psi(z) &= 1 + \phi_{1,1}z + \phi_{1,1}^2 z^2 + \ldots \\ &= \psi_0 + \psi_1 z + \psi_2 z^2 + \ldots\end{aligned}$$

i.e., $\psi_k = \phi_{1,1}^k$.

Hence,

$$\begin{aligned}X_t(l) &= \sum_{k=0}^{\infty}\delta_k\epsilon_{t-k} = \sum_{k=0}^{\infty}\psi_{k+l}\epsilon_{t-k} \\ &= \sum_{k=0}^{\infty}\phi_{1,1}^{k+l}\epsilon_{t-k} = \phi_{1,1}^l\sum_{k=0}^{\infty}\phi_{1,1}^k\epsilon_{t-k} \\ &= \phi_{1,1}^l X_t.\end{aligned}$$

The l-step prediction variance is

$$\sigma^2(l) = \sigma_\epsilon^2\left(\sum_{k=0}^{l-1}\psi_k^2\right) = \sigma_\epsilon^2\left(\sum_{k=0}^{l-1}\phi_{1,1}^{2k}\right) = \sigma_\epsilon^2\frac{(1-\phi_{1,1}^{2l})}{(1-\phi_{1,1}^2)}.$$

Alternatively,

$$X_t(l) = G^{(l)}(B)X_t,$$

with $G^{(l)}(z) = \Psi^{(l)}(z)\Psi^{-1}(z)$. But,

$$\Psi^{(l)}(z) = \sum_{k=0}^{\infty}\psi_{k+l}z^k = \sum_{k=0}^{\infty}\phi_{1,1}^{k+l}z^k,$$

and,

$$\Psi^{-1}(z) = 1 - \phi_{1,1}z,$$

so that

$$G^{(l)}(z) = (\phi_{1,1}^l + \phi_{1,1}^{l+1}z + \phi_{1,1}^{l+2}z^2 + \ldots)(1-\phi_{1,1}z) = \phi_{1,1}^l,$$

i.e., $X_t(l) = \phi_{1,1}^l X_t$ as before.

We have demonstrated that for the AR(1) model the linear least squares predictor of X_{t+l} depends only on the most recent observation, X_t, and does not involve $X_{t-1}, X_{t-2}, \ldots$, which is what we would expect bearing in mind the Markov nature of the AR(1) model. As $l \to \infty$, $X_t(l) \to 0$, since $X_t(l) = \phi_{1,1}^l X_t$ and $|\phi_{1,1}| < 1$. Also, the l-step prediction variance,

$$\sigma^2(l) \to \frac{\sigma_\epsilon^2}{(1-\phi_{1,1}^2)} = Var\left[X_t\right].$$

In fact the solution to the forecasting problem for the AR(1) model can be derived directly from the difference equation,

$$X_t - \phi_{1,1} X_{t-1} = \epsilon_t.$$

by setting future innovations ϵ_t to be zero:

$$\begin{aligned} X_t(1) &= \phi_{1,1} X_t + 0 \\ X_t(2) &= \phi_{1,1} X_t(1) + 0 \\ &\vdots \\ X_t(l) &= \phi_{1,1} X_t(l-1) + 0 \end{aligned}$$

so that,

$$X_t(l) = \phi_{1,1}^l X_t.$$

For general AR(p) processes it turns out that $X_t(l)$ depends only on the last p observed values of $\{X_t\}$, and may be obtained by solving the AR(p) difference equation with the future $\{\epsilon_t\}$ set to zero. For example for an AR(p) process and $l = 1$,

$$X_t(1) = \phi_{1,p} X_t + \ldots + \phi_{p,p} X_{t-p+1}.$$

Example: ARMA(1,1)

$$(1 - \phi_{1,1} B) X_t = (1 - \theta_{1,1} B)\epsilon_t.$$

Take $\phi_{1,1} = \phi$ and $\theta_{1,1} = \theta$,

$$X_t = \frac{(1 - \theta B)}{(1 - \phi B)} \epsilon_t = \Psi(B)\epsilon_t.$$

So,

$$\begin{aligned} \Psi(z) &= (1 - \theta z)(1 + \phi z + \phi^2 z^2 + \phi^3 z^3 + \ldots) \\ &= 1 + (\phi - \theta) z + \phi(\phi - \theta) z^2 + \ldots + \phi^{l-1}(\phi - \theta) z^l + \ldots \\ &= \psi_0 + \psi_1 z + \psi_2 z^2 + \ldots \end{aligned}$$

