Stanford STATS 207 Notes

Yikun Chi

September 26, 2022

Contents

Chapter 1

Baisc Time Series Model

1.1 Basic Definition

1.1.1 Stochastic Processes

Definition:
A discrete time stochastic process is a set of random variables indexed by $\mathbb{N} = \{1, 2, ...\}$

Notation:
$\{x_t\}, \{x_t\}_{t\in\mathbb{N}}$

Realization:
A realization is the observed value of the stochastic process.

Time Series Model:
A time series model specifies the joint distribution of the sequence of $\{x_t\}$ of random variables. i.e.:

$$Pr(x_1 \leq c_1, ..., x_t \leq c_t) \qquad \forall t, c_1, ..., c_t$$

But traditionally, we focus on second order properties. i.e.:

$$E[x_t], \qquad E\left[(x_s - E[x_s])(x_t - E[x_t])\right]$$

1.1.2 Stationary

Definition:
$\{x_t\}$ is strictly stationary if for all $k, t_1, ..., t_k, c_1, ..., c_k, h$, we have

$$Pr(x_{t_1} \leq c_1, ..., x_{t_k} \leq c_k) = Pr(x_{t_1+h} \leq c_1, ..., x_{t_k+h} \leq c_k)$$

i.e.: shifting time axis does not affect distribution

Second Order Properties - Mean Function:

$$\mu_{xt} = E[x_t]$$

Second Order Properties - Autocovariance function:

$$\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_{xs})(x_t - \mu_{xt})]$$

Weak stationarity:
x_t is weakly stationary if for all t, s, we have

$\mu_{xt} = \mu_{xs}$	mean does not vary with time
$\gamma_x(s,t) = \gamma_x(0, \lvert t-s \rvert)$	Only depends on distance b/t t and s

- For weak stationarity, we use shortcut representation $\gamma_x(h) = \gamma_x(0,h)$

Second Order Properties - Autocorrelation:

$$\rho_x(s,t) = corr(x_s, x_t) = \frac{cov(x_s, x_t)}{\sqrt{var(x_s)var(x_t)}} = \frac{\gamma_x(s,t)}{\sqrt{\gamma_x(s,s)\gamma_x(t,t)}}$$

- $\rho_x(t,t) = 1$
- $-1 \leq \rho_x(s,t) \leq 1$
- $|\rho_x(s,t) = 1| \rightarrow$ perfect linear relations between x_s and x_t
- If the process is weakly stationary, we have $\rho_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)}$

1.1.3 Estimating Autocovariance / Autocorrelation

Assume observations $x_1, ..., x_n$ and stationary process, we have

$$\text{Sample mean: } \hat{\mu_{xt}} = \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

$$\text{Sample autocov: } \hat{\gamma}_x(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})$$

$$\text{Sample autocorr: } \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

- Note for sample autocov, we are normalizing using $\frac{1}{n}$, and also subtracting from the full sample mean $\bar{x}$ instead of only restrict to $n-h$.

1.1.3.1 Testing for WN

If we can derive confidence limits for sample autocorrelation function, then we can test for white noise. We have:

$$SE[\hat{\rho}_{WN}(h)] \approx \frac{1}{\sqrt{n}}$$
$$\implies \sqrt{n}\hat{\rho}_{WN}(h) \overset{D}{\sim} N(0,1)$$

So for $h > 0$, we have $Pr(|\hat{\rho}_{WN}(h)| > 1.96/\sqrt{n}) \approx Pr(|N(0,1)| > 1.96/\sqrt{n}) = 0.05$

1.2 Basic Model: White Noise

$$x_t \sim WN(0, \sigma^2) \quad if$$

- Zero Mean: $E[x_t] = 0$ for all $t = 1, 2, ...$
- Finite & identical variance: $E[x_t^2] = \sigma^2$ for all t = 1,2,...
- Pairwise uncorrelated $E[x_s x_t] = 0$ for all $t \neq s$

1.2.1 White Noise Example 1: iid noise

$\{x_t\}$ is iid noise if x_t independent and identically distributed. So its joint distribution equals the product of the individual distribution. White noise is not necessarily iid noise, but iid noise is white noise. Because uncorrleated doesn't imply independent.

$$E[x_t] = 0$$
$$E[x_t^2] = \sigma^2$$
$$\gamma_x(t+h, t) = \begin{cases} \sigma^2 & h = 0 \\ 0 & o/w \end{cases}$$
$$\rho_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)} = \begin{cases} 1 & h = 0 \\ 0 & o/w \end{cases}$$

So this is a stationary process (actually true for all white noise process).

1.2.2 White Noise Example 2: Gaussian White noise

$\{x_t\}$ is Gaussian white noise if x_t iid noise with $x_t \sim N(0, \sigma^2)$

1.3 Basic Model: Random Walk

$$x_t = \sum_{i=1}^{t} w_i = x_{t-1} + w_t$$

- w_t is a 0-mean noise, e.g.: white noise
- $E[x_t] = 0$
- $var(x_t) = E[x_t^2] = E[\sum_{i=1}^{t} w_i \sum_{j=1}^{t} w_j] = \sum_{i=1}^{t} E[w_i^2] = t\sigma_w^2$ (Because all cross terms have $E[w_i * w_j] = 0$). So the variance grows with time.
- We can also add a drift term so $x_t = \delta_t + x_{t-1} + w_t$ to have non-zero expected value.
- $\gamma_x(t+h,t) = cov(x_{t+h}, x_t) = Cov(x_t + \sum_{i=1}^{h} w_{t+i}, x_t) = cov(x_t, x_t) = t\sigma^2$ Second last inequality is because w_i, w_{i+1} is independent.
- Not stationary because $\gamma_x(t+h,t) = t\sigma^2$ which is dependent on t, not just the distance between $t+h$ and t
- $\rho_x(t+h,t) = \frac{t\sigma^2}{\sqrt{(t\sigma^2)(t+h)\sigma^2}} = \frac{1}{1+\sqrt{h/t}}$

1.4 Basic Model: Autoregressive

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$$

- p is a number of lags considered in the model

Checking for stationary:

$$\begin{aligned}
\mu_{xt} &= E[x_t] = \phi E[x_{t-1}] + E[w_t] \\
&= \phi E[x_{t-1}] \\
&= 0 \qquad \text{If stationary} \\
\gamma_x(t+h,t) &= Cov(\phi x_{t+h-1} + w_{t+h}, x_t) \\
&= \phi Cov(x_{t+h-1}, x_t) \\
&= \phi \gamma_x(t+h-1,t) \\
&= \phi^{|h|}\gamma_x(t,t) \\
&= \phi^{|h|}E[x_t^2] \qquad \text{If stationary} \\
&= \phi^{|h|}(\phi^2 E[x_{t-1}^2] + \sigma^2) \\
&\text{With stationary assumption} \implies E[x_t^2] = E[x_{t-1}^2] \\
&\implies E[x_t^2] = E[x_{t-1}^2] = \frac{\sigma^2}{1-\phi^2} \\
&\implies \gamma_x(h) = \phi^{|h|} * \frac{\sigma^2}{1-\phi^2}
\end{aligned}$$

Thus for autocorrelation function, we have:

$$\rho_x(h) = \phi^{|h|}$$

1.5 Basic Model: Trend + seasonal

$$x_t = T_t + S_t + w_t$$

- T_t is the trend component. It could be a line $\beta_0 + \beta_1 t$
- S_t is the seasonal component. It could be $\sum_i (\beta_{i1} cos(\lambda_i t) + \beta_{i2} sin(\lambda_i t))$

Checking for Stationarity:

$$\mu_{xt} = E[x_t] = \beta_0 + \beta_1 t + \sum_i (\beta_{i1} \cos \lambda_i t + \beta_{i2} \sin \lambda_i t)$$

$$\gamma_x(t+h,t) = \begin{cases} \sigma^2 & h = 0 \\ 0 & o/w \end{cases}$$

So it is not stationary.

