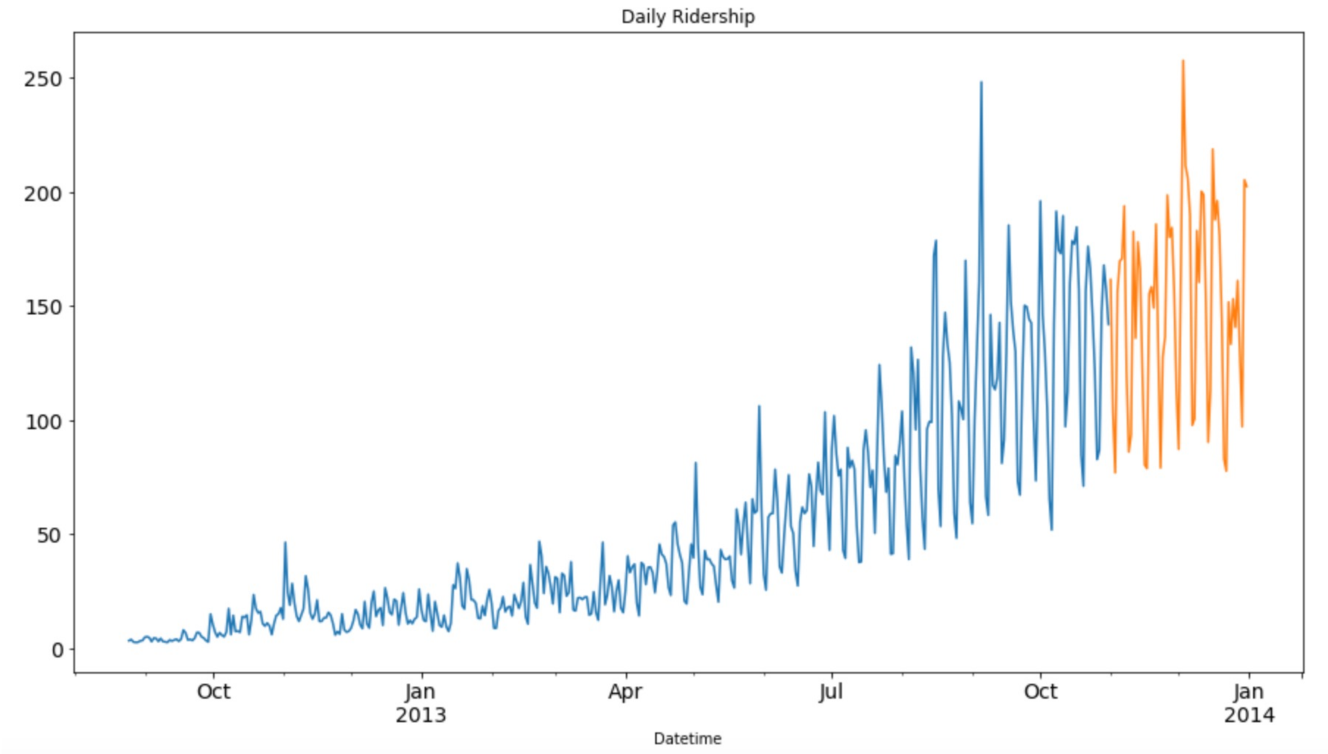
MATH 7339 - Machine Learning and Statistical Learning Theory 2

Section Forecasting

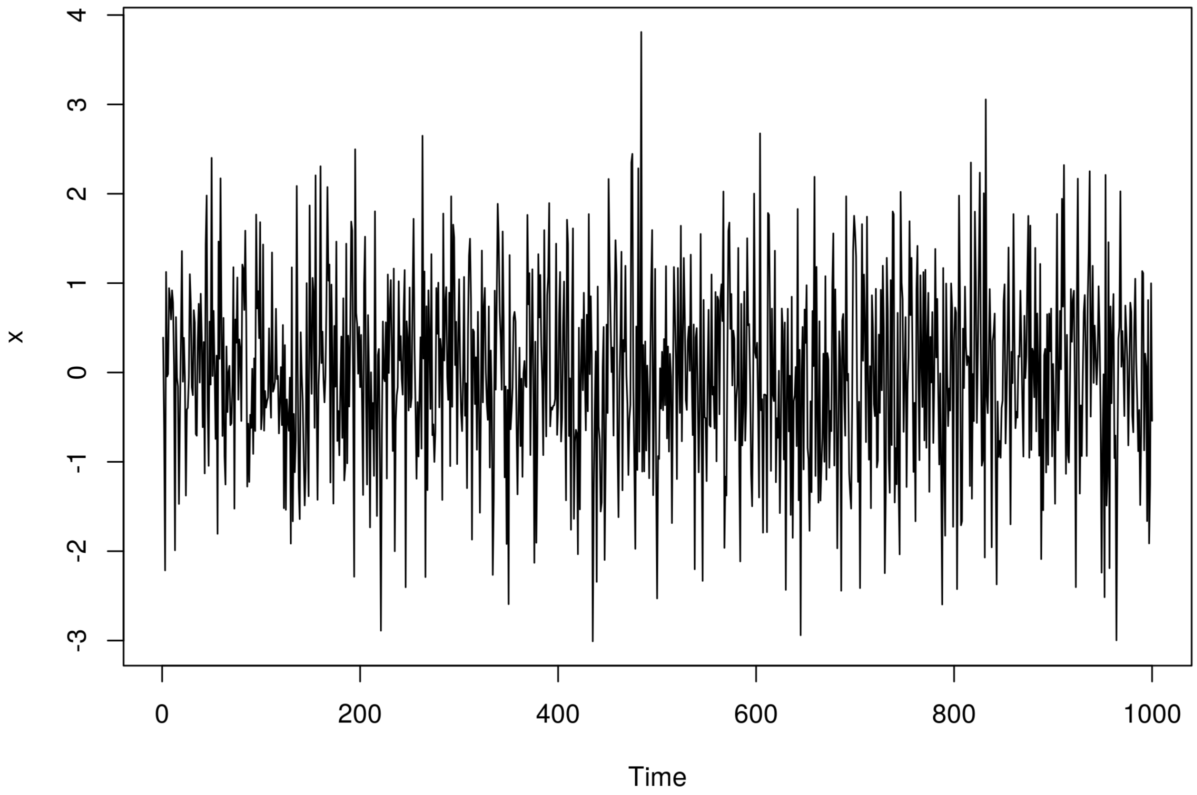
1. Linear prediction.
2. Best linear predictor
3. Forecasting AR and MA models.
4. Building ARMA models

➢ Forecasting

We explore **forecasting**: The **goal** is to predict future values of a time series x_{n+m}, based on the observed data $x = \{x_1, \dots, x_{n-1}, x_n\}$.