So,

$$\psi_l = \begin{cases} 1 & l = 0 \\ \phi^{l-1}(\phi - \theta) & l \geq 1 \end{cases}$$

The l-step prediction variance is

$$\begin{aligned} \sigma^2(l) &= \sigma_\epsilon^2 \left(\sum_{k=0}^{l-1} \psi_k^2 \right) = \sigma_\epsilon^2 \left(1 + \sum_{k=1}^{l-1} \psi_k^2 \right) \\ &= \sigma_\epsilon^2 \left(1 + (\phi - \theta)^2 \sum_{k=1}^{l-1} \phi^{2k-2} \right) \\ &= \sigma_\epsilon^2 \left(1 + (\phi - \theta)^2 \frac{(1 - \phi^{2l-2})}{(1 - \phi^2)} \right). \end{aligned}$$

Now,

$$\begin{aligned}
\Psi^{(l)}(z) &= \sum_{k=0}^{\infty} \psi_{k+l} z^k = \phi^{l-1}(\phi-\theta)\sum_{k=0}^{\infty}\phi^k z^k = \phi^{l-1}(\phi-\theta)(1-\phi z) \\
\Psi^{-1}(z) &= \frac{(1-\phi z)}{(1-\theta z)},
\end{aligned}$$

so therefore

$$\begin{aligned}
G^{(l)}(z) &= \Psi^{(l)}(z)\Psi^{-1}(z) = \phi^{l-1}(\phi-\theta)(1-\theta z)^{-1} \\
X_t(l) &= G^{(l)}(B)X_t = \phi^{l-1}(\phi-\theta)(1-\theta B)^{-1}X_t.
\end{aligned}$$

Consider $l = 1$,

$$\begin{aligned}
X_t(1) &= (\phi-\theta)(1-\theta B)^{-1}X_t \\
&= (\phi-\theta)(1+\theta B+\theta^2B^2+\theta^3B^3+\ldots)X_t \\
&\vdots \\
&= (\phi-\theta)X_t + \theta(\phi-\theta)X_{t-1} + \theta^2(\phi-\theta)X_{t-2} + \ldots \\
&= \phi X_t - \theta\left[X_t - (\phi-\theta)X_{t-1} - \ldots - \theta^{k-1}(\phi-\theta)X_{t-k} - \ldots\right]
\end{aligned}$$

But consider,

$$\begin{aligned}
\epsilon_t &= \Psi^{-1}(B)X_t = (1-\phi B)(1-\theta B)^{-1}X_t \\
&= (1-\phi B)(1+\theta B+\theta^2B^2+\theta^3B^3+\ldots)X_t \\
&\vdots \\
&= X_t - (\phi-\theta)X_{t-1} - \ldots - \theta^{k-1}(\phi-\theta)X_{t-k} - \ldots.
\end{aligned}$$

Therefore,

$$X_t(1) = \phi X_t - \theta\epsilon_t.$$

So can again be derived directly from the difference equation,

$$X_t = \phi X_{t-1} - \theta\epsilon_{t-1} + \epsilon_t,$$

by setting future innovations ϵ_t to zero.

