STAT 153 & 248 - Time Series
Lecture Twenty Three

Spring 2025, UC Berkeley

Aditya Guntuboyina

April 17, 2025

1 ARMA(p, q) Model

The ARMA(p, q) model is given by the equation:

$$(y_t - \mu) - \phi_1(y_{t-1} - \mu) - \cdots - \phi_p(y_{t-p} - \mu) = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}$$

where, as usual, $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. In backshift notation, this equation becomes

$$\phi(B)(y_t - \mu) = \theta(B)\epsilon_t \tag{1}$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA polynomials:

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad \text{and} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

applied to the backshift operator B. Another way of writing (1) is:

$$\phi(B)y_t = \delta + \theta(B)\epsilon_t,$$

where we now write the intercept term δ explicitly on the right hand side.

We can write the solution to (1) as

$$y_t - \mu = \frac{\theta(B)}{\phi(B)}\epsilon_t.$$

To make sense of the right hand side above, we can factorize $\phi(z)$ as:

$$\phi(z) = (1 - a_1 z)(1 - a_2 z)\ldots(1 - a_p z).$$

Here $1/a_1, \ldots, 1/a_p$ are the roots of $\phi(z)$. This gives

$$y_t - \mu = \frac{\theta(B)}{\prod_{k=1}^{p}(1 - a_k B)}\epsilon_t = \theta(B)(1 - a_1 B)^{-1}\ldots(1 - a_p B)^{-1}\epsilon_t$$

Each term $(1 - a_k B)^{-1}$ can be expanded via one of the following two formulae:

$$(1 - a_k B)^{-1} = \sum_{j=0}^{\infty} a_k^j B^j \quad \text{or} \quad (1 - a_k B)^{-1} = -\sum_{j=1}^{\infty} \frac{1}{a_k^j B^j}.$$

depending on whether $|a_k| < 1$ or $|a_k| > 1$. This allows us to write $y_t - \mu$ in terms of $\{\epsilon_t\}$. If $|a_k| < 1$ for every k, we can write

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

for some $\psi_0, \psi_1, \psi_2, \ldots$. This is a causal stationary process. We shall only work with ARMA(p, q) models in the causal stationary regime (which corresponds to $\phi(z)$ having all roots of modulus strictly larger than 1).

ARMA(p, q) is a more sophisticated model compared to pure AR(p) and MA(q). For AR(p), the theoretical PACF becomes zero for lags $h > p$. For MA(q), the theoretical ACF becomes zero for lags $h > q$. For ARMA(p, q) with both p and q at least one, one of these is true about the ACF and PACF. It is therefore to determine an appropriate choice for p and q for fitting an ARMA(p, q) model looking at the ACF and PACF. In practice, one usually searches over a range of p and q values using a model selection criterion such as AIC, BIC or Cross-Validation.

2 The Box-Jenkins Time Series Modeling Strategy

Box and Jenkins popularized the following strategy for modeling an observed time series $y_1, \ldots, y_n$:

1. Generally $y_1, \ldots, y_n$ will exhibit various kinds of trends. Preprocess the data to transform it to another series x_t which does not have any discernible trends.
2. Fit an ARMA(p, q) model for appropriate p and q to the transformed data x_t.

The preprocessing in the first step above is usually done in one of the following two ways:

1. **Differencing**. The first difference of $\{y_t\}$ is given by $\nabla y_t := y_t - y_{t-1}$ for $t = 2, \ldots, n$. The second difference is given by

$$\begin{aligned}\nabla^2 y_t &= \nabla(\nabla y_t) \\ &= \nabla(y_t - y_{t-1}) = \nabla y_t - \nabla y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}.\end{aligned}$$

   Higher order differences $\nabla^k y_t$ are defined recursively. Note that the length of the time series comes down after each successive differencing. For example, ∇y_t has length $n-1$, $\nabla^2 y_t$ has length $n-2$ and so on. Differencing usually eliminates increasing/decreasing trends. Usually one or two orders of differencing is enough to take care of increasing/decreasing trends.