We first assume $\{x_t\}$ is **stationary**.



Conditional expectations

Conditional expectations are almost always the way you want to use data to forecast/predict something else.

The **minimum mean square error predictor**

$$x_{n+m}^{n} = E(x_{n+m}|x_{1:n})$$

is the best way to forecast m steps into the future with the data you have in the sense that that point *minimizes mean square error*

$$E[x_{n+m} - g(x_{1:n})]^2$$

where $g(x)$ is any function of the observations.

Linear predictors

First, we restrict our attention to predictors that are linear functions of the observations, i.e.

$$x_{n+m}^{n} = \alpha_0 + \sum_{j=1}^{n} \alpha_j x_j = \sum_{j=0}^{n} \alpha_j x_j$$

where $\alpha_0, \alpha_1, \dots, \alpha_n \in \mathbb{R}\, and\ set\ x_0 = 1$.

Theorem: Linear predictors that minimize the mean square prediction error are called **best linear predictors (BLP),** which is solved by

$$E[(x_{n+m} - x_{n+m}^{n})x_k] = 0, \ \text{for } k = 0,1,2, \dots, n$$

where $x_0 = 1$.

Linear prediction depends on on the second-order moments of the process, which can be estimated from the data $\{x_1, \dots, x_{n-1}, x_n\}$.

The reason for the above theorem is similar as we did before. We want to minimize the mean square error

$$E[(x_{n+m} - x_{n+m}^n)^2] = E\left[\left(x_{n+m} - \sum_{j=0}^{n} \alpha_j x_j\right)^2\right]$$

Take partial derivative with respect to α_j, and find critical points, i.e.,

$$E\left[\left(x_{n+m} - \sum_{j=0}^{n} \alpha_j x_j\right) x_k\right] = 0 \qquad \text{for } k = 0,1,2,\ldots,n$$

So, we will find

$$x_{n+m}^n = \alpha_0 + \sum_{j=1}^{n} \alpha_j x_j$$

by solving the above equations.

Remarks:

- Every time we get a new data point, we have to recalculate the prediction.
- Every time we get a new data point, there is another equation in a new system of equations
- Prediction errors $x_{n+m} - x_{n+m}^{n}$ are are orthogonal/uncorrelated with the prediction variables $(1, x_1, \ldots, x_{n-1}, x_n)$.

- We want recursive formula: get x_{n+m}^{n} from x_{n-1+m}^{n-1}
- We also want our recursive formulas to have bounded memory
- footprints.
- Different algorithms work for different models: innovations algorithm, Durbin-Levinson, Kalman filter, particle filters, etc.
- There are algorithms that assume you have an infinite past of history.

From the first equation when $x_0 = 1$,

$$E\left[\left(x_{n+m} - \sum_{j=0}^{n} \alpha_j x_j\right)\right] = 0$$

we have

$$\alpha_0 = \mu\left[1 - \sum_{j=0}^{n} \alpha_j\right]$$

So, the general form of the forecast is

$$x_{n+m}^{n} = \mu + \sum_{j=1}^{n} \alpha_j (x_j - \mu)$$

So, we can assume without loss of generality that $\mu = 0$.

❑ One-step-ahead linear prediction

Now, we focus on **one-step-ahead**, and change the notations

$$x_{n+1}^{n} = \alpha_n x_n + \alpha_{n-1} x_{n-1} + \cdots + \alpha_1 x_1$$

$$= \phi_{n1} x_n + \phi_{n2} x_{n-1} + \cdots + \phi_{nn} x_1$$

The new notation $\phi_{nj} := \alpha_{n+1-j}$ more useful for pure AR(p) models.

$$E\left[\left(x_{n+m} - \sum_{j=1}^{n} \phi_{nj} x_{n+1-j}\right) x_{n+1-k}\right] = 0 \qquad \text{for } k = 1,2,\ldots,n$$

Equivalently,

$$\gamma_x(k) - \sum_{j=1}^{n} \phi_{nj} \gamma_x(k-j) = 0$$

Equivalently,

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix} \begin{bmatrix} \phi_{n1} \\ \phi_{n2} \\ \vdots \\ \phi_{nn} \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(n) \end{bmatrix}$$

Equivalently,

$$\Gamma_n \overrightarrow{\phi_n} = \overrightarrow{\gamma_n}$$

For each data set size n, we solve the above equation for $\overrightarrow{\phi_n}$

$$\overrightarrow{\phi_n} = \Gamma_n^{-1} \overrightarrow{\gamma_n}$$

Then, the **forecasting** is given by

$$x_{n+1}^n = \phi_{n1} x_n + \phi_{n2} x_{n-1} + \cdots + \phi_{nn} x_1 = \overrightarrow{\phi_n}^T \vec{x}$$

