STAT 153 - Introduction to Time Series Lecture Nineteen

Fall 2022, UC Berkeley

Aditya Guntuboyina

November 2, 2022

1 Last Class: Inference for $MA(1)$

In the last class, we studied inference in the $MA(1)$ model:

$$Y_t = \mu + Z_t + \theta Z_{t-1} \qquad \text{where } Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

This model has three parameters: $\mu.\theta, \sigma$, and we discussed how to estimate these parameters along with standard errors from observed data $y_1, \dots, y_n$. The first method is to simply to write down the likelihood of $Y_1, \dots, Y_n$ which is multivariate normal with mean vector $(\mu, \dots, \mu)^T$, and covariance matrix having the quantity $(1+\theta^2)\sigma^2$ on the diagonal and $\theta\sigma^2$ above and below the diagonal, and zero otherwise. We can thus write down the likelihood using the multivariate normal density. This likelihood involves the inversion of the $p \times p$ covariance matrix so can be computationally intensive. We can obtain MLEs for μ, θ, σ by numerically maximizing the likelihood.

From the form of the likelihood, it is clear that there is an identifiability issue with the MA(1) model. Specifically, the likelihood is identical for the two distinct parameter choices (μ, θ, σ) and $(\mu, \theta^{-1}, |\theta|\sigma)$. This motivates imposing the restriction $|\theta| < 1$.

The second method uses a simplfying assumption which makes the likelihood much easier to write down. Specifically, we assume that $Z_0 = 0$ (as opposed to $Z_0 \sim N(0, \sigma^2)$). This allows writing each $Z_1, \dots, Z_n$ in terms of $Y_1, \dots, Y_n$ and the parameters θ, μ. Basically, we recurse

$$Z_t = Y_t - \mu - \theta Z_{t-1}$$

for $t = 2, 3, \dots$ with the initialization $Z_1 = Y_1 - \mu$. We then substitute the observed data $y_1, \dots, y_n$ for $Y_1, \dots, Y_n$ and denote the resulting values of $Z_1, \dots, Z_n$ by $\hat{Z}_1(\mu, \theta), \dots, \hat{Z}_n(\mu, \theta)$. Specifically,

$$\begin{aligned} \hat{Z}_1(\mu, \theta) &= y_1 - \mu \\ \hat{Z}_2(\mu, \theta) &= y_2 - \mu - \theta \hat{Z}_1 \\ \hat{Z}_3(\mu, \theta) &= y_3 - \mu - \theta \hat{Z}_2 \\ \dots \end{aligned}$$

The approximate (conditional) likelihood is then given by

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{S(\mu, \theta)}{2\sigma^2}\right) \qquad \text{where } S(\mu, \theta) := \sum_{t=1}^n \hat{Z}_t(\mu, \theta)^2.$$

The quantity $S(\mu, \theta)$ is referred to as "Conditional Sum of Squares". Minimizing it leads to estimates of μ and θ; these estimates should be quite close to the Maximum Likelihood Estimates. One can obtain standard errors by taking the square roots of the diagonal entries of the matrix:

$$\frac{S(\hat{\mu}, \hat{\theta})}{n-2} \left(\frac{1}{2} HS(\hat{\mu}, \hat{\theta}) \right)^{-1}.$$

This is because, as we argued in the last class,

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

2 More Comments on $MA(1)$

2.1 Invertibility

The $MA(1)$ model in the regime $|\theta| < 1$ is known as **invertible**. This is because, under the assumption $|\theta| < 1$, we can write Z_t in terms of $Y_t, Y_{t-1}, Y_{t-2}, \dots$ which means the model can be inverted and the shocks Z_t can be written back in terms of $Y_t, Y_{t-1}, \dots$. To see why Z_t can be written in terms of $Y_t, Y_{t-1}, Y_{t-2}, \dots$, just note that

$$Y_t - \mu = \theta(B) Z_t \qquad \text{where } \theta(B) = I + \theta B$$

so that

$$\begin{aligned} Z_t &= \frac{1}{\theta(B)} (Y_t - \mu) \\ &= (I + \theta B)^{-1} (Y_t - \mu) \\ &= \left(I - \theta B + \theta^2 B^2 - \theta^3 B^3 + \dots \right) (Y_t - \mu) \\ &= (Y_t - \mu) - \theta (Y_{t-1} - \mu) + \theta^2 (Y_{t-2} - \mu) - \dots \end{aligned}$$

and the infinite sum above makes sense because of $|\theta| < 1$.