2. **Seasonal Differencing**. Seasonal differencing is used to eliminate seasonal trends. Suppose we have a dataset having seasonal trends with period s (for example, for monthly datasets, $s = 12$). The seasonal first difference of y_t with period s is defined as

$$\nabla_s y_t := y_t - y_{t-s}$$

   Note that $\nabla_s y_t$ is a time series of length $n - s$. The second order seasonal difference is

$$\nabla_s^2 y_t = \nabla_s(\nabla_s y_t) = y_t - 2y_{t-s} + y_{t-2s}$$

   and higher order seasonal differences are defined recursively. Seasonal differences eliminate seasonal trends. Usually, in datasets having seasonal and increasing/decreasing

trends, one first takes a seasonal difference. This often eliminates seasonality and might also eliminate the linear trend. If a linear trend still persists, one takes a regular difference of the seasonal differenced series. This will often give a series with no trend and seasonality.

To the transformed data x_t, one fits an ARMA(p, q) model which can be done via the `ARIMA` function from the statsmodels library. The order p and q can be determined via a model selection criterion such as AIC or BIC.

3 ARIMA models

ARIMA stands for AutoRegressive Integrated Moving Average. ARIMA is essentially differencing plus ARMA.

Definition 3.1 (ARIMA)**.** *A time series model y_t is said to be ARIMA(p, d, q) if*

$$\phi(B)((\nabla^d y_t) - \mu) = \theta(B)\epsilon_t,$$

where $\epsilon_t \overset{texti.i.d}{\sim} N(0, \sigma^2)$.

ARIMA models are fit by the function `ARIMA()` in statismodels. The mean μ above is taken to be zero by default when the order parameter d in `ARIMA` is strictly larger than zero.

4 Seasonal ARMA Models

Seasonal ARMA models are often useful while modeling datasets having seasonal features (e.g., monthly datasets). We say that $\{y_t\}$ is a seasonal ARMA(P, Q) process with period s if it satisfies the difference equation $\Phi(B^s)(y_t - \mu) = \Theta(B^s)\epsilon_t$ where $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ and

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

and

$$\Theta(B^s) = 1 + \Theta_1 B^2 + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}.$$

The seasonal ARMA(P, Q) model with period s is a special case of an ARMA(Ps, Qs) model. However the seasonal model has $P + Q + 1$ (the 1 is for σ^2) parameters while a general ARMA(Ps, Qs) model will have $Ps + Qs + 1$ parameters. So the seasonal models are much sparser.

Causal stationary solution exists when every root of $\Phi(z^s)$ (equivalently, $\Phi(z)$) has modulus strictly larger than one.

The ACF and PACF of seasonal ARMA models are **non-zero** only at the seasonal lags $h = 0, s, 2s, 3s, \ldots$. At these seasonal lags, the ACF and PACF of these models behave just as the case of the unseasonal ARMA model: $\Phi(B)X_t = \Theta(B)\epsilon_t$.

5 Multiplicative Seasonal ARMA Models

For the `co2` dataset (from the time series analysis textbook by Cryer and Chan), for the first and seasonal differenced data, we saw that the sample autocorrelations seem nonnegligible

at lags 0, 1, 11, 12, 13 and those at all other lags seem negligible. This behaviour can be produced in a MA(13) model but that model will have 14 parameters possibly leading to overfitting.

We can get a much more parsimonious model for this dataset by *combining* the MA(1) model with a seasonal MA(1) model of period 12. Specifically, consider the model

$$y_t = (1 + \Theta B^{12})(1 + \theta B)\epsilon_t = \left(1 + \theta B + \Theta B^{12} + \theta\Theta B^{13}\right)\epsilon_t = \epsilon_t + \theta\epsilon_{t-1} + \Theta\epsilon_{t-12} + \theta\Theta\epsilon_{t-13}.$$

It is easy to check that model has the autocorrelation function:

$$\rho(1) = \frac{\theta}{1+\theta^2} \quad \text{and} \quad \rho(12) = \frac{\Theta}{1+\Theta^2}$$

and

$$\rho(11) = \rho(13) = \frac{\theta\Theta}{(1+\theta^2)(1+\Theta^2)}.$$

At every other lag $h > 0$, the autocorrelation $\rho_X(h)$ equals zero. Based on this ACF (and the sample ACF calculated from the data), this model can be suitable for the first and seasonal differenced data in the co2 dataset.