MA(1) (invertible)

$$X_t = \epsilon_t - \theta_{1,1}\epsilon_{t-1} \qquad |\theta_{1,1}| < 1.$$

So,

$$\begin{aligned}
\Psi(z) &= \psi_0 + \psi_1 z + \psi_2 z^2 + \ldots \\
&= 1 - \theta_{1,1} z
\end{aligned}$$

Hence, $\psi_0 = 1$; $\psi_1 = -\theta_{1,1}$; $\psi_k = 0$, $k \geq 2$.

$$\begin{aligned}
X_t(l) &= \sum_{k=0}^{\infty}\psi_{k+l}\epsilon_{t-k} = \Psi^{(l)}(B)\epsilon_t \\
&= \psi_l\epsilon_t + \psi_{l+1}\epsilon_{t-1} + \ldots
\end{aligned}$$

So,

$$\Psi^{(l)}(z) = \sum_{k=0}^{\infty} \psi_{k+l} z^k = \psi_l z^0 + \psi_{l+1} z^1 = \begin{cases} -\theta_{1,1} & l = 1 \\ 0 & l \geq 2. \end{cases}$$

Hence,

$$G^{(l)}(z) = \Psi^{(l)}(z)\Psi^{-1}(z) = \begin{cases} -\theta_{1,1}(1-\theta_{1,1}z)^{-1} & l = 1 \\ 0 & l \geq 2. \end{cases}$$

Thus, for $l = 1$,

$$G^{(1)}(z) = -\theta_{1,1}(1 + \theta_{1,1}z + \theta_{1,1}^2 z^2 + \ldots),$$

and hence,

$$X_t(1) = G^{(1)}(B)X_t = -\sum_{k=0}^{\infty} \theta_{1,1}^{k+1} X_{t-k}$$

Forecast errors and updating

We have seen that when $\delta_k = \psi_{k+l}$ the forecast error is

$$\sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k}.$$

Let,

$$e_t(l) = X_{t+l} - X_t(l) = \sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k}.$$

Then,

$$e_t(l+m) = \sum_{j=0}^{l+m-1} \psi_j \epsilon_{t+l+m-j}.$$

Clearly,

$$E\{e_t(l)\} = E\{e_t(l+m)\} = 0.$$

Hence,

$$Cov\{e_t(l), e_t(l+m)\} = E\{e_t(l)e_t(l+m)\} = \sigma_\epsilon^2 \sum_{k=0}^{l-1} \psi_k \psi_{k+m}.$$

and

$$Var\{e_t(l)\} = \sigma_\epsilon^2 \sum_{k=0}^{l-1} \psi_k^2 = \sigma^2(l).$$

E.g.,

$$Cov\{e_t(1), e_t(2)\} = \sigma_\epsilon^2 \psi_1.$$

This could be quite large – should the forecast for a series wander of target, it is possible for it to remain there in the short run since forecast errors can be quite highly correlated. Hence, when X_{t+1} becomes available we should update the forecast.

$$\begin{aligned} X_{t+1}(l) &= \sum_{k=0}^{\infty} \psi_{k+l} \epsilon_{t+1-k} \\ &= \psi_l \epsilon_{t+1} + \psi_{l+1} \epsilon_t + \psi_{l+2} \epsilon_{t-1} + \dots, \end{aligned}$$

$$\begin{aligned} X_t(l+1) &= \sum_{k=0}^{\infty} \psi_{k+l+1} \epsilon_{t-k} \\ &= \psi_{l+1} \epsilon_t + \psi_{l+2} \epsilon_{t-1} + \psi_{l+3} \epsilon_{t-2} + \dots, \end{aligned}$$

and,

$$\begin{aligned} X_{t+1}(l) &= X_t(l+1) + \psi_l \epsilon_{t+1} \\ &= X_t(l+1) + \psi_l (X_{t+1} - X_t(1)). \end{aligned}$$

Hence, to forecast X_{t+l+1} we can modify the $l+1$- step ahead forecast at time t by producing an l-step ahead forecast at time $t+1$ using X_{t+1} as it becomes available.

Non-stationarity and Unit Roots

Many financial/econometric series are *trending*.

Two cases commonly considered;

1. Stationary process with **deterministic** trend (shocks have temporary effects)
2. Process with **stochastic** trend or **unit root** (shocks have permanent effects)

The distinction between the two cases is practically important for forecasting and statistical issues.