❑ Mean squared error of one-step-ahead linear prediction

Once we have the forecast from the weight vector solution, we can calculate its variance (mean square error):

$$E[(x_{n-1} - x_{n-1}^{n})^2] = E\left[\left(x_{n-1} - \overrightarrow{\phi_n}^T \vec{x}\right)^2\right]$$

$$= E[(x_{n-1})^2] - 2E\left[x_{n-1}\overrightarrow{\phi_n}^T \vec{x}\right] + E\left[\left(\overrightarrow{\phi_n}^T \vec{x}\right)^2\right]$$

$$= \gamma_x(0) - 2E[x_{n-1}\vec{x}^T]\overrightarrow{\phi_n} + \overrightarrow{\phi_n}^T E[\vec{x}\vec{x}^T]\overrightarrow{\phi_n}$$

$$= \gamma_x(0) - 2\overrightarrow{\gamma_n}^T\overrightarrow{\phi_n} + \overrightarrow{\phi_n}^T \Gamma_n\overrightarrow{\phi_n}$$

$$= \gamma_x(0) - 2\overrightarrow{\gamma_n}^T\overrightarrow{\phi_n} + \overrightarrow{\phi_n}^T \overrightarrow{\gamma_n}$$

$$= \gamma_x(0) - \overrightarrow{\gamma_n}^T\overrightarrow{\phi_n}$$

$$= \gamma_x(0) - \overrightarrow{\gamma_n}^T \mathit{\Gamma}_n^{-1}\, \overrightarrow{\gamma_n}$$

The term $\overrightarrow{\gamma_n}^T\overrightarrow{\phi_n}$ is the reduction in uncertainty and depends on how much autocorrelation you have.

Variance is reduced:

$$E[(x_{n-1} - x_{n-1}^{n})^2] = \gamma_x(0) - \overrightarrow{\gamma_n}^T \Gamma_n^{-1} \overrightarrow{\gamma_n}$$

$$= Var(x_{n+1}) - Cov(x_{n+1}, \vec{x})\, Cov(\vec{x}, \vec{x})^{-1} Cov(\vec{x}, x_{n+1})$$

$$= E[(x_{n-1} - 0)^2] - Cov(x_{n+1}, \vec{x})\, Cov(\vec{x}, \vec{x})^{-1} Cov(\vec{x}, x_{n+1})$$

Here $\vec{x} = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_1 \end{bmatrix}$

Example: Forecasting AR(1)

Consider an AR(1) model: $x_{t+1} = \phi_1 x_t + w_t$

The predictions/forecasts for $x_2^1 = \phi_{11} x_1$

Prediction equation: $\Gamma_n \overrightarrow{\phi_n} = \overrightarrow{\gamma_n}$

That is $\gamma(0)\phi_{11} = \gamma(1)$

$$\gamma(1) = Cov(x_0, x_1) = \phi_1 \gamma(0)$$

So, $\phi_{11} = \phi_1$

Example: AR(2)

Consider an AR(2) model: $x_{n+1} = \phi_1 x_n + \phi_2 x_{n-1} + w_n$

Find the predictions/forecasts for $n = 2, n = 3, \ldots$

At $n = 2$, forecasting time 3 requires solving

$$\Gamma_n \overrightarrow{\phi_n} = \overrightarrow{\gamma_n}$$

Equivalently,

$$\begin{bmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{bmatrix} \begin{bmatrix} \phi_{n1} \\ \phi_{n2} \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \end{bmatrix}$$

More explicitly,

$$E\left[\left(x_{n+1} - \sum_{j=1}^{n} \phi_{nj} x_{n+1-j}\right) x_{n+1-k}\right] = 0$$

for $k = 1,2$

Here, $\phi_{21} = \phi_1$ and $\phi_{22} = \phi_2$.

More generally, assume that n > 2. The prediction equations are now for $k \; = \; 1, \ldots, n$

$$E\left[\left(x_{n+1} - \sum_{j=1}^{n} \phi_{nj} x_{n+1-j}\right) x_{n+1-k}\right] = 0$$

Hence, the solution is

$$\begin{bmatrix} \phi_{n1} \\ \phi_{n2} \\ \phi_{n3} \\ \vdots \\ \phi_{nn} \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Example: Forecasting an AR(2) model

Consider the following AR(2) model.