2.2 Predictions

Forecasting with $MA(1)$ is done in the following way. We shall assume that $Z_0 = 0$ for simplicity. To forecast Y_{n+1}, we shall consider the distribution:

$$Y_{n+1} \mid Y_1 = y_1, \dots, Y_n = y_n, \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma}. \tag{1}$$

Note that we are conditioning on the parameters being equal to their point estimates. Strictly speaking, we need to average with respect to the posterior distribution of the parameters but fixing them at their point estimates is okay as the posterior distribution is, in general, quite concentrated.

Under the assumption $Z_0 = 0$, the conditioning event

$$Y_1 = y_1, \dots, Y_n = y_n, \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma}$$

is the same as

$$Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma}$$

where $\hat{Z}_i(\mu, \theta)$ is defined in the previous section. Therefore, the conditional distribution (1) is the same as

$$\begin{aligned}
&Y_{n+1} \mid Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&= \hat{\mu} + Z_{n+1} + \hat{\theta} Z_n \mid Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&= \hat{\mu} + Z_{n+1} + \hat{\theta} \hat{Z}_n(\hat{\mu}, \hat{\theta}) \mid Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&\overset{d}{=} \hat{\mu} + Z_{n+1} + \hat{\theta} \hat{Z}_n(\hat{\mu}, \hat{\theta}) \mid \mu = \hat{\mu}, \sigma = \hat{\sigma} \sim N(\hat{\mu} + \hat{\theta} \hat{Z}_n, \hat{\sigma}^2).
\end{aligned}$$

Our point prediction for Y_{n+1} is therefore $\hat{\mu} + \hat{\theta} \hat{Z}_n$ and the prediction standard error is $\hat{\sigma}$.

To predict Y_{n+i} for $i \geq 2$, we observe

$$\begin{aligned}
&Y_{n+i} \mid Y_1 = y_1, \dots, Y_n = y_n, \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&= Y_{n+i} \mid Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&= \hat{\mu} + Z_{n+i} + \hat{\theta} Z_{n+i-1} \mid Z_1 = \hat{Z}_1(\hat{\mu}, \hat{\theta}), \dots, Z_n = \hat{Z}_n(\hat{\mu}, \hat{\theta}), \theta = \hat{\theta}, \mu = \hat{\mu}, \sigma = \hat{\sigma} \\
&\overset{d}{=} \hat{\mu} + Z_{n+i} + \hat{\theta} Z_{n+i-1} \mid \sigma = \hat{\sigma} \sim N\left(\hat{\mu}, \hat{\sigma}^2(1 + \hat{\theta}^2)\right).
\end{aligned}$$

The predictions for the $MA(1)$ model are therefore quite simple. The prediction for the next time point is $\hat{\mu} + \hat{\theta} \hat{Z}_n$ with standard error $\hat{\sigma}$. The predictions for all other time points are just the mean estimate $\hat{\mu}$ with standard error $\hat{\sigma}\sqrt{1 + \hat{\theta}^2}$.

2.3 Initialization

In order to obtain the MLEs or the minimizers of the sum of squares $S(\mu, \theta)$, we need to use a numerical optimization routine (say the function `optim` in `R`). Such routine are usually iterative and they require specification of initial values. One can take random initialization or some specific value such as $\mu = 0, \theta = 0$. Sometimes, it makes sense to use more clever choices for the initial values. For μ, it makes sense to take $\hat{\mu}_{init} = \bar{y} = (y_1 + \cdots + y_n)/n$. For θ, one can use the fact that the autocorrelation function of $MA(1)$ at lag one equals $\theta/(1 + \theta^2)$. Thus we can use

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

where $\hat{\rho}(1)$ is the sample autocorrelation at lag one. We can solve the equation above (note we work under the assumption $|\theta| < 1$) to obtain

$$\hat{\theta}_{init} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}.$$

Observe that this only works when $\hat{\rho}(1) < 0.5$ (if $\hat{\rho}(1)$ exceeds 0.5, then there is no solution to (2)).