Trend Stationarity

Example: Consider an AR(1) model with *deterministic linear trend*

$$Y_t = \phi Y_{t-1} + \delta + \gamma t + \epsilon_t \qquad t = 1, \dots, N,$$

with $|\phi| < 1$. Then, as $N \longrightarrow \infty$,

$$E[Y_t] \longrightarrow \mu + \mu_1 t \qquad Var[Y_t] \longrightarrow \frac{\sigma^2}{1-\phi^2}$$

using the MA representation.

- Y_t is not stationary, but the deviation from the mean

$$X_t = Y_t - \mu - \mu_1 t$$

is stationary; Y_t is termed *trend-stationary*.
- The stochastic part is stationary, and shocks have transitory effects.
- Y_t is *mean-reverting*, with *attractor* $\mu + \mu_1 t$.

We can analyze X_t as a stationary process.

Unit Root Processes

Example: Consider an AR(1) model with a *unit root* $\phi = 1$

$$Y_t = Y_{t-1} + \delta + \epsilon_t$$

or

$$BY_t = \delta + \epsilon_t.$$

- $z = 1$ is a root of the AR polynomial $\Phi(z) = 1 - z$.
- Y_t is non-stationary.
- BY_t is stationary, Y_t termed a *difference stationary process*.
- Y_t is termed an *integrated first order process*, or an $I(1)$ process.
- A process of *integrated order* d is denoted $I(d)$.

Note that

$$Y_t = Y_0 + \sum_{i=1}^{t} BY_t = Y_0 + \delta t + \sum_{i=1}^{t} \epsilon_t$$

with moments

$$E[Y_t] = Y_0 + \delta t \qquad V[Y_t] = t\sigma^2$$

- Y_0 remains in the process.
- ϵ_t accumulates as a random walk, termed a *stochastic trend*. These shocks have a permanent effect.
- δ forms a deterministic linear trend.
- This model is termed a *random walk with drift*.
- Variance grows with t.
- Not mean-reverting.

Unit Root Tests

We consider null and alternative hypotheses to distinguish between stationarity and non-stationarity.

(1) Dickey-Fuller Test
   - H_0 is a unit root, H_1 is stationarity

(2) KPSS Test
   - H_0 is stationarity, H_1 is a unit root

Note: In practice, distinguishing $\phi = 0.99$ from $\phi = 1$ is often difficult ...

Dickey-Fuller Test Set up an AR model for de-trended process X_t and test $\phi = 1$.

- Consider AR(1) model

$$X_t = \theta X_{t-1} + \epsilon_t$$

We wish to test

$$H_0 : \phi = 1 \qquad \text{against} \qquad H_1 : \phi < 1.$$

- Rewrite model as

$$BX_t = (\phi - 1)X_{t-1} + \epsilon_t = \pi X_{t-1} + \epsilon_t$$

with $\pi = \phi - 1 = \Phi(1)$, say, and the hypotheses as

$$H_0 : \pi = 0 \qquad \text{against} \qquad H_1 : \pi < 0.$$

The Dickey-Fuller (DF) test is the Wald t-test for H_0 with test statistic t_{DF}

$$t_{DF} = \frac{\widehat{\phi} - 1}{se(\widehat{\phi})} = \frac{\widehat{\pi}}{se(\widehat{\pi})}$$

- The asymptotic null distribution is non-normal, and depends on the deterministic part of the model.
- The asymptotic null only holds if ϵ_t are IID.
- If not IID, need to include further terms in AR representation.
- MA and ARMA models handled similarly.

Extension to AR(p): The **Augmented Dickey-Fuller** (ADF) Test.

Example: AR(3).

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \epsilon_t$$

A unit root of

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3 = 0$$

corresponds to $\Phi(1) = 0$.