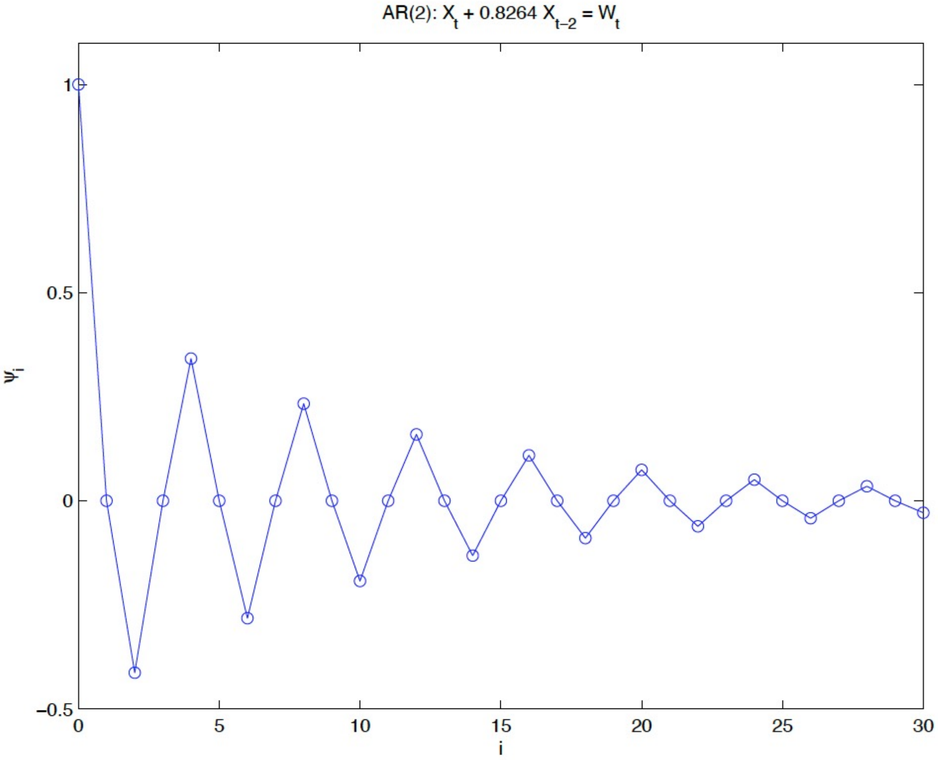
$$X_t + \frac{1}{1.21} X_{t-2} = W_t$$

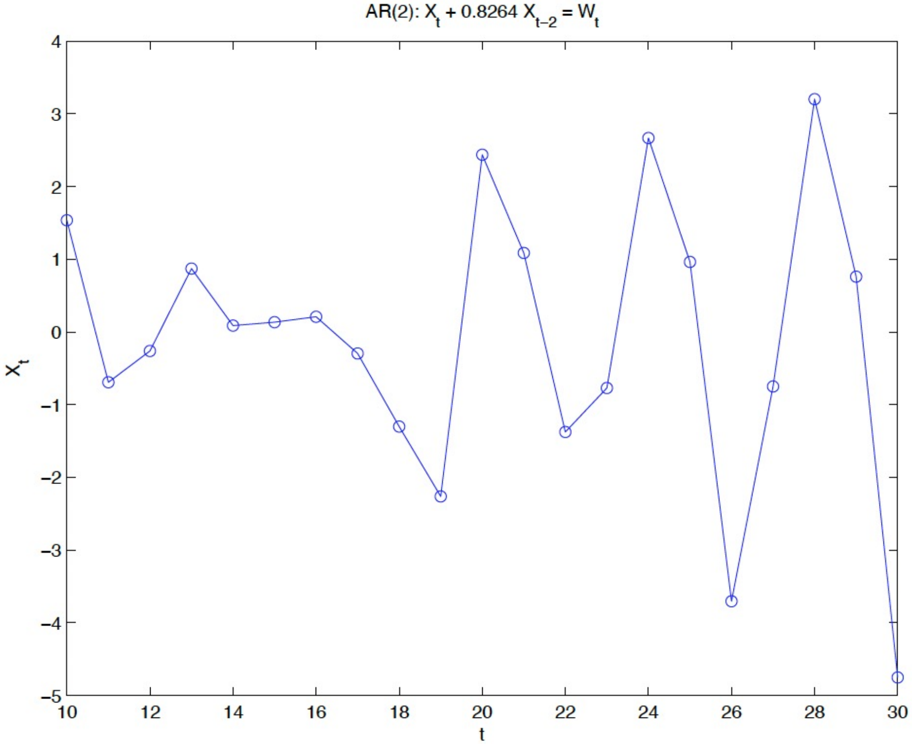
The zeros of the characteristic polynomial $z^2 + 1.21$ are at $\pm 1.1i$. We can solve the linear difference equations $\psi_0 = 1, \phi(B)\psi_t = 0$ to compute the $MA(\infty)$ representation:

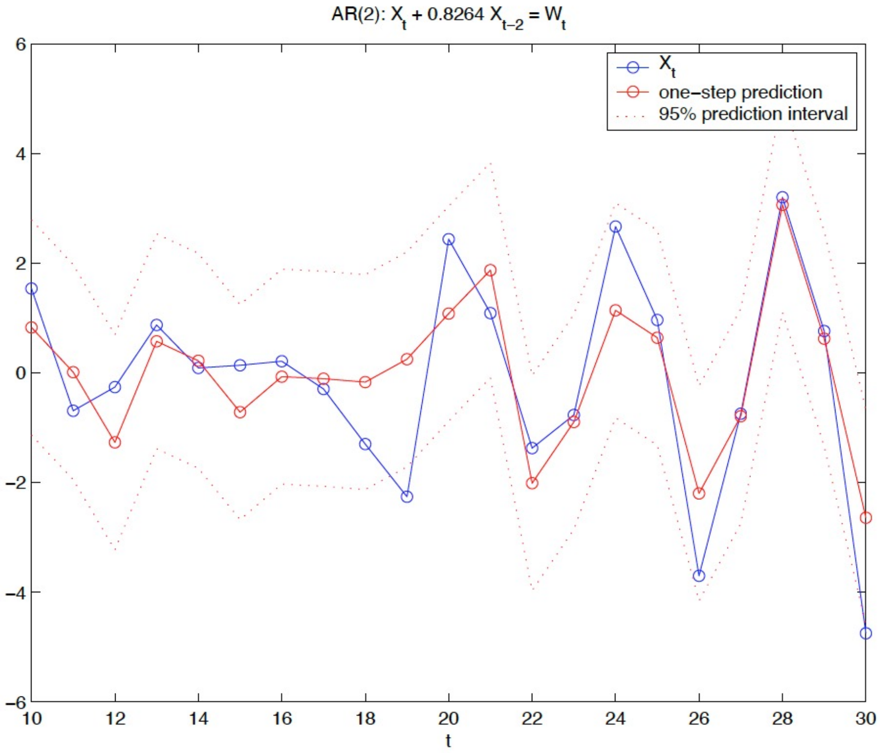
$$\psi_t = \frac{1}{2}(1.1)^{-t} \cos\left(\frac{\pi t}{2}\right)$$