More generally, we can combine, by multiplication, ARMA and seasonal ARMA models to obtain models which have special autocorrelation properties with respect to seasonal lags. The **Multiplicative Seasonal Autoregressive Moving Average Model** ARMA(p, q) $\times$ $(P, Q)_s$ is defined via the difference equation:

$$\Phi(B^s)\phi(B)(y_t - \mu) = \Theta(B^s)\theta(B)\epsilon_t.$$

The model we looked at above for the co2 dataset is ARMA(0, 1) $\times$ $(0, 1)_{12}$.

Another example of a multiplicative seasonal ARMA model is ARMA(0, 1) $\times$ $(1, 0)_{12}$ (this is same as $MA(1) \times AR(1)_{12}$)

$$(y_t - \mu) - \Phi(y_{t-12} - \mu) = \epsilon_t + \theta\epsilon_{t-1}.$$

The autocorrelation function of this model can be checked to be $\rho(12h) = \Phi^h$ for $h \geq 0$ and

$$\rho(12h-1) = \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h \qquad \text{for } h = 0, 1, 2, \ldots$$

and $\rho(h) = 0$ at all other lags.

When we have a dataset whose ACF and PACF show interesting patterns at seasonal lags, consider using a multiplicative seasonal ARMA model. You may use the `Statsmodels` functions `arma_acf` and `arma_pacf` to understand the autocorrelation and partial autocorrelation functions of these models.

6 SARIMA Models

These models are obtained by combining differencing with multiplicative seasonal ARMA models. These models are denoted by ARIMA(p, d, q) $\times$ $(P, D, Q)_s$. This means that after differencing d times and seasonal differencing D times (with period s), we get a multiplicative seasonal ARMA model. In other words, $\{y_t\}$ is ARIMA(p, d, q) $\times$ $(P, D, Q)_s$ if it satisfies the difference equation:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d(y_t - \mu) = \delta + \Theta(B^s)\theta(B)\epsilon_t.$$

Recall that $\nabla_s^d = (1 - B^s)^d$ and $\nabla^d = (1 - B)^d$ denote the differencing operators.

In the co2 example, we wanted to use the model ARMA$(0, 1) \times (0, 1)_{12}$ to the seasonal and first differenced data: $\nabla\nabla_{12}X_t$. In other words, we want to fit the SARIMA model with nonseasonal orders $0, 1, 1$ and seasonal orders $0, 1, 1$ with seasonal period 12 to the original co2 dataset. This model can be fit to the data using the function `ARIMA` with the `seasonal_order` argument.

7 Parameter Estimation in MA(1)

Estimating the parameters of ARMA (as well as ARIMA, SARIMA models) is much harder than parameter estimation in AR models which was handled by standard regression (ordinary least squares). We will not study this topic (and simply rely on the `ARIMA` function for fitting these models to data). But here, I will just illustrate the difficulties involved in the simplest case of an non-AR model: MA(1). Recall that the MA(1) model is given by

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \tag{2}$$

where $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. The joint density of $y_1, \ldots, y_n$ is multivariate normal with mean vector $m := (\mu, \ldots, \mu)^T$ and covariance matrix Σ where Σ equals the $n \times n$ matrix whose $(i, j)^{th}$ entry is given by

$$\Sigma(i, j) = \begin{cases} \sigma^2\left(1 + \theta^2\right) & \text{when } i = j \\ \sigma^2\theta & \text{when } |i - j| = 1 \\ 0 & \text{for all other } (i, j) \end{cases}$$

The likelihood is therefore

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n (\det\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(y - m)'\Sigma^{-1}(y - m)\right)$$

where y is the $n \times 1$ vector with components $y_1, \ldots, y_n$. This is a function of the unknown parameters μ, θ, σ which can be estimated by maximizing the logarithm of the likelihood. The presence of Σ^{-1} makes this computationally expensive. Some (exact or approximate) formula should be used for Σ^{-1} so that one does not need to invert an $n \times n$ matrix every time the log-likelihood is to be computed.