1.6 Basic Model: Moving Average Process

$$\begin{aligned}
x_t &= w_t + \theta w_{t-1} \\
w_t &\sim WN(0, \sigma^2)
\end{aligned}$$

We have:

$$\begin{aligned}
\mu_{xt} &= E[x_t] = 0 \\
\gamma_x(t+h,t) &= E[(w_{t+h} + \theta w_{t+h-1})(w_t + \theta w_{t-1})] \\
&= \begin{cases} Var(w_t + \theta w_{t-1}) & h = 0 \\ \sigma^2\theta & h = \pm 1 \\ 0 & o/w \end{cases} \\
&= \begin{cases} \sigma^2 + \theta^2\sigma^2 & h = 0 \\ \sigma^2\theta & h = \pm 1 \\ 0 & o/w \end{cases}
\end{aligned}$$

So the process is stationary. For autocorrelation function, we have:

$$\begin{aligned}
&h = 0 \Longrightarrow 1 \\
&h \pm 1 \Longrightarrow \frac{\theta}{1+\theta^2} \\
&o/w \Longrightarrow 0
\end{aligned}$$

1.7 Basic time series workflow

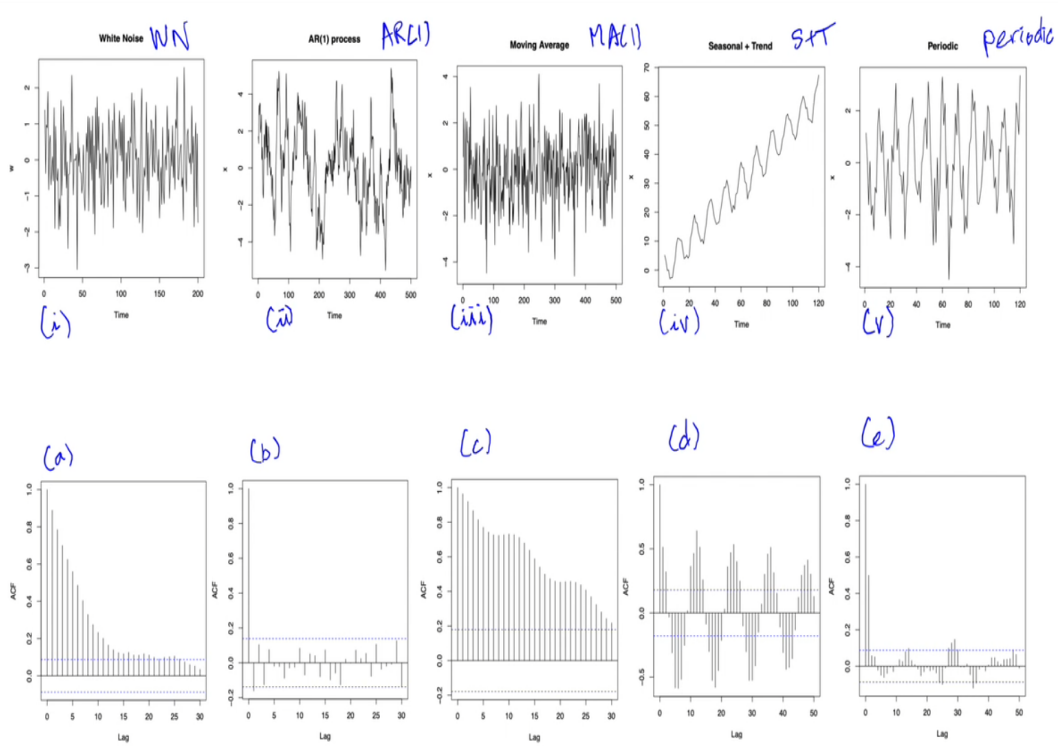
1. Plot time series
2. Transform data so residuals are stationary
   - Estimate and subtract trend and seasonal components
   - Differencing
   - Nonlinear transformations
3. Fit Model to residuals

Chapter 2

AutoCorrelation Function and Least Square Prediction

Intuition:
Sample autocorrelation function can inform some intuition about the structure of the time series. First, we can use sample ACF to test whether it is white noise process. Second, by examining the lags of high sample ACF, we can guess some potential structure. Looking at sample ACF at different lag is somewhat equivalent to have a series of scatterplot, e.g.: observation in t vs. t+1 ($\rho(h)$), t vs. t+2 ($\rho(h)$), and etc.

2.1 Matching Example



For White Noise process, everything should be uncorrelated except for $\rho(0)$, so the ACF should be plot b).
For AR(1) process, we have $\phi^{|h|}$, so it just decrease as h increases. Hence it matches to a).
For MA(1) process, we have the autocorrelation to be 0 outside of the MA order, so it matches to e).
For seasonal + trend, we know that the trend will induce correlation at long lags, then we have periodic struction, hence c).
Finally periodic will have only periodic ACF.

2.2 ACF and Least Square Prediction

2.2.1 Primer with Gaussian Process

Best LS estiamte of $x_{n+h}|x_n$:
Suppose $x = (x_1, ..., x_{n+h})$ is jointly Gaussian, then we have

$$f_x(x) = \frac{1}{(2\pi)^{\frac{n+h}{2}}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- Σ matrix is a matrix with variance on the diagonal, and $cov(x_i, x_j) = \rho(i,j) * \sigma_i * \sigma_j$ on other entries.

Given the jointly Gaussian, we have that the distribution of pair (x_n, x_{n+h}) is just a bivariate Gaussian distribution. i.e.:

$$\begin{pmatrix} x_n \\ x_{n+h} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_n \\ \mu_{n+h} \end{pmatrix}, \begin{pmatrix} \sigma_n^2 & \rho(n,n+h)\sigma_n\sigma_{n+h} \\ \rho(n,n+h)\sigma_n\sigma_{n+h} & \sigma_{n+h}^2 \end{pmatrix}\right)$$

The conditional distribution of $x_{n+h}|x_n$ is:

$$N(\mu_{n+h} + \rho(n,n+h)\frac{\sigma_{n+h}}{\sigma_n}(x_n - \mu_n), \sigma_{n+h}^2(1-\rho(n,n+h)^2))$$

Hence, if $\{x_t\}$ is Gaussian and stationary, the best estiamte of x_{n+h} given $x_n = c_n$ is the conditional mean, hence

$$\begin{aligned} &\mu_{n+h} + \rho(n,n+h)\frac{\sigma_{n+h}}{\sigma_n}(x_n - \mu_n) \\ &= \mu + \rho(h)(c_n - \mu_n) \end{aligned}$$

- $\rho(n, n+h) = \rho(h)$ due to stationarity
- $\frac{\sigma_{n+h}}{\sigma_n} = 1$ due to variance being time insensitive.

and the MSE is

$$\begin{aligned} E[(x_{n+h} - f(x_n))^2] &= E[(x_{n+h} - E[x_{n+h}|x_n])^2|x_n] \\ &= Var(x_{n+h}|x_n) \end{aligned}$$

So notice that

- predictive accuracy improves as $|\rho(h)| \to 1$
- prediction is linear in x

2.2.2 General Linear Prediction

Goal: Predicting $x_{n+h}|x_n = c_n$ with $f(x_n) = a(x_n - \mu) + b$. Assume $\{x_t\}$ stationary with $E[x_t] = \mu$ and $var(x_t) = \sigma^2$. The best linear predictor has the form

$$f(x_n) = \rho(h)(x_n - \mu) + \mu$$

and the MSE is

$$\sigma^2(1 - \rho(h)^2)$$

Chapter 3

Data Wrangling

Motivation
The goal is to

- Remove trend and periodic components
- Ohter transformations of data
- Fit model (for stationary time series) to resulting data

3.1 Parametric Detrending and data transformations

3.1.1 Parametric Trend Model

The trend model is

$$x_t = y_t + T_t$$

- y_t is the stationary process
- T_t is a deterministic trend parameterized by B.
- The detrended series will be defined as $\hat{y}_t = x_t - f(t; \hat{B})$
- Some examples of trend functions are
  - Polynomial regression $f(t; B) = \beta_0 + \beta_1 t + \beta_2 t^2 + ...$
  - Periodic regression over known period T: $f(t; A) = \alpha_1 cos(\frac{2\pi t}{T}) + \alpha_2 sin(\frac{2\pi t}{T})$
  - Hybrid: polynomial + periodic
- It is easy to fit and predict, but may be unrealistic

3.1.2 Differencing

1st order

$$\begin{aligned}\nabla x_t &= x_t - x_{t-1} = (1-B)x_t && \text{(B is a backshift operator)}\\ &= (y_t - y_{t-1}) + (T_t - T_{t-1}) && \text{(Hope } T_t - T_{t-1} \text{ is constant)}\\ &\text{If } y_t \text{ is stationary, then } y_t - y_{t-1} \text{ is also stationary}\end{aligned}$$

Order d

$$\nabla^d x_t = (1-B)^d x_t$$

Example:

$$\begin{aligned}&x_t = t^2 + y_t\\ &\nabla^2 x_t\\ &= \nabla(\nabla x_t)\\ &= \nabla(x_t - x_{t-1})\\ &= \nabla x_t - \nabla x_{t-1}\\ &= (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})\\ &= (t^2 + y_t - (t-1)^2 - y_{t-1}) - (x_{t-1} - x_{t-2})\\ &= [(2t-1) + (y_t - y_{t-1})] - [2(t-1) - 1 + (y_{t-1} - y_{t-2})]\\ &= 2 + y_t - 2y_{t-1} + y_{t-2}\\ &\quad \text{(Constant + filtering of } y_t \text{ (approximation of second derivative))}\end{aligned}$$

3.1.3 Log and Power Transformation

Log

$$y_t = log x_t$$

- Applies to non-negative data
- tends to suppress large fluctuations occuring over portions of the series
- Improves approximation to normality (including in transforming large count data to be approximately normal)
- Improves linearity in predicting one series from another