3 Differencing

The MA(q) model is stationary (i.e., the mean $\mathbb{E}(Y_t)$ is constant over time and the covariance $\mathrm{Cov}(Y_{t_1}, Y_{t_2})$ depends only on $|t_1 - t_2|$). Observed time series data usually have levels that change over time so the MA(q) model would not directly be applicable. Thus, before fitting the MA(q) model, a preprocessing step is performed where the data is **differenced**. More precisely, instead of working with the raw data $\{x_t\}$, one looks at $y_t = x_t - x_{t-1}, t = 2, \dots, n$.

Quite often (and this can be empirically checked), the differenced observations have levels that can be treated as constant. Sometimes, even after differencing, one can notice a trend in the data. In that case, just difference again.

Note that predictions for the original time series $\{x_t\}$ can be obtained from predictions for the differenced time series $\{y_t\}$. Indeed, suppose we fit the $MA(q)$ model to the differenced time series y_t leading to predictions $y_{n+1}, y_{n+2}, \dots$. Then the predictions for the original time series $x_{n+1}, x_{n+2}, \dots$ can be obtained as follows:

$$\begin{aligned}\text{prediction for } x_{n+1} &= (\text{observed value of } x_n) + (\text{prediction for } x_{n+1} - x_n)\\ &= (\text{observed value of } x_n) + y_{n+1}.\end{aligned}$$

Further

$$\begin{aligned}\text{prediction for } x_{n+2} &= (\text{prediction for } x_{n+1}) + (\text{prediction for } x_{n+2} - x_{n+1})\\ &= (\text{prediction for } x_{n+1}) + y_{n+2}\\ &= (\text{observed value of } x_n) + y_{n+1} + y_{n+2}\end{aligned}$$

and so on leading to

$$\text{prediction for } x_{n+k} = (\text{observed value of } x_n) + y_{n+1} + y_{n+2} + \dots + y_{n+k}.$$

The following strategy is often adopted while working with $MA(q)$ models (and also if one only wants to work with causal stationary $AR(p)$ models):

1. Difference the original time series (once or twice or, even, thrice) to obtain data whose levels seem to be constant across time.

2. Fit the $MA(q)$ model to the differenced dataset.

3. Rewrite the model in terms of the original time series. For example, if we fitted the model
$$Y_t = \hat{\mu} + Z_t + \hat{\theta} Z_{t-1}$$
to the once difference time series $Y_t = X_t - X_{t-1}$, then the model in terms of the original data would be
$$X_t = X_{t-1} + \hat{\mu} + Z_t + \hat{\theta} Z_{t-1}. \tag{3}$$

4. If predictions are desired, first obtain predictions for $\{Y_t\}$ and then convert them into predictions for $\{X_t\}$ (alternatively, obtain predictions directly using the model equation (3)).

4 ACF Diagnostic for $MA(q)$

While fitting $MA(q)$ models to data, one needs to determine a suitable choice of the order q. One heuristic way of doing this is to look at the Sample AutoCorrelation Function (ACF) of the observed time series data.

The theoretical AutoCorrelation Function of the MA(q) model equals zero for all lags that are strictly larger than q. To see this, note first that the definition of theoretical AutoCorrelation for a stationary time series model is given by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

where $\gamma(h)$ is the AutoCovariance Function defined as $\gamma(h) = \text{Cov}(Y_t, Y_{t+h})$ (note that stationarity implies that $\gamma(h)$ only depends on h and not on t). For $MA(q)$, we saw in the Lecture 17 that $\gamma(h)$ equals zero for $h > q$ which implies that $\rho(h)$ also equals zero for $h > q$.

Based on this observation, if the sample acf exhibits the property that the values for lags larger than the particular lag q are relatively small, then we can try fitting the $MA(q)$ model to the data. As a concrete example, consider the US GDP growth rate time series data (this is the percent change in the GDP each year compared to the previous year). The sample ACF for this dataset has a significant valueu at lag 1 but is fairly small for larger lags. It makes sense therefore to use the $MA(1)$ model for this dataset.