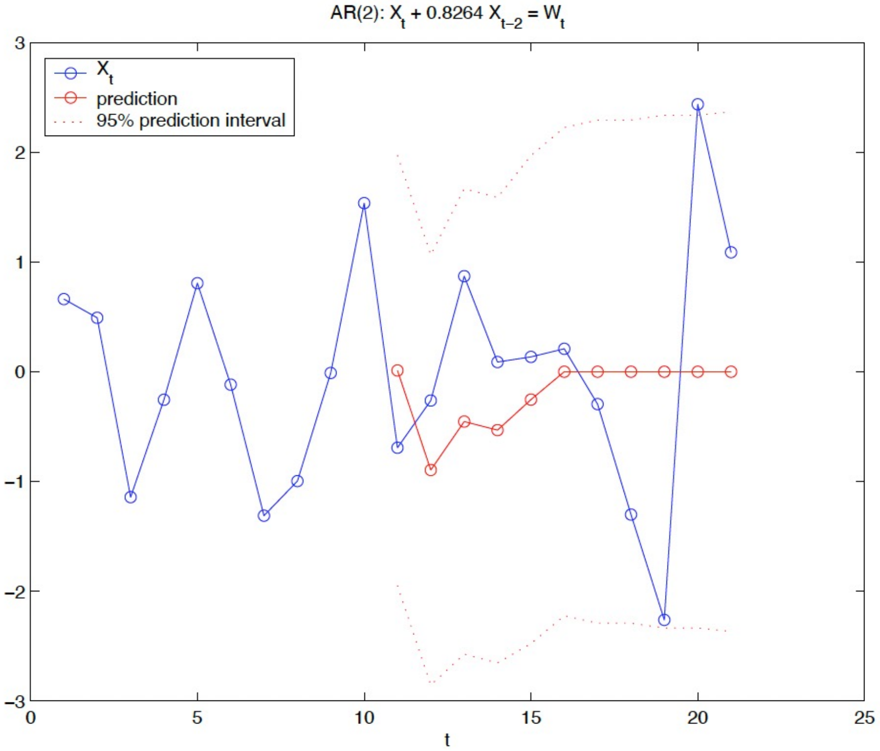
Thus, the m −step-ahead estimates have mean squared error

$$E\left(X_{n+m} - \tilde{X}_{n+m}\right)^2 = \sum_{j=0}^{m-1} \psi_j^2$$


AR(2): X_t + 0.8264 X_{t-2} = W_t
ψ_i
1
0.5
0
−0.5
0
5
10
15
20
25
30
i


AR(2): X_t + 0.8264 X_{t-2} = W_t
X_t
4
3
2
1
0
−1
−2
−3
−4
−5
10
12
14
16
18
20
22
24
26
28
30
t


AR(2): $X_t + 0.8264\, X_{t-2} = W_t$
X_t
one−step prediction
95% prediction interval
6
4
2
0
−2
−4
−6
10
12
14
16
18
20
22
24
26
28
30
t


AR(2): $X_t + 0.8264\, X_{t-2} = W_t$
X_t
prediction
95% prediction interval
3
2
1
0
−1
−2
−3
0
5
10
15
20
25
t

Example: AR(p)

Consider an AR(p) model: $x_{t+1} = \phi_1 x_t + \phi_2 x_{t-1} + \cdots + \phi_p x_{t-p+1} + w_t$

Assume at time $t \geq p$ we want to predict/forecast time $t + 1$ with an AR(p) model for which we know the parameters $\phi_1, \ldots, \phi_p$. We would just use the previous p time points $t, t-1, \ldots, t-p+1$, and use prediction $\phi_1 x_t + \cdots + \phi_p x_{t-p+1}$:

$$
\begin{aligned}
x_{n+1}^{n} &= P(X_{n+1} | X_1, \ldots, X_n) \\
&= P\left(\textstyle\sum_{i=1}^{p} \phi_i X_{n+1-i} + W_{n+1} \middle| X_1, \ldots, X_n\right) \\
&= \sum_{i=1}^{p} \phi_i \, P(X_{n+1-i} | X_1, \ldots, X_n) \\
&= \sum_{i=1}^{p} \phi_i \, X_{n+1-p} \qquad \text{for } n \geq p.
\end{aligned}
$$

In more details, the **Durbin-Levinson** algorithm gives the calculation for AR(p) models

Remark (**Innovations Algorithm** for MA(q))

The Innovations Algorithm is more useful for pure MA(q) models, and can be extended to ARMA(p,q) models. The idea is to write predictions in terms of

$$\hat{x}_{n+1}^{n} = \sum_{j=1}^{n} \theta_{nj}(x_{n+1-j} - \hat{x}_{n+1-j})$$

instead of

$$\hat{x}_{n+1}^{n} = \sum_{j=1}^{n} \phi_{nj}\; x_{n+1-j}$$

The terms $x_{n+1-j} - \hat{x}_{n+1-j}$ are called the innovations, and are kind of like w_{n+1-j}

The prediction operator

For random variables $Y, Z_1, \ldots, Z_n$, the ***best linear prediction of Y given*** $\vec{\boldsymbol{Z}}$ is the operator $P(-|\vec{Z}\ )$ applied to Y

$$P\left(Y\middle|\vec{Z}\right) = \mu_Y + \vec{\phi}^T(\vec{Z} - \mu_{\vec{Z}})$$

with $\Gamma\vec{\phi} = \vec{\gamma}$

where

$$\gamma = Cov\left(Y, \vec{Z}\right)$$

$$\Gamma = Cov\left(\vec{Z}, \vec{Z}\right)$$

Properties of the prediction operator

$$E\left(Y - P\left(Y\middle|\vec{Z}\right)\right) = 0$$

$$E\left[\left(Y - P\left(Y\middle|\vec{Z}\right)\right)\vec{Z}\right] = \vec{0}$$

$$E\left(Y - P\left(Y\middle|\vec{Z}\right)^2\right) = Var(Y) - \vec{\phi}^T\vec{\gamma}$$

$$P\left(Z_i\middle|\vec{Z}\right) = Z_i$$

$$P\left(Y\middle|\vec{Z}\right) = E[Y] \text{ if } \gamma = 0$$

$$P(\alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2 \mid Z) = \alpha_0 + \alpha_1 P(Y_1|Z) + \alpha_2 P(Y_2|Z)$$

Example: predicting m steps ahead

$$x_{n+m}^{n} = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \cdots + \phi_{nn}^{(m)} x_1 = \overrightarrow{\phi_n}^{(m)^T} \vec{x}$$

Solve equation $\Gamma_n \overrightarrow{\phi_n}^{(m)} = \overrightarrow{\gamma_n}^{(m)}$

where $\Gamma_n = Cov(\vec{x}, \vec{x})$ and $\overrightarrow{\gamma_n}^{(m)} = Cov(x_{n+m}, \vec{x}) = \begin{bmatrix} \gamma(m) \\ \gamma(m+1) \\ \vdots \\ \gamma(m+n-1) \end{bmatrix}$

In addition, the variance (error)

$$E[(x_{n+m} - x_{n+m}^{n})^2] = \gamma(0) - \overrightarrow{\phi_n}^{(m)^T} \overrightarrow{\gamma_n}^{(m)}$$

Review: (Real Hilbert Space)

Let V be a **real vector space**. For example, V is a subspace of $\mathbb{R}^n$.

