Linear Prediction

Introduction

STAT 464 / 864 | Fall 2024

Discrete Time Series Analysis

Skyepaphora Griffith, Queen's University

Nerd time! [don't write this down]

Hey Skye, when would we actually want to model an ARMA?

ARMA(p,q) are just really good at modelling (3.1 of Brocky & Davey):

- ARMA ACVFs can effectively approximate a huge class of ACVFs
- For any positive integer K, there's an ARMA process {Xt} such that

$$\gamma_X(h) = \gamma(h) \text{ for } h = 0, 1, \ldots, K.$$

We can leverage cool properties

Given AR(p) and MA(q) parameters, say we build an ARMA(p,q) model.

Describes **(unique) stationary** {Xt} if and only if
[no roots on unit circle → can model as stationary]

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \quad \text{for all } |z| = 1.$$

ARMA(p,q) process {Xt} is **causal** if and only if
[no roots in unit disk → can model using only past values]

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \quad \text{for all } |z| \leq 1.$$

- Later, we'll see similar tools regarding something called **invertibility**
- Also we'll see data-driven examples in workshop eventually

Nerd time! [don't write this down]

Also uhh what about Yule-Walker?

- Not time for that yet (that's chapter 5) but nice pull
- We're wrapping up a unit on linear filters → thesis of the lesson:

We can model an ARMA(1,1) process using a linear filter

$$\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{j+h} = \begin{cases} \sigma^2\left(1 + \frac{(\phi+\theta)^2}{1-\phi^2}\right) & h = 0 \\ \sigma^2(\phi+\theta)\phi^{|h|-1} + \frac{\sigma^2(\phi+\theta)^2\phi^{|h|}}{1-\phi^2} & |h| > 0 \end{cases}$$

- Filter method's *derivation* is messy
- Easier to compute than Y.W. for large lags
- In practice, without infinitely many past values, Filter method is an *approximation*

(The past is far behind us, but not that far)

Further reading: Great discussion in the intro to section 5.1 (Brocky & Davey)

Problem Setup | Review from Friday

Let **Y** be an RV with finite variance. We want to:

- 🕑 Predict **Y** based on some sequence of RVs, $\mathbf{W} = \{W_N, \ldots, W_1\}$
- 🕑 Find a function $\mathbf{g(W_1, \ldots, W_N)}$ that gives a "good" prediction of **Y**

We'll measure that "goodness" using **Mean Squared Error (MSE)**

$$MSE \stackrel{\text{def}}{=} \mathrm{E}\left[\left(Y - \underbrace{g(W_1, \ldots, W_N)}_{\text{prediction}}\right)^2\right]$$

Linearity of the Predictor | Review from Friday

Optimal Prediction:

Conditional expectation $\mathbf{E[Y \mid W_1, \ldots, W_N]}$

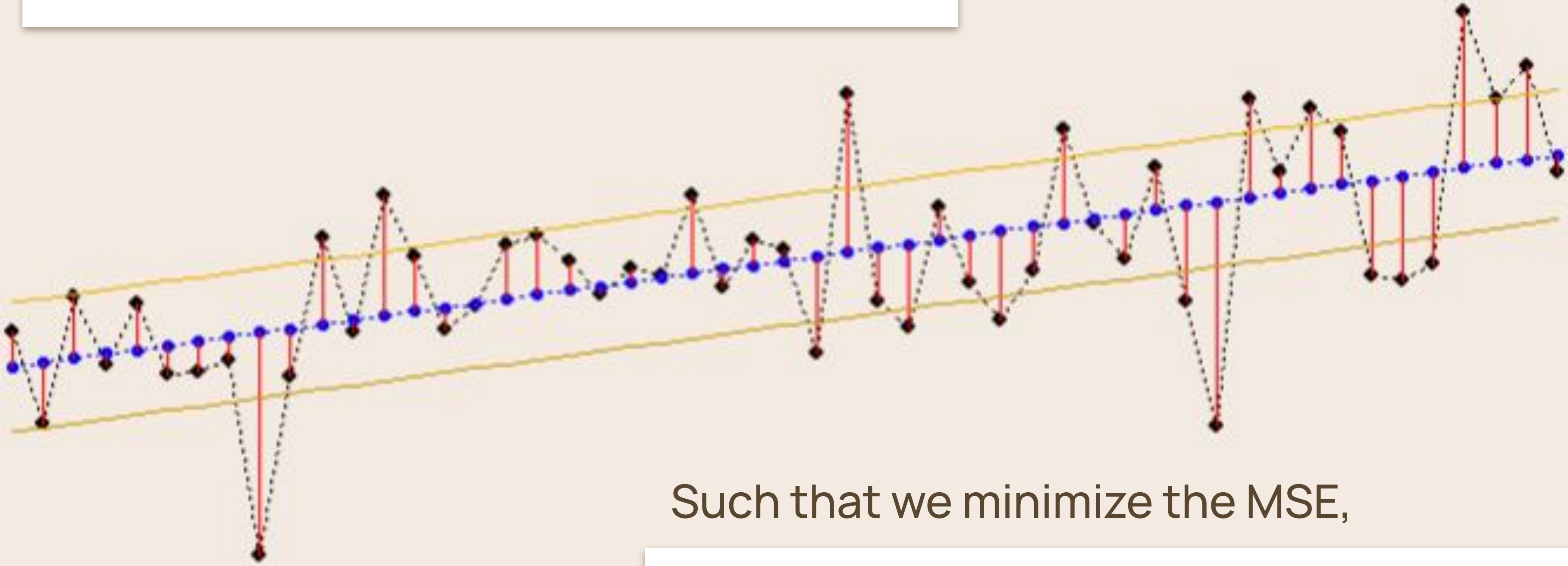
Not computable without joint distribution of Y and W