Test is achieved by rewriting the model as

$$BX_t = \pi X_{t-1} + c_1 BX_{t-1} + c_2 BX_{t-2} + \epsilon_t$$

where

$$\begin{aligned} \pi &= \phi_1 + \phi_2 + \phi_3 - 1 = -\phi(1) \\ c_1 &= -(\phi_2 + \phi_3) \\ c_2 &= -\phi_3 \end{aligned}$$

- Null hypothesis $\Phi(1) = 0$ corresponds to

$$H_0 : \pi = 0 \qquad \text{against} \qquad H_1 : \pi < 0.$$

- The ADF test is the Wald t-test of this hypothesis.
- Need model selection to choose number of lags.
- Can correct for autocorrelation in ϵ_t - use the *Phillips-Perron* test that uses a standard ergodic estimate of the autocorrelation (*Newey-West*).

Note: The deterministic terms in the ADF specification are important, as they influence the asymptotic null distribution.

- if X_t has a non-zero level, use

$$BY_t = \pi Y_{t-1} + c_1 BX_{t-1} + c_2 BX_{t-2} + \delta\epsilon_t$$

- if X_t has a deterministic trend level, use

$$BY_t = \pi Y_{t-1} + c_1 BX_{t-1} + c_2 BX_{t-2} + \delta + \gamma t + \epsilon_t$$

In both cases, can fit model using regression methods.

In both cases, the null distribution changes.

Note: consider the factor representation

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \\ Y_t &= X_t + \mu \end{aligned}$$

so that

$$Y_t = \phi Y_{t-1} + (1-\phi)\mu + \epsilon_t = \phi Y_{t-1} + \delta + \epsilon_t$$

so there is a common factor restriction; if $\phi = 1$,

$$\delta = (1-\phi)\mu = 0.$$

This is not imposed by the standard t-test; consider

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

The hypotheses

$$H_0 : \phi = 1 \qquad \text{against} \qquad H_1 : \phi < 1.$$

imply

$$\begin{aligned} H_1 &: \quad Y_t = \mu + \text{ stationary process} \\ H_0 &: \quad Y_t = Y_0 + \delta t + \text{ stochastic trend.} \end{aligned}$$

that is, two fundamentally different models.

Need to consider the combined null hypothesis

$$H_0^C : \pi = \delta = 0$$

which can be tested by fitting two regressions

$$\begin{aligned} H_1 &: \quad BY_t = \pi Y_{t-1} + \delta + \epsilon_t \\ H_0^C &: \quad BY_t = \epsilon_t. \end{aligned}$$

and carrying out a likelihood ratio test to compare the fits.

Again, the null distribution is non-standard.

Alternatively, consider the model with a trend

$$BY_t = \pi Y_{t-1} + \delta + \gamma t + \epsilon_t$$

where the common factor restriction implies that if $\pi = 0$ then $\gamma = 0$. Under the standard null H_0, the trend will accumulate.

Again need to impose the combined null hypothesis

$$H_0^C : \pi = \gamma = 0$$

which can be tested by fitting two regressions

$$\begin{aligned} H_1 &: \quad BY_t = \pi Y_{t-1} + \delta + \gamma t \epsilon_t \\ H_0^C &: \quad BY_t = \delta + \epsilon_t. \end{aligned}$$

and carrying out a likelihood ratio test to compare the fits.

Again, the null distribution is non-standard.

Special Events: Large shocks (breaks, changepoints) have potentially large, permanent effects.

- **One large shock**: may lead to bias toward accepting unit root hypothesis, event of series is stationary.
- **Many large shocks**: may lead to bias toward accepting stationarity hypothesis. Series may appear mean-reverting even if it is not.

Kwiatkowski, Phillips, Schmidt and Shin (KPSS) Test

- Assume
$$Y_t = \xi_t + e_t$$
where e_t is stationary and ξ_t is a random walk
$$\xi_t = \xi_{t-1} + v_t$$
where $v_t \sim N(0, \sigma_v^2)$ i.i.d..
- If $\sigma_v^2 = 0$, $\xi_t = \xi_0$ and Y_t is stationary. Thus can test the hypothesis
$$H_0 : \sigma_v^2 = 0 \qquad \text{against} \qquad H_1 : \sigma_v^2 > 0.$$

The KPSS Test is a (score) test of this hypothesis.