Box Cox Power Transformation

$$y_t = \begin{cases}(x_t^\lambda - 1)/\lambda & \lambda > 0\\ \log(x_t) & \lambda = 0\end{cases}$$

3.2 Nonparametric Trend Estimation: Smoothing

General goals of nonparametrics:

- Flexibility
- Make few assumptions about $f(x)$
- Complexity can grow with the number of observation

3.2.1 k-NN Regression

Given dataset of pairs $(t_1, x_1), ..., (t_n, x_n)$ and query point t_q. We

1. Order data by distance from t_q. $(t_{NN1}, ..., t_{NNK})$ such that for any t_i not in the nearest neighbor (NN) set, $|t_i - t_q| \geq |t_{NNK} - t_q|$
2. Estimate $\hat{f}(t_q) = \frac{1}{K} \sum_{j=1}^{K} x_{NNj}$

KNN have discontinuity, and also boundary issue. We can use weighted k-NN

$$\hat{f}(t_q) = \frac{C_{qNN1} X_{NN1} + C_{qNN2} X_{NN2} + ... + C_{qNNk} X_{NNk}}{\sum_{j=1}^{k} C_{qNNj}}$$

The weight can be decided using Kernel, e.g. a Gaussian kernel :

- $C_{qNNj} = Kernel_\lambda(|t_{NNj} - t_q|) = exp(-(t_{NNj} - t_q)^2/\lambda)$

3.2.2 Kernel Regression

K-NN has a fixed set of K nearest neighbors. We can weight all points and use Kernel weights. (Nadaraya-Watson Kernel Weighted Average). Notice here λ is the distance away from t_q where we apply the kernel. So if a point t_i is more than λ away from t_q, the Kernel will be 0.

$$\hat{f}(t_q) = \frac{\sum_{i=1}^{n} Kernel_\lambda(distance(t_i, t_q)) * x_i}{\sum_{i=1}^{n} Kernel_\lambda(distance(t_i, t_q))}$$

Generally, choice of kernel matters much less than choice of λ. We can pick λ through cross validation.

3.2.3 Local Linear Regression

Notice the kernel regression is a solution to

$$\hat{f}(t_q) = \overset{argmin}{a} \sum_{i=1}^{n} K_\lambda(|t_i - t_q|)(x_i - a)^2$$

So we are fitting constant function locally at each point (finding an average / weighted average locally). What if we fit a line or polynomial locally at each point?

$$\beta_{0q} + \beta_{1q}(t - t_q) \qquad \text{(Centered at } t_q\text{)}$$

So our objective is to at each point, we fit a local regrression

$$\overset{argmin}{\beta} \sum_i K_\lambda(|t_i - t_q|)(x_i - (\beta_{0q} + \beta_{1q}(t - t_q)))^2$$

$$\hat{f}(t_q) = \hat{\beta}_{0q} \qquad \text{(Fit at } t_q \text{ is just intercept)}$$

Local regression rules of thumb:

- Local linear fit reduces bias at boundaries with minimum increase in variance
- Local quadratic fit doesn't help at boundaries and increase variance, but does help capture curvature in the interior
- with sufficient data, local polynomials of odd degree dominate those of even degree

3.2.4 Splines

Consider generic function forms

$$\overset{min}{f} ||x - f(t)||^2$$

If we constrained to linear forms, we have least square solution. But if f can be arbitrary, we have interpolator function (just set f to be those points). So we can introduce a term to penalize complexity. hence

$$\overset{min}{f} ||x - f(t)||^2 + \lambda \int f''(t)^2 dt$$

The result is a natural cubic spline with knots at data points (smoothing splines). Cublic spline with two knots example below are:

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + b_1(t - \xi_1)_+^3 + b_2(t - \xi_2)_+^3$$

Overall, an order-M spline with knots $\xi_1 < ... < \xi_K$ is a piecewise $M - 1$ degree polynomial with $M - 2$ continuous derivative at the knots. A spline that is linear beyond the boundary knots is called a natural spline. The choices are

Chapter 4

AR, MA, and ARMA Processes

4.1 Linear Processes, causality and invertibility

4.1.1 Linear Processes

A linear process is one that can be written as

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, w_t \sim WN(0, \sigma_w^2)$$
$$\text{where } \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

4.1.2 Backshift Operator

The backshift operator is defined as $Bx_t = x_{t-1}$.

4.1.3 Causality

A linear process $\{x_t\}$ is causal (formally a causal function of $\{w_t\}$) if

$$x_t = \psi(B)w_t$$
$$\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + ...,$$
$$\sum_{j=0}^{\infty} |\psi_j| \leq 0$$

i.e.: we can represent x_t as a linear combination of past w_i

4.1.4 Invertibility

A linear process $\{x_t\}$ is invertible iff

$$w_t = \pi(B)x_t$$
$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + ...$$
$$\sum_{j=0}^{\infty} |\pi_j| < \infty$$

i.e.: we can represent w_t as a linear combination of past x_i

4.2 Autoregressive Process

4.2.1 Definition

An order-p autoregressive process, $AR(p)$ is defined as

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t$$
$$w_t \sim WN(0, \sigma_w^2)$$

Alternative, $AR(p)$ process can be defined with a mean

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + ... + \phi_p(x_{t-p} - \mu)) + w_t$$

or

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t$$
$$\alpha = \mu(1 - \phi_1 - ...\phi_p)$$

4.2.2 AR(1) Processes

$$x_t = \phi x_{t-1} + w_t$$

What if |phi| < 1

$$x_t = \phi x_{t-1} + w_t$$
$$= \phi^k x_{t-k} + \sum_{j=1}^{k-1} \phi^j w_{t-j}$$

If $|\phi| < 1$ and $\overset{sup}{t}\, var(x_t) < \infty$

as $k \to \infty$, we have:

$$\implies \phi^k x_{t-k} \to 0, \qquad \sum_{j=0}^{\infty} \phi^j w_{t-j} \text{converges}$$

$$\implies x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$$

So this is a stationary, linear process. It is equivalent to $MA(\infty)$

$$
\begin{aligned}
E[x_t] &= \sum_{j=0}^{\infty} \phi^j E[w_{t-j}] = 0 \\
\gamma(h) &= cov(x_{t+h}, x_t) \\
&= E\left[\sum_{j=0}^{\infty} \phi^j w_{t+h-j} * \sum_{k=0}^{\infty} \phi^k w_{t-k}\right] \\
&= E\left[(w_{t+h} + \phi^1 w_{t+h-1} + ... + \phi^h w_t + \phi^{h+1} w_{t-1} + ...)(w_t + \phi^1 w_{t-1} + ...)\right] \\
&= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j}\phi^j \qquad (E[w_i w_j] = 0 \text{ if } i \neq j) \\
&= \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} \\
&= \sigma_2^2 \frac{\phi^h}{1-\phi^2} \\
\rho(h) &= \frac{\gamma(h)}{\gamma(0)} = \phi^h = \phi * \rho(h-1)
\end{aligned}
$$

So overall, we say an AR(1) process is causal iff $|\phi| < 1$

What if phi = 1
We have a random walk process, which is not stationary.

What if |phi| > 1
We have a non-stationary exploding variance AR process. But there is a stationary solution with $|\phi| > 1$. I.e.: is there a stationary sequence with recursion like $x_t = \phi x_{t-1} + w_t$

$$
\begin{aligned}
& x_{t+1} = \phi x_t + w_{t+1} \\
& \implies w_t = \phi^{-1} x_{t+1} - \phi^{-1} w_{t+1} \\
& = \phi^{-k} x_{t+k} - \sum_{j=1}^{k-1} \phi^{-j} w_{t+j} \\
& \text{If } |\phi|^{-1} < 1 \\
& \implies x_t = -\sum_{j=1}^{\infty} \phi^{-j} w_{t+j}
\end{aligned}
$$

So this is a linear process, and it is stationary. This is not causal. (past depended on the future)

4.2.3 Leveraging the backshift Operator and Manipulation

Using this, we can rewrite AR(1) processes as

$$\phi(B)x_t = w_t \qquad\qquad \phi(B) = 1 - \phi * B$$

Let $|\phi| < 1$, we have

$$\begin{aligned} x_t &= \sum_{j=0}^{\infty} \phi^j w_{t-j} \\ &= \sum_{j=0}^{\infty} \phi^j B^j * w_t \\ &= \pi(B) w_t \end{aligned}$$

We can show that $\pi(B) = \phi(B)^{-1}$

$$\begin{aligned} \pi(B) * \phi(B) &= \sum_{j=0}^{\infty} \phi^j B^j * (1 - \phi * B) \\ &= \sum_{j=0}^{\infty} \phi^j B^j - \sum_{j=1}^{\infty} \phi^j B^j \\ &= \phi^0 B^0 = 1 \end{aligned}$$

4.2.4 Causal Processes and backshift Operator

We can say AR(1) process is causal iff of the following two equivalent statement

- $|\phi| < 1$
- $\phi(z) = 1 - \phi z$ satisfies root $|z_1| > 1$

An AR(1) process is stationary iff of the following two equivalent statement

- $|\phi| \neq 1$
- $\phi(z) = 1 - \phi z$ satisfies root $|z_1| \neq 1$

4.2.5 AR(p) Process Summary

Definition
AR(p) process can be defined as :

$$\begin{aligned} &x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t \\ &(1 - \phi_1 B - \phi_2 B^2 - ... - \phi^p B^p) x_t = w_t \\ &\phi(B) x_t = w_t \end{aligned}$$

Its characteristic polynomial is $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^P$

Stationarity and causality for AR(p):
A unique stationary solution to

$$\phi(B)x_t = w_t$$

exists iff

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^P = 0 \Longrightarrow |z| \neq 1$$

This AR(p) is causal iff

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^P = 0 \Longrightarrow |z| > 1$$

4.3 Moving Average Processes

4.3.1 Definition

An order-q moving average process, MA(q) is defined as

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}$$

Equivalently, in backshift operator form

$$x_t = \theta(B)w_t \qquad \theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$$

This process is a stationary process for all choices of θ. The characteristic polynomial is $\theta(z) = 1 + \theta_1 z + ... + \theta_p z^p$. We normally pick the representation that is invertible.