An alternative approach is to try to write the likelihood (approximately) without using an explicit Σ^{-1}. One way of doing this is to use the connection to AR models. The MA(1) model (2) $y_t = \mu + \theta(B)\epsilon_t$ (with $\theta(B) = 1 + \theta B$) can be converted to an AR model as follows:

$$\epsilon_t = \frac{1}{\theta(B)}(y_t - \mu) = \frac{1}{1 + \theta B}(y_t - \mu) = \left(1 - \theta B + \theta^2 B^2 - \theta^3 B^3 + \ldots\right)(y_t - \mu)$$

so that

$$y_t - \theta y_{t-1} + \theta^2 y_{t-2} - \theta^3 y_{t-3} + \cdots = \frac{\mu}{1 + \theta} + \epsilon_t$$

This requires the assumption that $|\theta| < 1$. For this AR model, we can write the likelihood:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(y_t - \frac{\mu}{1 + \theta} - \theta y_{t-1} + \theta^2 y_{t-2} - \theta^3 y_{t-3} + \ldots\right)^2\right).$$

This formula involves $y_0, y_{-1}, y_{-2}, \ldots$ for which we have no data. We can deal with them by simply setting them to be zero (you can think of writing the conditional likelihood of the data $y_1, \ldots, y_n$ given $y_0, y_{-1}, y_{-2}, \ldots$ as zero). The likelihood then becomes:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{S(\mu, \theta)}{2\sigma^2}\right)$$

where

$$S(\mu,\theta) = \left(y_1 - \frac{\mu}{1+\theta}\right)^2 + \left(y_2 - \frac{\mu}{1+\theta} - \theta y_1\right)^2 + \left(y_3 - \frac{\mu}{1+\theta} - \theta y_2 + \theta^2 y_1\right)^2 + \cdots + \left(y_n - \frac{\mu}{1+\theta} - \theta y_{n-1} + \theta^2 y_{n-2} - \cdots + (-1)^{n-1}\theta^{n-1} y_1\right)^2.$$

The MLEs of μ and θ are obtained by minimizing $S(\mu,\theta)$:

$$\hat{\mu}, \hat{\theta} \text{ minimize } S(\mu,\theta).$$

This is a nonlinear minimization that can be done via some optimization routines in Python (say in `scipy`). The MLE for σ is easily seen to be

$$\hat{\sigma} = \frac{S(\hat{\mu},\hat{\theta})}{n}.$$

For uncertainty quantification, we can take a Bayesian approach and combine the likelihood with a prior on θ, μ, σ. Here is how this is done. **I did not cover the following in lecture, and this material is optional. It is included here just for completeness**.

We assume that θ, μ, σ are independent with:

$$\theta \sim \text{Unif}(-1,1) \quad \mu \sim \text{Unif}(-C,C) \quad \log\sigma \sim \text{Unif}(-C,C)$$

for a large $C \to \infty$. Note that we have restricted the range of θ to $(-1,1)$ because we assumed that $|\theta| < 1$. The posterior is then

$$\begin{aligned} f_{\mu,\theta,\sigma|\text{data}}(\mu,\theta,\sigma) &\propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{S(\mu,\theta)}{2\sigma^2}\right) \times \frac{1}{\sigma} I\{-1<\theta<1, -C<\mu, \log\sigma<C\} \\ &\propto \sigma^{-n-1} \exp\left(-\frac{S(\mu,\theta)}{2\sigma^2}\right) I\{-1<\theta<1, -C<\mu, \log\sigma<C\}. \end{aligned}$$