Models For Changing Variance

Objective: obtain better estimates of local variance.

***p*'th order ARCH(*p*)**

ARCH stands for **autoregressive conditionally heteroscedastic**

Assume we have a derived time series $\{Y_t\}$ that is (approximately) uncorrelated but has a variance (volatility) that changes through time,

$$Y_t = \sigma_t \varepsilon_t \tag{1}$$

where $\{\varepsilon_t\}$ is a white noise sequence with zero mean and unit variance.

Here, σ_t represents the local conditional standard deviation of the process. Note that σ_t is not observable directly.

$\{Y_t\}$ is ARCH(p) if it satisfies equation (1) and

$$\sigma_t^2 = \alpha + \beta_{1,p}y_{t-1}^2 + \ldots + \beta_{p,p}y_{t-p}^2, \quad (2)$$

where $\alpha > 0$ and $\beta_{j,p} \geq 0, j = 1, \ldots, p$ (to ensure the variance remains positive), and y_{t-1} is the observed value of the derived time series at time $(t-1)$

Note

(a) the absence of the error term in equation (2).

(b) unconstrained estimation often leads to violation of the non-negativity constraints that are needed to ensure positive variance.

(c) quadratic form (i.e. modelling σ_t^2) prevents modelling of asymmetry in volatility (i.e. volatility tends to be higher after a decrease than after an equal increase and ARCH cannot account for this).

ARCH(1)

$$\sigma_t^2 = \alpha + \beta_{1,1}y_{t-1}^2$$

Define, $v_t = y_t^2 - \sigma_t^2 \Rightarrow \sigma_t^2 = y_t^2 - v_t$. The model can also be written:

$$y_t^2 = \alpha + \beta_{1,1}y_{t-1}^2 + v_t,$$

i.e. an AR(1) model for $\{y_t^2\}$ where the errors, $\{v_t\}$, have zero mean, but as $v_t = \sigma_t^2(\epsilon_t^2 - 1)$ the errors are heteroscedastic.

(p, q)'th order generalized autoregressive conditionally heteroscedastic model GARCH(p, q)

$\{Y_t\}$ is GARCH(p, q) if it satisfies equation (1) and

$$\sigma_t^2 = \alpha + \beta_{1,p}y_{t-1}^2 + \ldots + \beta_{p,p}y_{t-p}^2 + \gamma_{1,q}\sigma_{t-1}^2 + \ldots \gamma_{q,q}\sigma_{t-q}^2,$$

where the parameters are chosen to ensure positive variance.

Stochastic volatility models SV

Stochastic volatility models treat σ_t as an unobserved random variable which is assumed to follow a certain stochastic process. The specification for the derived series $\{Y_t\}$ is:

$$Y_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = \exp(h_t),$$

where ε_t is white noise with zero mean and unit variance, and let h_t, for example, be an AR(1) process:

$$h_t = \alpha + \beta_{1,1} h_{t-1} + \eta_t,$$

where $\{\eta_t\}$ is a white noise process with variance σ_η^2. If $|\beta_{1,1}| < 1$, h_t is stationary $\Rightarrow$ Y_t stationary.

Notes:

(a) unlike the GARCH specification, h_t (which defines in turn σ_t) is NOT deterministic.

(b) the exponential specification ensures positive conditional variance.

(c) can be further generalized by assuming, for example, h_t follows an ARMA(p,q) model.

Harmonic with additive white noise

Here $\{X_t\}$ is expressed as

$$X_t = \cos(2\pi f_0 t + \phi) + \epsilon_t$$

f_0 is a fixed frequency and $\{\epsilon_t\}$ is zero mean white noise with variance σ_ϵ^2.

Case (a) ϕ is constant.

$$E[X_t] = E[\cos(2\pi f_0 t + \phi)] + E[\epsilon_t] = \cos(2\pi f_0 t + \phi).$$

so, mean depends on $t \Rightarrow$ not stationary.