4.3.2 MA(1)

$$E[x_t] = 0$$

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0 \\ \theta\sigma_w^2 & h = 1 \\ 0 & h > 1 \end{cases}$$

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta}{1+\theta^2} & h = 1 \\ 0 & h > 1 \end{cases}$$

It is not unique. $\tilde{x_t} = \tilde{w_t} + \frac{1}{\theta}\tilde{w}_{t-1}, \tilde{w_t} \sim WN(0, \theta^2\sigma_w^2)$ has the exact same second order statistics.

Invertibility
The backshift operator form of MA(1) is

$$x_t = w_t + \theta w_{t-1} = (1 + \theta B) w_t$$

If $|\theta| < 1$, we can write it as

$$\begin{aligned} &(1 + \theta B)^{-1} x_t = w_t \\ &(1 - \theta B + \theta^2 B - \theta^3 B^3 + ...) x_t = w_t \qquad \text{(geometric series)} \\ &\sum_{j=0}^{\infty} (-\theta)^j x_{t-j} = w_t \end{aligned}$$

Note we can write w_t as a causal function of x_t, we call this invertible. In addition, notice that the third line is an infinity order AR process.
If $|\theta| > 1$, then $sum_{j=0}^{\infty}(-\theta)^j x_{t-j}$ diverges. But we can write it as

$$w_{t-1} = -\theta^{-1} w_t + \theta^{-1} x_x t \qquad = \sum_{j=0}^{\infty} (-\theta)^j x_{t+j}$$

That is we can write w_t as linear, but non-causal function of x_t. so now MA(1) is not invertible.
In summary, an MA(1) process is invertible

- iff $|\theta| < 1$
- iff $\theta(z) = 1 + \theta z$ has root outside of unit circle.

4.3.3 Summary of MA(q) Process

An MA(q) process is invertible iff the root of the characteristics polynomial is outside the unit circle.

4.4 ARMA

An ARMA(p,q) process $\{x_t\}$ is a stationary process that satisfies

$$\begin{aligned} &x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \\ &w_t \sim WN(0, \sigma_w^2) \end{aligned}$$

In backshift form, we have

$$\phi(B) x_t = \theta(B) w_t$$

4.4.1 Stationary AR models that are not causal

An ARMA(p,q) process $\{x_t\}$ is causal if it can be written as a one-sided linear process / past white noise terms

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t_j} = \psi(B) w_t$$

where

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$$

with

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \qquad \psi_0 = 1$$

An ARMA process is causal iff the AR characteristics polynomial

$$\phi(z) \neq 0 \forall |z| \leq 1$$

That is, $\phi(z)$ has roots only outside the unit circle.
In the case of a causal ARMA(p,q) process, we can write

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$$

4.4.2 Non-unique MA models / Invertibility

An ARMA(p,q) process $\{x_t\}$ is invertible if it can be written as

$$\pi(B) x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t$$

Where

$$\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$$

with

$$\sum_{j=0}^{\infty} |\pi_j| < \infty \qquad \pi_0 = 1$$

An ARMA process is invertible iff MA char polynomial

$$\theta(z) \neq 0 \forall |z| \leq 1$$

That is, $\theta(z)$ has roots only outside the unit circle.
In the case of an invertible ARMA(p,q) process, we can write

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, |z| \leq 1$$

4.4.3 Parameter redundancy

Assuming $\phi_p \neq 0$ and $\theta_p \neq 0$

$$\begin{aligned} &AR : \phi(z) = 1 - \phi_1 z - ... - \phi_p z^P \\ &MA : \theta(z) = 1 + \theta_1 + ... + \theta_p z^q \\ &\psi(z) = \frac{\theta(z)}{\phi(z)} \\ &\pi(z) = \frac{\phi(z)}{\theta(z)} \end{aligned}$$

If AR and MA polynomials $\phi(z), \theta(z)$ have common factors, they cancel out when solving for the ARMA causal or invertible operator. That is, there exists a reduced order, yet stochastically equivalent ARMA process.

4.4.4 Example

$$\begin{aligned} &x_t = 1.5x_{t-1} + w_t + 0.2w_{t-1} \\ &\implies \phi(B) = 1 - 1.5z && z = \frac{2}{3} \\ &\theta(z) = 1 + 0.2z && z = -5 \end{aligned}$$

$\frac{2}{3} < 1$ so the process is not causal. $|-5| > 1$ so the process is invertible.

4.4.5 Solving for MA infinite representation and AR infinite representation

How to solve $\phi(z)\psi(z) = \theta(z)$ or $\theta(z)\pi(z) = \phi(z)$.

In general, we have

$$\begin{aligned} &\phi(B)x_t = \theta(B)w_t x_t = \psi(B)w_t \\ &\implies \theta_j = \phi(B)\psi_j \end{aligned}$$

with $\theta_0 = 1$, $\theta_j = 0$ for $j < 0, j > q$

4.4.5.1 Example

$$x_t = 0.4x_{t-1} + 0.45x_{t-2} + w_t + w_{t-1} + 0.25w_{t-2}$$
$$\phi(z) = 1 - 0.4z - 0.45z^2 = (1 + 0.5z)(1 - 0.9z)$$
$$\theta(z) = 1 + z + 0.25z^2 = (1 + 0.5z)^2$$

- ϕ and θ have common factors, so we need to take out the common factor. So it is a ARMA(1,1) process
- the root for ϕ is $\frac{10}{9}$, so it is causal (outside the unit circle)
- the root for θ is -2, so it is invertible.

Recall the ARMA form is

$$\phi(B)x_t = \theta(B)w_t$$

Solving for MA infinity form (causal polynomial)
If ARMA is causal, it means we can write $x_t = \psi(B)w_t$. So we have $\psi(z) = \frac{\theta(z)}{\phi(z)}$. Alternatively, we have

$$\phi(z) * \psi(z) = \theta(z)$$
$$\implies (1 - 0.9z)(1 + \psi_1 z + \psi_2 z^2 + ...) = (1 + 0.5z)$$

Once we multiply out the left term, we can do coefficient matching and have

$$\begin{cases} \psi_1 - 0.9 = 0.5 \\ \psi_j - 0/9\psi_{j-1} = 0 \quad j > 1 \end{cases}$$

So overall, we have

$$x_t = w_t + 1.4\sum_{j=1}^{\infty} 0.9^{j-1} w_{t-j}$$

4.5 ACF and partial ACF

4.5.1 Autocovariance of an ARMA process

Causal linear process representation

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$$

Autocovariance is

$$\gamma(h) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} h \geq 0$$

Special case: MA(q) (Because moving average is already in casual form)

$$\gamma(h) = \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} h \geq 0$$

4.5.2 Solving linear difference equation directly for autocov fcn

The ARMA form is

$x_t - \phi_1 x_{t-1} - ... - \phi_p x_{t-p} = w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$
$\implies E[(x_t - \phi_1 x_{t-1} - ... - \phi_p x_{t-p}) * x_{t-h}] = E[(w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}) * x_{t-h}]$
$\implies \gamma(h) - \phi_1 \gamma(h-1) - ... - \phi_p \gamma(h-p) = \sigma_w^2 \sum_{j=0}^{q-h} \theta_{h+j} \psi_j$

Some more detail on the right hand side:

Recall $w_t, ..., w_{t-h+1}$ is uncorrelated with x_{t-h}
$\implies E[(w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}) * x_{t-h}] = E[(\theta_h w_{t-h} + ...\theta_q w_{t-q}) * x_{t-h}]$
Examine a single term $E[\theta_{h+i} * w_{t-h-i} * x_{t-h}]$
$= \theta_{h+i} E[w_{t-h-i} \sum_{j=0}^{\infty} \psi_j w_{t-h-j}]$
The only non zero product is when $j = i$
$\implies = \theta_{h+i} \psi_i E[w_{t-h-i}^2]$
$= \theta_{h+i} \psi_i \sigma_w^2$