Definition (Inner Product). An **inner product** on V is a binary function

$$\langle -,- \rangle : V \times V \longrightarrow \mathbb{R}$$

such that for vectors $\vec{u}, \vec{v}, \vec{w} \in V$ and a scalar $c \in \mathbb{R}$, the following hold:

(1.) $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$
(2.) $\langle \vec{u} + \vec{v},\ \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle + \langle \vec{v}, \vec{w} \rangle$
(3.) $\langle c\vec{u}, \vec{v} \rangle = c\langle \vec{v}, \vec{u} \rangle$
(4.) $\langle \vec{u}, \vec{u} \rangle \geq 0$
(5.) $\langle \vec{u}, \vec{u} \rangle = 0$ if and only if $\vec{u} = \vec{0}$

- We call V an **inner product space** with inner product $\langle -,- \rangle$.
- **Hilbert spaces** is a **complete** inner product space.

A sequence $\{x_n\}$ is called **Cauchy sequence** if for every $\epsilon \in \mathbb{R}$, there is a positive integer N such that $|x_m - x_n| < \epsilon$ all natural numbers $m, n > N$.

Complete means that the limits of **Cauchy sequences** $\{x_n\}$ are in the space

Examples:

1. Euclidean inner product space $\mathbb{R}^n$

2. {Random Variables X | $E(X^2) < \infty$} with inner product $\langle X, Y\rangle \coloneqq E(XY)$

3. Rational numbers $\mathbb{Q}$ is not complete.

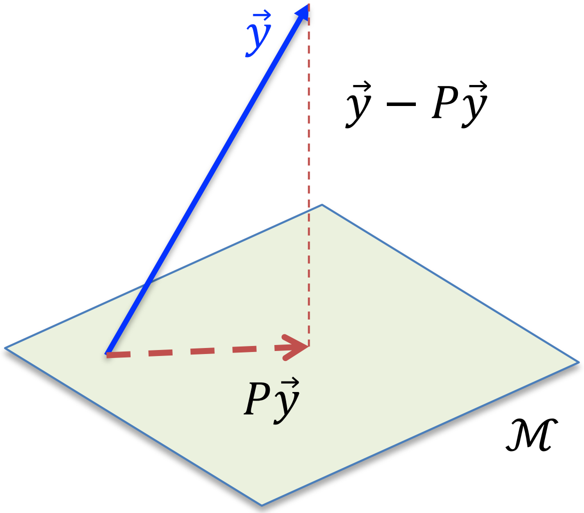
Theorem (Orthogonal Projection Theorem)

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{M}$ be a *closed* linear subspace of $\mathcal{H}$, and $\vec{y} \in \mathcal{H}$.

There exist an **unique** point $P\vec{y} \in \mathcal{M}$ such that

1. $\|\vec{y} - P\vec{y}\| \leq \|\vec{w} - P\vec{y}\|$ for any $\vec{w} \in \mathcal{M}$

2. $\langle \vec{y} - P\vec{y}, \vec{w}\rangle = 0$ for any $\vec{w} \in \mathcal{M}$



Application: Linear prediction

Let $W = Span\{1, X_1, \dots, X_n\}$ be subspace of the Hilbert space of all random variables Z such that $E(Z^2) < \infty$.

So, $W = \{\alpha_0 + \sum_{j=1}^{n} \alpha_j x_j\}$

Inner product $\langle X, Y \rangle \coloneqq E(XY)$

$\vec{y} := X_{n+m}$

Apply the Projection theorem, we know the best linear predictor X_{n+m}^{n} is give by the orthogonal projection of $\vec{y}$ onto W. From condition 2, we have

$$E[(X_{n+m}^{n} - X_{n+m})X_i] = 0 \text{ for any } i = 0,1, \dots, n$$

So, the prediction errors $(X_{n+m}^{n} - X_{n+m})$ are orthogonal to the prediction variables $(1, X_1, \dots, X_n)$.

Review: Time series modelling and forecasting

1. Plot the time series.

   Look for trends, seasonal components, step changes, outliers.

2. Transform data so that residuals are stationary.

   (a) Remove trend and seasonal components.

   (b) Differencing.

   (c) Nonlinear transformations (log, $\sqrt{-}$, etc .

3. Fit model to residuals.

4. Forecast time series by forecasting residuals and inverting any transformations.

Stationary time series models: ARMA(p,q).

- p = 0: MA(q),
- q = 0: AR(p).

We have seen that any **causal**, **invertible** linear process has:

  - an MA(∞) representation (from causality), and
  - an AR(∞) representation (from invertibility).

Real data cannot be exactly modelled using a finite number of parameters.

We choose p, q to give a simple but accurate model.

Question: How do we use data to decide on p, q?

1. Use sample ACF/PACF to make preliminary choices of model order.

2. Estimate parameters for each of these choices.

3. Compare predictive accuracy/complexity of each (using, e.g., AIC).

❑ Parameter estimation

We need to compute parameter estimates for several different model orders. Thus, recursive algorithms for parameter estimation are important. Some of these are identical to the recursive algorithms for forecasting.

Estimate the parameters of an ARMA(p,q) model.

$$\phi(B)X_t = \theta(B)W_t$$

We will assume (for now) that:

1. The model order (p and q) is known, and
2. The data has zero mean. (We can subtract the sample mean if not.)

We explore a couple of ways to estimate the parameters for ARMA models: Method of Moments (MOM) estimation and Maximum Likelihood (ML) estimation.