Case (b): $\phi \sim U[-\pi, \pi]$ and independent of $\{\epsilon_t\}$.

$$E[X_t] = E[\cos(2\pi f_0 t + \phi) + \epsilon_t] = E\{\cos(2\pi f_0 t + \phi)\}$$

Now,

$$\begin{aligned} E\{\cos(2\pi f_0 t + \phi)\} &= \int_{-\pi}^{\pi} \cos(2\pi f_0 t + \phi) \frac{1}{2\pi}\, d\phi \\ &= \left[\frac{\sin(2\pi f_0 t + \phi)}{2\pi} \right]_{-\pi}^{\pi} = 0. \end{aligned}$$

So $E[X_t] = 0$, and, using the fact that $\{e_t\}$ and ϕ are independent.

$$\begin{aligned} E[X_t X_{t+\tau}] &= E[[\cos(2\pi f_0 t + \phi) + \epsilon_t][\cos(2\pi f_0(t+\tau) + \phi) + \epsilon_{t+\tau}]] \\ &= E[\cos(2\pi f_0 t + \phi)\cos(2\pi f_0 t + \phi + 2\pi f_0 \tau)] + E[\epsilon_t \epsilon_{t+\tau}]. \end{aligned}$$

Recall, as $\{\epsilon_t\}$ is white noise we have,

$$E\{\epsilon_t \epsilon_{t+\tau}\} = \begin{cases} \sigma_\epsilon^2 & \text{if } \tau = 0, \\ 0 & \text{if } \tau \neq 0, \end{cases}$$

So, for $\tau = 0$,

$$Cov\{X_t, X_t\} = s_0 = E\{\cos^2(2\pi f_0 t + \phi)\} + \sigma_\epsilon^2.$$

Now,

$$\begin{aligned} E\{\cos^2(2\pi f_0 t + \phi)\} &= \int_{-\pi}^{\pi} \cos^2(2\pi f_0 t + \phi)\frac{1}{2\pi}\, d\phi \\ &= \frac{1}{2}\int_{-\pi}^{\pi} [1 + \cos(4\pi f_0 t + 2\phi)]\frac{1}{2\pi}\, d\phi = \frac{1}{2}. \end{aligned}$$

So, $s_0 = \frac{1}{2} + \sigma_\epsilon^2$, and for $\tau > 0$,

$$\begin{aligned} Cov[X_t, X_{t+\tau}] &= s_\tau = E[\cos(2\pi f_0 t + \phi)\cos(2\pi f_0 t + \phi + 2\pi f_0 \tau)] \\ &= \frac{1}{2}E[\cos(4\pi f_0 t + 2\phi + 2\pi f_0 \tau) + \cos(2\pi f_0 \tau)] \\ &= \frac{1}{2}\int_{-\pi}^{\pi} \cos(2\pi f_0 \tau)\frac{1}{2\pi}\, \mathrm{d}\phi \\ &= \frac{\cos(2\pi f_0 \tau)}{2}\left[\frac{\phi}{2\pi}\right]_{-\pi}^{\pi} = \frac{\cos(2\pi f_0 \tau)}{2} \end{aligned}$$

which does not depend on $t \Rightarrow X_t$ is stationary.

Trend removal and seasonal adjustment

There are certain, quite common, situations where the observations exhibit a trend – a tendency to increase or decrease slowly steadily over time – or may fluctuate in a periodic manner due to seasonal effects. The model is modified to

$$X_t = \mu_t + Y_t$$

- μ_t = time dependent mean.
- Y_t = zero mean stationary process.