4.5.3 Autocorrelation Function (ACF) of an ARMA process

As always, $\rho(h) = \gamma(h)/\gamma(0)$. From previous result, we have $\gamma(h) - \phi_1 \gamma(h-1) - ... - \phi_p \gamma(h-p) = \sigma_w^2 \sum_{j=0}^{q-h} \theta_{h+j} \psi_j$

4.5.3.1 ACF for AR(p)

For ARMA(p,0) process, there is no moving average component. So

$\gamma(h) - \phi_1\gamma(h-1) - ... - \phi_p\gamma(h-p) = 0$
$\implies \rho(h) - \phi_1\rho(h-1) - ... - \phi_p\rho(h-p) = 0$
$\implies \rho(h) = z_1^{-h}P_1(h) + z_2^{-h}P_2(h) + ... + z_r^{-h}P_r(h)$
$r =$ number of distinct roots, each z_i is a root of the characteristics polynomial
$P_j(h)$ is a polynomial of degree $m_j - 1$ where m_j is multiplicity of root z_j, constant if distinct root

So

- If z_j are all real, $\rho(h)$ dampens exponentially fast to 0 as $h \to \infty$
- if some z_j are complex (in complex conjugate pairs), $\rho(h)$ dampens exponentially fast in sinusoidal manner.

4.5.4 Partial ACF

ACF can help with determine the q for $MA(q)$ process because for any $h > q$, it goes to 0. But ACF cannot determine order of ARMA(p,q) or AR(p)

Partial correlation

$$ParCorr(x, y|z_1, ..., z_j) = Corr(x - \hat{x}(z_1, ..., z_h), y - \hat{y}(z_1, ..., z_h))$$

where $\hat{x}, \hat{y}$ is regression of x or y on $z_1, ..., z_h$

Partial Autocorrelation Function
PartialACF of a stationary process $\{x_t\}$ is

$$ParCorr(x_{t+h}, x_t|x_{t+1}, ..., x_{t+h-1}$$

In summary, we have

	AR(p)	MA(q)	ARMA(p,q)
ACF	decays	cutoff q	decays
PACF	cutoff p	decays	decays

Chapter 5

Forecasting

5.1 Primer: Least squares estimation

Best LS estimate of $y|x$ is $E[Y|X]$

$$\begin{aligned} \overset{min}{f} E_{x,y}[(y - f(x))^2] &= \overset{min}{f} E_x[E_{y|x}[[(y - f(x))^2|x]] \\ &= \overset{min}{f} E_x[E_{y|x}[[(y - E[Y|X])^2|x]] \\ &= var(Y|X) \end{aligned}$$

So the best LS estimate of $x_{n+h}|x_n$ is

$$E[x_{n+h}|x_n]$$

If we constraint the LS predictor to be linear in the form of $f(x_n) = a(x_n - \mu) + b$. Assume $\{x_t\}$ is stationary with $E[x_t] = \mu$. Then the best linear predictor has the form

$$f(x_n) = \rho(h)(x_n - \mu) + \mu$$

and with MSE

$$\sigma^2(1 - \phi(h)^2)$$

And if the process is also Gaussian, f is the optimal overall predictor.

5.2 Projection Theorem

Let $\mathcal{H}$ is a Hilbert space (complete inner product space). For random variable, $< x, y >= E[xy]$ Let $\mathcal{M}$ be a closed linear subspace of $\mathcal{H}$. So for $y \in \mathcal{H}$, there is a point $P_y \in \mathcal{M}$ such that

$$\begin{aligned} &\|P_y - y\| \leq \|w - y\| \text{ for } w \in \mathcal{M} \\ &\|P_y - y\| < \|w - y\| \text{ for } w \in \mathcal{M}, w \neq y \\ &< y - P_y, w >= 0 \text{ for } w \in M \end{aligned}$$

In translation, given $y \in \mathcal{H}$, P_y is a point (project of P_y onto M) such that the distance between y and P_y is smaller than y to all other points in M.

5.3 The m-step-ahead predictor Definition

The m-step-ahead predictor ($n + m$ is the goal, and n is the length of the sequence given) is defined as

$$x_{n+m}^n = E[x_{n+m}|x_1, ..., x_n]$$

and minimizes MSE (here x_{n+m}^n is the predictor)

$$P_{n+m}^n = E[(x_{n+m} - x_{n+m}^n)^2]$$

We want to constrain to linear form

$$x_{n+m}^n = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i$$

Note we have $a_0 = 0$ if we assume stationary process with mean $E[x_i] = \mu$, we can get $a_0 = \mu$. So normally, we can just assume $E[x_i] = 0$ for simplicity.

5.4 Prediction Equation and Projection Theorem

Given history $x_1, ..., x_n$, the best linear predictor satisfies the prediction equations

$$E[x_{n+m} - x_{n+m}^n] = 0 \qquad \text{(Unbiased predictor)}$$
$$E[(x_{n+m} - x_{n+m}^n)x_i] = 0 \qquad \forall i = 1, ..., n$$
(Error uncorrelated to observed value (projection theorem))

We can think of H as the space of random variables with inner product defined as $E[xy]$, and the linear space M is span of $\{1, x_1, ..., x_n\}$ and $y = x_{n+m}$

5.5 1-step-ahead linear prediction

Form of forecast:

$$x_{n+1}^n = \phi_{n+1,1}^n x_n + ... + \phi_{n+1,n}^n x_1$$

The prediction equations gives us

$$
\begin{aligned}
&E[(x^n_{n+1} - x_{n+1})x_{n+1-i}] = 0, i = 1, ..., n \\
&\implies E[(\sum_{j=1}^{n} \phi^n_{n+1,j} x_{n+1-j} - x_{n+1}) * x_{n+1-i}] = 0 \\
&\implies \sum_{j=1}^{n} \phi^n_{n+1,j} E[x_{n+1-j} x_{n+1-i}] = E[x_{n+1} x_{n+1-i}] \\
&\implies \sum_{j=1}^{n} \phi^n_{n+1,j} \gamma(i-j) = \gamma(i) \\
&\implies \Gamma_n \tilde{\phi}^n_{n+1} = \tilde{\gamma}(n)
\end{aligned}
$$

$$
\Gamma_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix} \tilde{\phi}^n_{n+1} = \begin{bmatrix} \phi^n_{n+1,1} \\ \phi^n_{n+1,2} \\ \vdots \\ \phi^n_{n+1,n} \end{bmatrix} \tilde{\gamma}(n) = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(n) \end{bmatrix}
$$

Note so getting best linear predictor still uses second order statistics

MSE of 1-step-ahead linear prediction
So our MSE is

$$
\begin{aligned}
P^n_{n+1} &= E[(x_{n+1} - x^n_{n+1})^2] \\
&= E[(x_{n+1}(x_{n+1} - x^n_{n+1}) - x^n_{n+1}(x_{n+1} - x^n_{n+1})] \\
&= E[(x_{n+1}(x_{n+1} - x^n_{n+1})]
\end{aligned}
$$

(By projection theorem, x^n_{n+1} in linear space M, and $x_{n+1} - x^n_{n+1}$ is the error)

$$
\begin{aligned}
&= \gamma(0) - E[x_{n+1} \tilde{\phi^n_{n+1}}^T x] \qquad (x \text{ is a vector of } x_i) \\
&= \gamma(0) - E[x_{n+1} x^T] \Gamma_n^{-1} \tilde{\gamma}(n) \\
&= \gamma(0) - \tilde{\gamma}(n)^T \Gamma_n^{-1} \tilde{\gamma}(n) \\
&= Var(x_{n+1}) - Cov(x_{n+1}, x) Cov(x, x)^{-1} Cov(x, x_{n+1})
\end{aligned}
$$

So variance is reduced by conditioning on the history.

5.6 The m-step-ahead predictor

Form (using $x_1, ..., x_n$ to predict x_{n+m}):

$$
\begin{aligned}
x^n_{n+m} &= \phi^n_{n+m,1} x_n + \phi^n_{n+m,2} x_{n-1} + ... + \phi^n_{n+m,n} x_1 \\
&= \sum_{i=1}^{n} \phi^n_{n+m,i} x_{n+1-i}
\end{aligned}
$$

The prediction equations are

$$E\left[(x_{n+m}^n - x_{n+m}) * (x_{n+1-i})\right] = 0, i = 1, ..., n$$

The simplified matrix form is

$$\tilde{\phi}_{n+m}^n = \Gamma_n^{-1}\tilde{\gamma}_n^m$$
$$\tilde{\gamma}_n^m = \begin{bmatrix} \gamma(m) \\ \vdots \\ \gamma(m+n-1) \end{bmatrix}$$

And the MSE is

$$P_{n+m}^n = E\left[(x_{n+m} - x_{n+m}^n)^2\right] = \gamma(0) - (\tilde{\gamma}_n^{(m)})^T\Gamma_n^{-1}\gamma_n^{(m)}$$

Computation Alternative
How to compute $\tilde{\phi}_{n+m}^n$ without solving the linear system

- Durbin-Levinson recursive algorithm for AR(p)
- Innovations algorithm for MA(q)

Prediction Intervals

$$x_{n+m}^n \pm c_{\alpha/2}\sqrt{P_{n+m}^n}$$

For Gaussian process, the prediction error has $\mathcal{N}(0, P_{n+m}^n)$. But we need to adjust for multiple testing to get PI for more than one time period, such as using Bonferroni.