❑ Parameter estimation: Method of Moments

Let's start with the method of moments (MOM) estimation. The idea behind this is to equate population moments to sample moments and then solve for the parameters in terms of the sample moments. *We re-use a lot of the same equations from the previous forecasting.*

Let's first assume that we have a **causal** AR(p) model

$$\phi(B)(X_t - \mu) = W_t$$

where the white noise W_t has variance σ_W^2 and

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

Given n observations $x_1, x_2, \dots, x_n$, we are interested in estimating the parameters $\phi_1, \dots, \phi_p$ and σ_W^2 . Initially we assume that the order p is known.

$E[X_t] = \mu$ can always be estimated with the first sample moment $\bar{x}$.

We then transform the data before estimating as follows mean zero by $x_t - \bar{x}$.

The method of moments works well when estimating causal AR(p) models.

Consider a casual AR(p) model:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t$$

For $h = 1, \ldots, p$, multiply both sides by x_{t-h} and take **expectations**:

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2) + \cdots + \phi_p \gamma(h-p)$$

When $h = 0$, we do the same thing and get

$$\gamma(0) = \phi_1 \gamma(1) - \phi_2 \gamma(2) - \cdots - \phi_p \gamma(p)$$

We call these the **Yule-Walker equations**

We can also write them in matrix notation that should look familiar:

$$\Gamma_p \vec{\phi} = \vec{\gamma}$$

$$\sigma_W^2 = \gamma(0) - \vec{\phi}^T \vec{\gamma}$$

Use the sample data, solve for the desired parameters:

$$\hat{\vec{\phi}} = \widehat{\Gamma_p}^{-1} \hat{\vec{\gamma}}$$

$$\sigma_W^2 = \hat{\gamma}(0) - \hat{\vec{\gamma}}^T \widehat{\Gamma_p}^{-1} \hat{\vec{\gamma}}$$

The asymptotic behavior of the Yule-Walker estimators for causal AR(p) processes is (when p large)

$$\sqrt{n}\left(\hat{\vec{\phi}} - \vec{\phi}\right) \longrightarrow N\left(0, \sigma_W^2 \widehat{\Gamma_p}^{-1}\right)$$

$$\hat{\sigma}_W^2 \longrightarrow \sigma_W^2$$

So, we have Confidence intervals for Yule-Walker estimation

Method of Moments Estimation for MA(q)

Consider an invertible MA(1) process $X_t = W_t + \theta W_{t-1}$ with $|\theta| < 1$

We know that

$$\rho(1) = \frac{\theta}{1+\theta^2}$$

Using method of moments, we equate $\hat{\rho}(1)$ to $\rho(1)$ and solve a quadratic equation in θ.

For higher order MA(q) models, the method of moments quickly gets complicated. The equations are non-linear in $\theta_1, \dots, \theta_q$, so numerical methods need to be used.

❑ Parameter estimation: Maximum likelihood estimator

Assume that $\{X_t\}$ is Gaussian, that is, $\phi(B)X_t = \theta(B)W_t$, where W_t is i.i.d. Gaussian.

Choose ϕ_i, θ_j to maximize the likelihood:

$$L(\vec{\phi}, \vec{\theta}, \sigma^2) = f_{\vec{\phi},\vec{\theta},\sigma^2}(X_1, \dots, X_n)$$

Here $f_{\vec{\phi},\vec{\theta},\sigma^2}$ is the joint (Gaussian) density for the given ARMA model.
(c.f. choosing the parameters that maximize the probability of the data.)

Suppose that $X_1, X_2, \ldots, X_n$ is drawn from a zero mean Gaussian ARMA(p,q) process.

$$L(\vec{\phi}, \vec{\theta}, \sigma^2) = \frac{1}{(2\pi)^{n/2} |\Gamma_n|^{1/2}} \exp\left(-\frac{1}{2} X^T \Gamma_n^{-1} X\right)$$

Here Γ_n is the variance/covariance matrix of X with the given parameter values. The maximum likelihood estimator (MLE) of $\vec{\phi}, \vec{\theta}, \sigma^2$ maximizes this quantity.

Advantages of MLE:
Efficient (low variance estimates).
Often the Gaussian assumption is reasonable.
Even if $\{X_t\}$ is not Gaussian, the asymptotic distribution of the estimates of $\vec{\phi}, \vec{\theta}, \sigma^2$ is the same as the Gaussian case.
Disadvantages of MLE:
Difficult optimization problem.
Need to choose a good starting point (often use other estimators for this).

Example

Consider the AR(1) model with nonzero mean

$$X_t = \mu + \phi(X_{t-1} - \mu) + W_t$$

where $|\phi| < 1$ and W_t is iid $Normal(0, \sigma^2)$

The likelihood:

$$L(\vec{\phi}, \vec{\theta}, \sigma^2) = f_{\vec{\phi}, \vec{\theta}, \sigma^2}(X_1, \dots, X_n)$$

is functionally equivalent to the joint probability distribution of the observed data $x_1, \dots, x_n$

For a given data set, think of the likelihood as a function of the parameters (not the data).

$$L(\vec{\phi}, \vec{\theta}, \sigma^2) = f(x_1, \ldots, x_t)$$

$$= f(x_1) f(x_2|x_1) f(x_3|x_2, x_1) \ldots f(x_t|x_{t-1}, \ldots, x_1)$$

$$= f(x_1) f(x_2|x_1) f(x_3|x_2) \ldots f(x_t|x_{t-1})$$

These are all the same:

$$X_t = \mu + \phi(X_{t-1} - \mu) + W_t \qquad W_t \sim Normal(0, \sigma^2)$$

So, $X_t|X_{t-1} \sim Normal(\mu + \phi(X_{t-1} - \mu), \sigma^2)$

$$L(\vec{\phi}, \vec{\theta}, \sigma^2) = f_{x_1}(x_1)\,(2\pi\sigma^2)^{-(n-1)/2} \exp\left(-\frac{1}{2}\frac{\sum_{t=2}^{n}\left(x_t - \mu - \phi(x_t - \mu)\right)^2}{\sigma^2}\right)$$

Take log likelihood and partial derivatives.

Building ARMA models

1. Plot the time series.

Look for trends, seasonal components, step changes, outliers.