In time series, we typically only specify up to 2^{nd} order properties (Var/Cov)

If we specify the prediction to be a <u>linear</u> function **g** of **W**,
we can compute optimal prediction + MSE using only 2^{nd} order properties

We wish to find the best linear combination

Remember that "residual sum of squares" idea, from linear regression?



Such that we minimize the MSE,

Notation

Note that we reverse the time indices in **W**, by convention

- 🕑 In the spirit of filtering
- 🕑 convenient linear algebra, down the road

Minimizing the MSE:

$$\mathrm{E}\left[\left(Y-(a_0+\underbrace{a_1 W_N+\cdots+a_N W_1}_{\mathbf{a}^{\mathsf T}\mathbf{W}})\right)^2\right]$$

Assume we can differentiate MSE with respect to each a_i by taking the derivative inside the expectation

Set equal to 0 for all i,

solve for coefficients $\{a_i\}_{i=0}^N$

<u>i = 0</u> Divide both sides by **-2** → **E[Y-(a$_0$ + a$^\mathsf{T}$W)] = 0** → Use linearity of E!

Minimizing the MSE (i > 0)

$$0 = E\Big[-2\big(Y - (a_0 + a^T W)\big)W_{N+1-i}\Big]$$

Minimizing the MSE

$$\boxed{0 = \mathrm{E}\big[(Y-\mu_Y)W_{N+1-i}\big] - a^T\mathrm{E}\big[(W-\mu_W)W_{N+1-i}\big]}$$

$$0 = \mathrm{E}\big[(Y-\mu_Y)(W_{N+1-i}-\mu_{N+1-i})\big] - a^T\mathrm{E}\big[(W-\mu_W)(W_{N+1-i}-\mu_{N+1-i})\big]$$

Why can we subtract μ_{N+1-i}?

When you take the expectation through,

Then you're left with

This is , the i$^{\text{th}}$ component of the vector γ

This is the vector of covariances between W_{N+1-i} and components of W

ie) the i$^{\text{th}}$ row (or column) of

Minimizing the MSE

$$\boxed{0 = \mathrm{E}\left[(Y - \mu_Y)W_{N+1-i}\right] - a^T\mathrm{E}\left[(W - \mu_W)W_{N+1-i}\right]}$$

$$0 = \underline{\mathrm{E}\left[(Y - \mu_Y)(W_{N+1-i} - \mu_{N+1-i})\right]} - a^T\underline{\mathrm{E}\left[(W - \mu_W)(W_{N+1-i} - \mu_{N+1-i})\right]}$$

🕑 This is $\mathrm{Cov}(Y, W_{N+1-i})$, the ith component of the vector γ

🕑 This is the vector of covariances between W_{N+1-i} and components of W ie) the ith row (or column) of Γ

Minimum MSE

[warning: skye straight up didn't proof-read the LaTeX]

Conclusion

The best linear predictor of Y, based on W, that minimizes MSE is:

Where the vector a satisfies

Remark:

This solution is theoretical. Assumes we know $\mu_Y, \mu_W, \gamma,$ and Γ

In practice, we must estimate these quantities.