5.7 Backcasting

Given $x_1, ..., x_n$, predict $x_{1-m}, m > 0$

1-step linear backcasting
Form of backcast

$$x_0^n = \phi_{0,1}^n x_1 + \phi_{0,2}^n x_2 + ... + \phi_{0,n}^n x_n = (\tilde{\phi}_0^n)^T x$$
(Note now x is flipped from $x_n, ..., x_1$ to $x_1, ..., x_n$)

Prediction equations

$$E[(x_0^n - x_0)x_i] = 0, i = 1, ..., n \Longrightarrow \Gamma_n\tilde{\phi}_0^n = \tilde{\gamma}(n)$$

So it is the same, except now the coefficients are lined up from x_1 to x_n instead of the reveres order.

5.8 Forecasting ARMA

5.8.1 Recursive Solution

Linear prediction based on infinite past:
Assume infinite past $x_n, x_{n-1}, ...,$

$$\tilde{x}_{n+m} = E[x_{n+m}|x_n, x_{n-1}, ...] = \sum_{i=1}^{\infty} \alpha_i x_{n+1-i}$$

Orhtogonality property of optimal linear predictor implies

$$E[(\tilde{x}_{n+m} - x_{n+m})x_i] = 0, i = 1, 2, 3, ...$$

So if $\{x_t\}$ is a 0 mean stationary process, we have

$$\sum \alpha_i \gamma(i-j) = \gamma(m-1+i), i = 1, 2, 3, ...$$

If $\{x_t\}$ is causal, invertible, linear process, we can write

$$x_{n+m} = \sum_{j=1}^{\infty} \psi_j w_{n+m-j} + w_{n+m} \qquad w_{n+m} = \sum_{j=1}^{\infty} \pi_j x_{n+m-j} + x_{n+m}$$

Taking expectation of the second equation, we have

$$\begin{aligned} E[x_{n+m}|x_n, ...] &= E[w_{n+m}|x_n, ...] - \sum_{j=1}^{\infty} \pi_j x_{n+m-j} \\ &= -\sum_{j=1}^{\infty} \pi_j x_{n+m-j} \\ &= -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j} \end{aligned}$$

(Notice the second part contains observed value)

So we can iteratively predict:

$$\tilde{x}_{n+1} = -\sum_{j=1}^{\infty} \pi_j x_{n+1-j}$$

$$\tilde{x}_{n+2} = -\sum_{j=1}^{1} \pi_j \tilde{x}_{n+2-j} - \sum_{j=2}^{\infty} \pi_j x_{n+2-j} = -\pi_1 \tilde{x}_{n+1} - \sum_{j=2}^{\infty} \pi_j x_{n+2-j}$$

$$\tilde{x}_{n+3} = -\sum_{j=1}^{2} \pi_1 \tilde{x}_{n+3-j} - \sum_{j=2}^{\infty} \pi_j x_{n+3-j} = -\pi_1 \tilde{x}_{n+2} - \pi_2 \tilde{x}_{n+1} - \sum_{j=3}^{\infty} \pi_j x_{n+3-j}$$

5.8.2 MSE

$$
\begin{aligned}
x_{n+m} &= \sum_{j=1}^{\infty} \psi_j w_{n+m-j} + w_{n+m} \\
E[\tilde{x}_{n+m}|x_n, ...] &= \sum_{j=1}^{\infty} \psi_j E[w_{n+m-j}|x_n, ...] + E[w_{n+m}|x_n, ...] \\
&= \sum_{j=1}^{\infty} \psi_j E[w_{n+m-j}|x_n, ...] \\
&= \sum_{j=m}^{\infty} \psi_j w_{n+m-j} \\
MSE &= E[(x_{n+m} - E[\tilde{x}_{n+m}|x_n, ...])^2] \\
&= E[(x_{n+m} - \sum_{j=m}^{\infty} \psi_j w_{n+m-j})^2] \\
&= E[(\sum_{j=0}^{\infty} \psi_j w_{n+m-j} - \sum_{j=m}^{\infty} \psi_j w_{n+m-j})^2] \\
&= E[(\sum_{j=0}^{m-1} \psi_j w_{n+m-j})^2] \\
&= \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2
\end{aligned}
$$

So for $m = 1$, we have the MSE being the variance σ_w^2.
As $m \to \infty$, we have

$$
\begin{aligned}
&\tilde{x}_{n+m} = 0 && \text{(The mean)} \\
&P_{n+m}^n \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma_x^2
\end{aligned}
$$

So for far in the future, we just predict the mean (history doesn't matter), and the MSE is the variance of the process.

5.8.3 Truncated Forecasts

For large n, truncating infinite-past forecasts gives a good approximations:

$$\tilde{x}_{n+m} = \sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}$$
$$\tilde{x}^n_{n+m} = \sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}$$

The approximation is exact for AR(p) when $n \geq p$ because $\pi_j = 0$ for $j > p$. In general, a good approximation if π_j converge quickly to 0.

5.8.4 Example: Forecasting ARMA(p,q) model Summary

$$x_t - \sum_{i=1}^{p} \phi_i x_{t-i} = w_t + \sum_{i=1}^{q} \theta_i w_{t-i}$$

We want to forecast x_{n+m} from $x_1, ..., x_n$. The exact predictor is trivial for AR(p) process using ϕ_i. Otherwise, we can compute truncated forecast recursively in time $O((n+m)(p+q))$. The algorithm is
For $t = n+1, ..., n+m$, calculate:

$$\tilde{x}^n_t = \phi_1 \tilde{x}^n_{t-1} + ... + \phi_p \tilde{x}^n_{t-p} + \theta_1 \tilde{w}^n_{t-1} + ...\theta_q \tilde{w}^n_{t-q}$$

For $1 \leq t \leq n$:

$$\tilde{w}^n_t = \phi(B)\tilde{x}^n_t - \theta_1 \tilde{w}^n_{t-1} - ... - \theta_q \tilde{w}^n_{t-q}$$

Boundary condition:

$$\tilde{x}^n_t = x_t, 1 \leq t \leq n$$
$$\tilde{x}^n_t = 0, t \leq 0$$
$$\tilde{w}^n_t = 0, t \leq 0, t > n$$

5.8.5 Forecasting Metrics

Mean absolute percent error (MAPE)

$$\frac{1}{m} \sum_{j=1}^{m} \frac{|x^n_{n+j} - x_{n+j}|}{|x_{n+j}|} * 100\%$$

- have issue if x_{n+j} is 0

Weighted Average Percentage Error(WAPE)

$$\frac{\frac{1}{m}\sum_{j=1}^{m} |x^n_{n+j} - x_{n+j}|}{\frac{1}{m}\sum_{j=1}^{m} |x_{n+j}|}$$

Chapter 6

Estimation

6.1 Estimating ARMA(p,q) parameters given p and q

6.1.1 Method of Moments

Choose parameters for which the moments equal the empirical moments.