2. Nonlinearly transform data, if necessary

3. Identify preliminary values of p, and q.

4. Estimate parameters.

5. Use diagnostics to confirm residuals are white/iid/normal.

6. Model selection: Choose p and q.

Question: How do we check that a model fits well?

The residuals (innovations, $x_t - x_t^{t-1}$) should be white noise.

Consider the standardized innovations,

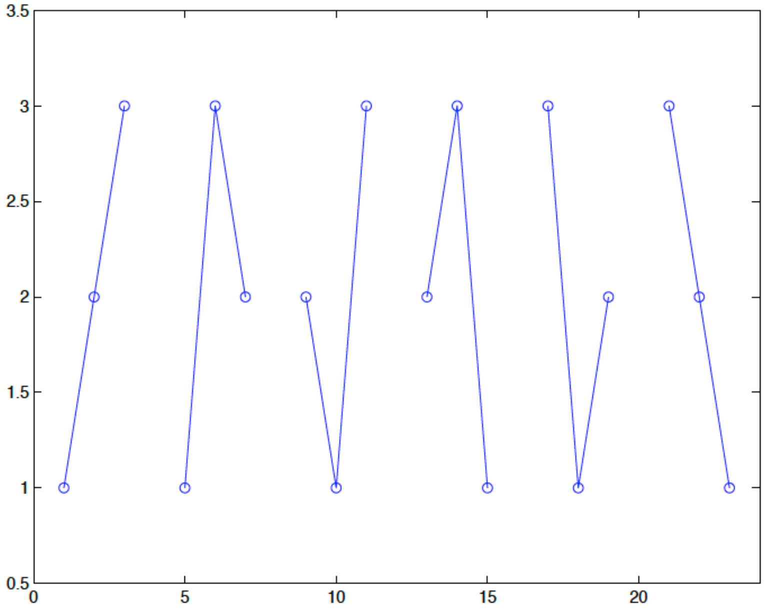
$$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\widehat{P}_t^{\,t-1}}}$$

This should behave like a mean-zero, unit variance, iid sequence.

- Check a time plot
- Turning point test
- Difference sign test
- Rank test
- Q-Q plot, histogram, to assess normality

Testing i.i.d.: Turning point test

$\{X_t\}$ i.i.d. implies that $X_t, X_{t+1}\ and\ X_{t+2}$ are equally likely to occur in any of six possible orders:



provided X_t, X_{t+1}, X_{t+2} are distinct

Four of the six are turning points.

Define $T = \{t : X_t, X_{t+1}, X_{t+2}$ is a turning point$\}$.

$ET = (n-2)2/3$

Can show $T \sim N\left(\frac{2n}{3}, \frac{8n}{45}\right)$

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left|T - \frac{2n}{3}\right| > 1.96\sqrt{\frac{8n}{45}}$$

Tests for positive/negative correlations at lag 1.

Testing i.i.d.: Difference-sign test

$$S = \{i | X_i > X_{i-1}\} = \{i | (\nabla X)_i > 0\}$$

$$E(S) = \frac{n-1}{2}$$

Can show $S \sim N(n/2, n/12)$.

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left|S - \frac{n}{2}\right| > 1.96\sqrt{\frac{n}{12}}$$

Tests for trend.

(But a periodic sequence can pass this test...)

Testing i.i.d.: Rank test

$$N \;=\; \{(i,j)|X_i \;>\; X_j \; and \; i \;>\; j\}|.$$

$$EN = n(n-1)/4$$

$$Can\ show\ N \sim Normal(n^2/4, n^3/36)$$

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left|N - \frac{n^2}{4}\right| > 1.96\sqrt{\frac{n^3}{36}}$$

Tests for linear trend.

Testing if an i.i.d. sequence is Gaussian: qq plot

Plot the pairs $(m_1, X_{(1)})$, …, $(m_n, X_{(n)})$, where $m_j = EZ_{(j)}$

$Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ are order statistics from N(0, 1) sample of size n, and

$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are order statistics from series $X_1, \dots, X_n$

Idea: If $X_i \sim N(\mu, \sigma)$

$$EX_{(j)} = \mu + \sigma m_{(j)}$$

so $(m_j, X_{(j)})$ should be linear.

There are tests based on how far correlation of $(m_j, X_{(j)})$ is from 1

❑ Model Selection

We have used the data x to estimate parameters of several models. They all fit well (the innovations are white). We need to choose a single model to retain for forecasting. How do we do it?

If we had access to independent data y from the same process, we could compare the likelihood on the new data, $L_y(\vec{\phi}, \vec{\theta}, \sigma^2)$

We could obtain y by leaving out some of the data from our model-building, and reserving it for model selection. This is called **cross-validation**. It suffers from the drawback that we are not using all of the data for parameter estimation.

Model Selection: AIC

We can approximate the likelihood defined using independent data: asymptotically

$$-\ln\ L_y\left(\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}^2\right) \approx -\ln L_x\left(\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}^2\right) + \frac{(\mathrm{p}+\mathrm{q}+1)\mathrm{n}}{\mathrm{n}-\mathrm{p}-\mathrm{q}-2}$$

AIC_c : corrected Akaike information criterion.

Notice that:

- More parameters incur a bigger penalty.
- Minimizing the criterion over all values *p,q,* $\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}^2$ corresponds to choosing the optimal $\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}^2$ w for each p, q, and then comparing the penalized likelihoods.

There are also other criteria: BIC.

Textbook[Shumway-Stoffer]: Chapter 3.4.

MATLAB Examples:

Time Series Regression: Forecasting

https://www.mathworks.com/help/econ/time-series-regression-vii-forecasting.html

Lagged Variables and Estimator Bias

https://www.mathworks.com/help/econ/time-series-regression-viii-lagged-variables-and-estimator-bias.html

Model the United States Economy

https://www.mathworks.com/help/econ/modeling-the-united-states-economy.html

Time Series Prediction and Forecasting for Prognosis

https://www.mathworks.com/help/ident/ug/time-series-prediction-and-forecasting-for-prognosis.html