To obtain the posterior of μ and θ alone, we integrate the above with respect to σ. Integrating from 0 to ∞ (assuming C is large so $e^{-C} \approx 0$ and $e^C \approx \infty$), we obtain (as in Lecture Three):

$$f_{\mu,\theta|\text{data}}(\mu,\theta) \propto \left(\frac{1}{S(\mu,\theta)}\right)^{n/2} I\{-1<\theta<1, -C<\mu<C\}.$$

This posterior can be evaluated numerically over a grid of values of μ and θ and approximated by the appropriate discrete distribution over the grid. Alternatively, we can approximate this posterior by a suitable t-distribution by doing a Taylor expansion of $S(\mu,\theta)$ near the minimizer $\hat{\mu}, \hat{\theta}$. To illustrate this, let $\alpha = (\mu,\theta)$ and $\hat{\alpha} = (\hat{\mu},\hat{\theta})$. Taylor expansion for α near $\hat{\alpha}$ gives

$$\begin{aligned} S(\alpha) &= S(\hat{\alpha}) + \langle \nabla S(\hat{\alpha}), \alpha - \hat{\alpha}\rangle + (\alpha-\hat{\alpha})^T \left(\frac{1}{2} HS(\hat{\alpha})\right)(\alpha-\hat{\alpha}) \\ &= S(\hat{\alpha}) + (\alpha-\hat{\alpha})^T \left(\frac{1}{2} HS(\hat{\alpha})\right)(\alpha-\hat{\alpha}) \end{aligned}$$

where we used $\nabla S(\hat{\alpha}) = 0$ because $\hat{\alpha}$ minimizes $S(\alpha)$. Here $HS(\hat{\alpha})$ denotes the Hessian of

S at $\hat{\alpha}$. Therefore

$$
\begin{aligned}
& f_{\mu,\theta|\text{data}}(\mu,\theta) \\
& \propto \left(\frac{1}{S(\mu,\theta)}\right)^{n/2} I\{-1<\theta<1,-C<\mu<C\} \\
& \propto \left(\frac{S(\hat{\alpha})}{S(\alpha)}\right)^{n/2} I\{-1<\theta<1,-C<\mu<C\} \\
& = \left(\frac{S(\hat{\alpha})}{S(\hat{\alpha})+(\alpha-\hat{\alpha})^T\left(\frac{1}{2}HS(\hat{\alpha})\right)(\alpha-\hat{\alpha})}\right)^{n/2} I\{-1<\theta<1,-C<\mu<C\} \\
& = \left(\frac{1}{1+(\alpha-\hat{\alpha})^T\left(\frac{1}{2S(\hat{\alpha})}HS(\hat{\alpha})\right)(\alpha-\hat{\alpha})}\right)^{n/2} I\{-1<\theta<1,-C<\mu<C\} \\
& = \left(\frac{1}{1+\frac{1}{n-2}(\alpha-\hat{\alpha})^T\left(\frac{n-2}{2S(\hat{\alpha})}HS(\hat{\alpha})\right)(\alpha-\hat{\alpha})}\right)^{\frac{n-2+2}{2}} I\{-1<\theta<1,-C<\mu<C\}.
\end{aligned}
$$

Comparing the above with the formula:

$$\left(\frac{1}{1+\frac{1}{k}(x-m)^T\Sigma^{-1}(x-m)}\right)^{\frac{k+p}{2}}$$

for the p-variate t-density $t_{k,p}(\mu,\Sigma)$, we see that (ignoring the indicator function $I\{-1<\theta<1,-C<\mu<C\}$)

$$\alpha \mid \text{data} \sim t_{n-2,2}\left(\hat{\alpha}, \frac{S(\hat{\alpha})}{n-2}\left(\frac{1}{2}HS(\hat{\alpha})\right)^{-1}\right).$$

This t-density can be used for uncertainty quantification of μ and θ.

8 Additional Optional Reading

1. Sections 3.5, 3.6 and 3.9 of Shumway-Stoffer 4th edition.