6.1.1.1 AR(p) and Yule-Walker estimation

$$\gamma(h) = \phi_1\gamma(h-1) + ... + \phi_p\gamma(h-p), h = 1, ..., p$$
$$\sigma_w^2 = \gamma(0) - \phi_1\gamma(1) - ... - \phi_p\gamma(p)$$

The matrix form is

$$\Gamma_p\tilde{\phi} = \tilde{\gamma}_p$$
$$\sigma_w^2 = \gamma(0) - \tilde{\phi}^T\tilde{\gamma}_p$$

Notice here the $\tilde{\phi}$ and σ_w^2 is what we want to solve. The method of moments replace $\gamma(h)$ by $\hat{\gamma(h)}$

$$\hat{\phi} = \hat{\Gamma}_p^{-1}\hat{\gamma}_p$$
$$\hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\gamma}_p^T\hat{\Gamma}_p^{-1}\hat{\gamma}_p$$

Convergence property:

$$\sqrt{n}(\hat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(0, \sigma_w^2\Gamma_p^{-1})$$

6.1.2 Maximum likelihood estimation

6.1.2.1 MLE for AR(1)

$$x_t = \phi x_{t-1} + w_t$$
$$|\phi| < 1 \qquad\qquad w_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$$

The likelihood of sequence $x_1, ..., x_n$:

$$\begin{aligned}
L(\phi, \sigma_w^2) &= f(x_1) * \prod_{t=2}^{n} f(x_t|x_{t-1}) \\
&= f(x_1) * \prod_{t=2}^{n} \mathcal{N}(\phi x_{t-1}, \sigma_w^2) \\
&= \mathcal{N}(0, \frac{\sigma_w^2}{1-\phi^2}) * \prod_{t=2}^{n} \mathcal{N}(\phi x_{t-1}, \sigma_w^2) \qquad \text{(use } \gamma(0)\text{)} \\
&= \frac{1}{(2\pi\sigma_w^2)^{\frac{n}{2}}(1-\phi^2)^{\frac{-1}{2}}} * \exp\left(\frac{-(1-\phi^2)x_1^2 + \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2}{2\sigma_w^2}\right) \\
&= \frac{1}{(2\pi\sigma_w^2)^{\frac{n}{2}}(1-\phi^2)^{\frac{-1}{2}}} * \exp\left(\frac{-S(\phi)}{2\sigma_w^2}\right) \\
S(\phi) &= (1-\phi^2)x_1^2 + \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2
\end{aligned}$$

Recall:

$$\gamma(h) = \frac{\sigma_w^2}{1-\phi^2}\phi^{|h|}$$

The log likelihood is

$$-\frac{1}{2}(n\log 2\pi + n\log\sigma_w^2 - \log(1-\phi^2)) - \frac{S(\phi)}{2\sigma_w^2}$$

Taking partial derivative and set it 0 we have

$$\hat{\sigma_w}^2 = \frac{S(\hat{\phi})}{n}$$

So now if we plut in $\hat{\sigma}_w^2$ into the original log likelihood equation, and ignore constant, we have

$$\begin{aligned}
\hat{\phi} &= \overset{argmax}{\phi} \; -\frac{n}{2}\log\frac{S(\phi)}{n} + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2}\frac{S(\phi)}{S(\phi)/n} \\
&= \overset{argmin}{\phi} \; \log\frac{S(\phi)}{n} - \frac{1}{n}\log(1-\phi^2) \qquad \text{(also divided everything by } n\text{)}
\end{aligned}$$

The unconditional least square of σ_w^2 is just $\hat{s}(\phi)/2$ (Sample variance divided by 2)

Overall, the asymptotic distribution of parameter estimates same as Gaussian case. But it is a difficult optimization problem, and normally require a good starting point.

6.1.2.2 MLE for ARMA(p,q)

$$x_t - \sum_{i=1}^{p} \phi_i x_{t-i} = w_t + \sum_{i=1}^{q} \theta_i w_{t-i}$$

Let $\beta = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)^T$
The likelihood can be represented as Gaussian with mean of 1 step ahead forecasting and variance being the variance of the forecasting.

$$\begin{aligned}
L(\beta, \sigma_w^2) &= \prod_{t=1}^{n} f(x_t | x_{t-1}, ..., x_1) \\
&= \prod_{t=1}^{n} \mathcal{N}(x_t^{t-1}, P_t^{t-1}) \\
&= \frac{1}{(2\pi)^{n/2}(P_1^0 * ... * P_n^{n-1})^{1/2}} * \exp\left(-\frac{1}{2}\sum_{t=1}^{n}\frac{(x_t - x_t^{t-1})^2}{P_t^{t-1}}\right) \\
&= \frac{1}{(2\pi * \sigma_w^2)^{n/2}(r_1^0 * ... * r_n^{n-1})^{1/2}} * \exp\left(-\frac{S(\beta)}{2\sigma_w^2}\right) \\
r_t^{t-1} &= \frac{{P_t}^{t-1}}{\sigma_w^2} && \text{(a function of } \beta \text{ and } \sigma_w^2\text{)} \\
S(\beta) &= \sum_{t=1}^{n}\frac{(x_t - x_t^{t-1})^2}{r_t^{t-1}} && (x_t^{t-1} \text{ is a function of } \beta)
\end{aligned}$$

So the log likelihood is

$$-\frac{n}{2}\log 2\pi\sigma_w^2 - \frac{1}{2}\sum_{i=1}^{n}\log r_t^{t-1} - \frac{S(\beta)}{2\sigma_w^2}$$

And we have

$$\begin{aligned}
\hat{\sigma_w}^2 &= \frac{S(\hat{\beta})}{n} \\
\hat{\beta} = (\hat{\phi}, \hat{\theta}) &= argmin \log\frac{S(\beta)}{n} + \frac{1}{n}\sum_{t=1}^{n}\log r_t^{t-1}
\end{aligned}$$

If we do unconditional LS, we can drop the $\log r_t^{t-1}$ terms in the log likelihood. In conditional LS (condition on first $m = max(p, q)$ values of time series). Now $r_t^{t-1} = r$ is a constant because we can forecast any x_t equally well. So we have

$$L = \frac{1}{(2\pi\sigma_w^2)^{n-m}} e^{-\frac{1}{2\sigma_w^2 r}} \sum_{t=m+1}^{n} (x_t - x_t^{t-1})^2$$

$$\hat{\sigma}_w^2 = \frac{S(\beta)}{n - m - p - q - 1}$$

$$\hat{\beta} = argmin \qquad S(\beta)$$

6.1.3 Conditional Least Square

6.1.3.1 Conditional Least Square for AR(1)

Conditioning on x_1 removes complicated terms from the log likelihood, so our loglikelihood can be simplified.

$$-\frac{1}{2}(n\log 2\pi + n\log\sigma_w^2 - \log(1-\phi^2)) - \frac{-(1-\phi^2)x_1^2 + \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2}{2\sigma_w^2}$$

$$= -\frac{1}{2}(n\log 2\pi + n\log\sigma_w^2) - \frac{-\sum_{t=2}^{n}(x_t - \phi x_{t-1})^2}{2\sigma_w^2}$$

$$\implies \hat{\sigma_w^2} = \frac{s_c(\hat{\phi})}{n-1} \text{ or } \frac{s_c(\hat{\phi})}{n-2-1}$$

$$\hat{\phi} = \overset{argmax}{\phi} \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2 \qquad \text{(Same obj as Least Squares)}$$

$$= \frac{\sum_{t=2}^{n}(x_t - \frac{1}{n-1}\sum_{k=2}^{n} x_k)(x_{t-1} - \frac{1}{n-1}\sum_{k=1}^{n-1} x_k)}{\sum_{t=2}^{n}(x_{t-1} - \frac{1}{n-1}\sum_{k=1}^{n-1} x_k)}$$

$$\approx \hat{\phi}(1)$$

Chapter 7

ARIMA Model

7.1 Motivation and IMA Models

Interated Moving Average:
A process $\{x_t\}$ is IMA(d,q) if

$$\nabla^d x_t = \theta(B) w_t$$

where w_t is white noise, $\theta(B) = \sum_{j=0}^{q} \theta_j B^j, \theta_0 = 1$. I.e.: the process's dth order difference is a moving average process.

- e.g.: $\nabla^2 x_t = (1-B)^2 x_t = x_t - 2x_{t-1} + x_{t-2}$

7.1.1 EWMA as IMA(1,1)

$$\begin{aligned}
&\nabla x_t = w_t - \lambda w_{t-1} \\
&\implies x_t = x_{t-1} + w_t - \lambda w_{t-1} \\
&\text{Assume } |\lambda| < 1 \text{ and } x_0 = 0
\end{aligned}$$

Alternative form:

$$
\begin{aligned}
& x_t - x_{t-1} = \theta(B) w_t && \theta(B) = 1 - \lambda B \\
& \implies w_t = \theta^{-1}(B)(x_t - x_{t-1}) \\
& = \frac{1}{1-(\lambda B)}(x_t - x_{t-1}) \\
& = \sum_{j=0}^{\infty} \lambda^j B^j (x_t - x_{t-1}) \\
& = \sum_{j=0}^{\infty} \lambda^j x_{t-j} - \sum_{j=0}^{\infty} \lambda^j x_{t-j-1} \\
& = x_t + \sum_{j=1}^{\infty} \lambda^j x_{t-j} - \sum_{j=0}^{\infty} \lambda^j x_{t-j-1} \\
& = x_t + \lambda \sum_{j=1}^{\infty} \lambda^{j-1} x_{t-j} - \sum_{j=1}^{\infty} \lambda^{j-1} x_{t-j} \\
& \implies x_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{t-j} + w_t
\end{aligned}
$$

So the 1 step ahead prediction given infinite history is just

$$\tilde{x}_{n+1} = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j}$$

We also have a very simple updating formula:

$$
\begin{aligned}
\tilde{x}^n_{n+1} &= \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j} \\
&= (1-\lambda)x_n + \sum_{j=2}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j} \\
&= (1-\lambda)x_n + \lambda \sum_{j=2}^{\infty} (1-\lambda)\lambda^{j-2} x_{n+1-j} \\
&= (1-\lambda)x_n + \lambda \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n-j} \\
&= (1-\lambda)x_n + \lambda \tilde{x}^{n-1}_n
\end{aligned}
$$

7.2 ARIMA Models and Workflow

A process $\{x_t\}$ is ARIMA(p,d,q) if

$$\nabla^d x_t = (1-B)^d x_t$$

is ARMA(p,q), hence

$$\phi(B)(1-B)^d x_t = \theta(B) w_t$$

7.2.1 ARIMA forecasting

Theoretical and computational aspects of ARIMA forecasting are better handled using state space models. Intuitively, since $y_t = \nabla^d x_t$, we can use ARMA forecasts y_{n+m}^n and transform back.