Trend adjustment for CO^2 data: $\{X_t\}$ is monthly atmospheric CO^2 concentrations expressed in parts per million (ppm) derived from in situ air samples collected at Mauna Loa observatory, Hawaii. Monthly data from May 1988 – December 1998, giving $N = 128$. Model suggested by plot:

$$X_t = \alpha + \beta t + Y_t.$$

(a) Estimate α and β by least squares, and work with the residuals

$$\hat{Y}_t = X_t - \hat{\alpha} - \hat{\beta} t.$$

(b) Take first differences:

$$X_t^{(1)} = X_t - X_{t-1} = \alpha + \beta t + Y_t - (\alpha + \beta(t-1) + Y_{t-1}) = \beta + Y_t - Y_{t-1}.$$

Note: if $\{Y_t\}$ is stationary so is $\{Y_t^{(1)}\}$ In the case of linear trend, if we difference again:

$$\begin{aligned} X_t^{(2)} &= X_t^{(1)} - X_{t-1}^{(1)} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= (\beta + Y_t - Y_{t-1}) - (\beta + Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2}, \quad (\equiv Y_t^{(1)} - Y_{t-1}^{(1)} = Y_t^{(2)}), \end{aligned}$$

so that the effect of $\mu_t (= \alpha + \beta t)$ has been completely removed.

If μ_t is a polynomial of degree $(d-1)$ in t, then dth differences of μ_t will be zero ($d = 2$ for linear trend). Further,

$$X_t^{(d)} = \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k} = \sum_{k=0}^{d} \binom{d}{k} (-1)^k Y_{t-k}.$$

There are other ways of writing this. Define the difference operator

$$\Delta = (1 - B)$$

where $BX_t = X_{t-1}$ is the *backward shift operator* (sometimes known as the *lag operator* L – especially in econometrics). Then,

$$X_t^{(d)} = \Delta^d X_t = \Delta^d Y_t.$$

For example, for $d = 2$:

$$\begin{aligned} X_t^{(2)} &= (1-B)^2 X_t = (1-B)(X_t - X_{t-1}) \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= (\beta + Y_t - Y_{t-1}) - (\beta + Y_{t-1} - Y_{t-2}) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= (1-B)^2 Y_t = \Delta^2 Y_t. \end{aligned}$$

This notation can be incorporated into the ARMA set up; if $\{X_t\}$ is ARMA(p, q),

$$X_t = \phi_{1,p}X_{t-1} + \ldots + \phi_{p,p}X_{t-p} + \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

$$X_t - \phi_{1,p}X_{t-1} - \ldots - \phi_{p,p}X_{t-p} = \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q}$$

$$(1 - \phi_{1,p}B - \ldots - \phi_{p,p}B^p)X_t = (1 - \theta_{1,q}B - \ldots - \theta_{q,q}B^q)\epsilon_t$$

That is,

$$\Phi(B)X_t = \Theta(B)\epsilon_t$$

where

$$\Phi(B) = 1 - \phi_{1,p}B - \phi_{2,p}B^2 - \ldots - \phi_{p,p}B^p$$

$$\Theta(B) = 1 - \theta_{1,q}B - \theta_{2,q}B^2 - \ldots - \theta_{q,q}B^q$$

are known as the *associated* or *characteristic polynomials*.

Further, we can generalize the class of ARMA models to include differencing to account for certain types of non-stationarity, namely,

- X_t is called **ARIMA(**p, d, q**)** if

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\epsilon_t,$$

$$\Phi(B)\Delta^d X_t = \Theta(B)\epsilon_t.$$

Seasonal adjustment
The model is modified to

$$X_t = s_t + Y_t$$

where

- $\{s_t\}$ is the **seasonal** component,
- $\{Y_t\}$ is zero mean **stationary** process.

Presuming that the seasonal component maintains a constant pattern over time with period s, there are again several approaches to removing s_t. A popular approach used by Box & Jenkins is to use the operator $(1 - B^s)$:

$$
\begin{aligned}
X_t^{(s)} &= (1 - B^s)X_t = X_t - X_{t-s} \\
&= (s_t + Y_t) - (s_{t-s} + Y_{t-s}) \\
&= Y_t - Y_{t-s}
\end{aligned}
$$

since s_t has period s (and so $s_{t-s} = s_t$).