7.2.2 ARIMA Workflow

- Plotting the data
- Possibly transforming the data
- Identifying possible order of the AR, differencing, and MA components
- Parameter estiamtion
- Diagonstics
- Model Choice

Diagnostic Fitted Model
Standardized residuals:
Should behave like 0-mean unit variance IID sequence:

$$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}_t^{t-1}}}$$

- Time plot
- QQ plot to detect departure from normality : $\Phi^{-1}(\frac{i-1/2}{n})$ vs. ith order statistics
- Sample ACF for the standardized residuals. Check any large values or patterns
- Ljung-Box-Pierce test: compute Q statistics $Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_e^2(h)}{n-h}$, $Q \sim \chi^2_{H-p-q}$. Often $H = 20$. Detect any significant value at confidence level α

Model Selection Criteria
How to choose if diagnostic for many models are all good? Just as in linear regression case.

AIC: Akaike's information criterion for a model with k parameters

$$\log \hat{\sigma_k}^2 + \frac{n+2k}{n} \qquad \hat{\sigma}_k^2 = \frac{\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}{n}$$

AICc: Bias corrected AIC

$$\log \hat{\sigma_k}^2 + \frac{n+k}{n-k-2}$$

Bayesian Information Criterion (BIC)

$$\log \hat{\sigma_k}^2 + \frac{k \log n}{n}$$

7.3 Seasonal ARIMA

7.3.1 SARMA

A process $\{x_t\}$ is $SARMA(P,Q)_s$ if

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t$$

Where

$$\Phi_p(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_p B^{Ps}$$
$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + ... + \Theta_Q B^{Qs}$$

So essentially we are focusing on past value at the seasonal period.

7.3.2 Multiplicative Seasonal ARIMA Model

A process $\{x_t\}$ is $SARMA(p,q) \times (P,Q)_s$ if

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t$$
$$\Phi_p(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_p B^{Ps}$$
$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + ... + \Theta_Q B^{Qs}$$
$$\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$$
$$\theta(B) = 1 + \sum_{i=1}^{q} \theta_i B^i$$

7.3.3 Multiplicative Seasonal ARIMA Model

A process $\{x_t\}$ is $SARIMA(p,d,q) \times (P,D,Q)_s$ if

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t$$
$$\nabla_s^D = (1-B^s)^D \qquad \text{(Seasonal Differencing)}$$
$$\nabla^d = (1-B)^d$$

Chapter 8

Spectral Analysis

8.1 Representing Stationary Processes as Random Sum of Sines and Cosines

Any stationary process $\{x_t\}$ can be somewhat approximately represented by

$$x_t = \sum_{k=1}^{q} u_k \cos(2\pi\omega_k t) + v_k \sin(2\pi\omega_k t)$$

- u_k and v_k are uncorrelated 0 mean random variable with variance σ_k^2
- q could be quite large

8.1.1 Example: Periodic Stationary Processes with 1 period

$$\begin{aligned}
x_t &= u\cos(2\pi\omega t) + v\sin(2\pi\omega t) \\
E[x_t] &= 0 \\
\gamma(t, t+h) &= cov(u\cos_t + v\sin_t, u\cos_{t+h} + v\sin_{t+h}) \\
&= E[u^2\cos_t\cos_{t+h}] + E[v^2\sin_t\sin_{t+h}] \\
&= \sigma^2(\cos_t\cos_{t+h} + \sin_t\sin_{t+h}) \\
&= \sigma^2(\cos(2\pi\omega(t+h) - 2\pi\omega t)) \qquad \text{(trig identity of } cos(a-b)) \\
&= \sigma^2(\cos(2\pi\omega h) \qquad\qquad \text{(Stationary, because no } t)
\end{aligned}$$

8.1.2 Periodic Stationary Processes with sum over frequency

$$\gamma(h) = \sum_{k=1}^{q} \sigma_k^2 \cos(2\pi\omega_k h)$$

$$var(x_t) = \gamma(0) = \sum_{k=1}^{q} \sigma_k^2$$

8.2 Spectral density

8.2.1 Definition

If the autocovariance function $\gamma(h)$ of a stationary process $\{x_t\}$ satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$$

then the $\gamma(h)$ can be written as the inverse transform of the spectral density $f(\omega)$

$$\gamma(h) = \int_{-0.5}^{0} .5e^{2\pi i\omega h} f(\omega)\, d\omega$$

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i\omega h} \qquad i = \sqrt{-1}$$

- $f(\omega)$ is the Fourier transform of $\gamma(h)$
- $h = 0, \pm 1, \pm 2, ...$
- $-\frac{1}{2} \leq \omega \leq \frac{1}{2}$

8.2.2 Properties

Spectral density is:

- Non-negative
- symmetric
- Perod = 1, hence $f(\omega + 1) = f(\omega)$
- Defines variance $\gamma(0) = var(x_t) = \int_{-0.5}^{0.5} f(\omega)\, d\omega$

8.3 Linear Filtering

For some specified set of filter coefficients a_j (impulse response of filter) with $\sum_{j=-\infty}^{\infty} |a_j| \leq \infty$, linear filtering y_t is

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}$$

Let the frequency response function be the Fourier transform of the impulse response function

$$A(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-2\pi i \omega j}$$

We also know that if $\{x_t\}$ has spectrum $f_x(\omega)$, then the filtered output $\{y_t\}$ has spectrum

$$f_y(\omega) = |A(\omega)|^2 f_x(\omega)$$

8.4 Examples

8.4.1 Spectral density of white noise

Given $\gamma(0) = \sigma_w^2$ and $\gamma(h) = 0$ for $h \neq 0$, we have

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} = \sigma_w^2$$

This means spectral density constant across frequencies. In another word, each frequency contributes equally to variance.

8.4.2 Spectral density of ARMA

For a causal ARMA process, we can write

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \sum_{j=0}^{\infty} |\psi_j| \leq \infty$$

where $\psi(z) = \frac{\theta(z)}{\phi(Z)}$ AND $f_w(\omega) = \sigma_w^2$. Notice this is already in the form of linear filter where $a_j = \psi_j$. So we know

$$\begin{aligned}
A(\omega) &= \sum_{j=0}^{\infty} \psi_j e^{-2\pi i \omega j} \\
&= \sum_{j=0}^{\infty} \psi_j (e^{-2\pi i \omega})^j \qquad \text{(Looks like } \textstyle\sum_{j=0}^{\infty} \psi_j B^j = \psi(B)\text{)} \\
&= \psi(e^{-2\pi i \omega} \\
&= \frac{\theta(e^{-2\pi i \omega})}{\phi(e^{-2\pi i \omega})} \\
\implies f_x(\omega) &= \left| \frac{\theta(e^{-2\pi i \omega})}{\phi(e^{-2\pi i \omega})} \right|^2 \sigma_w^2
\end{aligned}$$

8.4.3 Spectral density of MA(1)

For MA(1), we just have $\theta(z) = 1 + \theta z$. so

$$\begin{aligned}
f_x(\omega) &= \sigma_w^2 \left| 1 + \theta e^{-2\pi i \omega} \right|^2 \\
&= \sigma_w^2 (1 + 2\theta \cos(2\pi\omega) + \theta^2)
\end{aligned}$$

8.4.4 Spectral density of AR(1)

For AR(1), we just have $\phi(z) = 1 - \phi z$

$$\begin{aligned}
f_x(\omega) &= \frac{\sigma_w^2}{\left| 1 - \phi e^{-2\pi i \omega} \right|^2} \\
&= \frac{\sigma_w^2}{1 - 2\phi \cos(2\pi\omega) + \phi^2}
\end{aligned}$$

8.4.5 Spectral density of AR(2)

$$\begin{aligned}
& x_t - x_{t-1} + 0.9 x_{t-2} = w_t \qquad \sigma_w^2 = 1 \\
& \implies \phi(z) = 1 - z + 0.9 z^2 \\
& \implies f_x(\omega) = \left| \phi(e^{-2\pi i \omega} \right|^{-2} \\
& = \left| 1 - e^{-2\pi i \omega} + 0.9 e^{-4\pi i \omega} \right|^{-2} \\
& = \left| 2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega) \right|^{-1}
\end{aligned}$$