This ACF diagnostic cannot be used for figuring the order p for fitting the AR(p) model. This is because the theoretical ACF for $AR(p)$ does not sharply cut off after a certain lag but instead exhibits a tapering-off behavior. For example, the theoretical ACF for $AR(1)$ equals $\phi^{|h|}$ which is strictly non-zero for every h. A better diagnostic for AR order determination is given by the PACF (Partial Autocorrelation Function) which we shall discuss in the next class.

5 Parameter Estimation for $MA(q)$

Inference for $MA(q)$ for $q \geq 1$ is similar to that for $q = 1$. Before discussing the details, let us first define the notion of “invertibility” for $MA(q)$ for $q \geq 1$.

5.1 Invertibility for $MA(q)$

Invertibility is defined in terms of the MA polynomial:

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q.$$

We say that the $MA(q)$ model with parameters $\theta_1, \dots, \theta_q$ is invertible if all q roots of the the MA polynomial $\theta(z)$ have modulus strictly larger than 1. In other words, we can represent

$$\theta(z) = (1 - \lambda_1 z)(1 - \lambda_2 z) \dots (1 - \lambda_q z)$$

with $|\lambda_i| < 1$ for each $i = 1, \dots, q$. For invertible $MA(q)$ models, we can represent Z_t as an infinite linear combination of $Y_t, Y_{t-1}, Y_{t-2}, \dots$ in the following way. First note that

$$Y_t - \mu = \theta(B) Z_t$$

so that

$$\begin{aligned}
Z_t &= \frac{1}{\theta(B)}(Y_t - \mu) \\
&= \frac{1}{(I - \lambda_1 B)(I - \lambda_2 B) \dots (I - \lambda_q B)}(Y_t - \mu) \\
&= (I - \lambda_1 B)^{-1}(I - \lambda_2 B)^{-1} \dots (I - \lambda_q B)^{-1}(Y_t - \mu) \\
&= \left(I + \lambda_1 B + \lambda_1^2 B^2 + \dots\right) \dots \left(I + \lambda_q B + \lambda_q^2 B^2 + \dots\right)(Y_t - \mu).
\end{aligned}$$

By multiplying out all the terms above, we obtain a representation of Z_t in terms of $Y_t, Y_{t-1}, \dots$. The coefficients of this representation will involve higher powers of λ_j and their products over j. Because each $|\lambda_j| < 1$, the resulting infinite sum will be meaningful (convergent).

While fitting $MA(q)$ models to data, we shall restrict attention to parameter values which lead to invertibility. This is because for every non-invertible $MA(q)$ model, there exists an invertible $MA(q)$ model which has the same mean and autocovariance function. The idea behind the construction of the equivalent invertible model is the following. Write the original $MA(q)$ model as

$$Y_t - \mu = (I - \lambda_1 B)(I - \lambda_2 B)\dots(I - \lambda_q B)Z_t$$

The roots of the MA polynomial are $\lambda_1^{-1}, \dots, \lambda_q^{-1}$. Suppose that this is not an invertible model so that some of the λ_i's have modulus strictly larger than 1. For concreteness, assume that $|\lambda_i| < 1$ for $i = 1, \dots, s$ and $|\lambda_i| > 1$ for $i = s+1, \dots, q$. We can then define the alternative $MA(q)$ model:

$$\tilde{Y}_t - \mu = (I - \lambda_1 B)\dots(I - \lambda_s B)(I - \lambda_{s+1}^{-1} B)\dots(I - \lambda_q^{-1} B)\tilde{Z}_t$$

where $\tilde{Z}_t, t = \dots, -3, -2, -1, 0, 1, 2, 3, \dots$ are i.i.d $N(0, \tilde{\sigma}^2)$ where

$$\tilde{\sigma}^2 = \sigma^2 \lambda_{s+1}^2 \lambda_{s+2}^2 \dots \lambda_q^2.$$

Clearly $\tilde{Y}_t$ is an invertible MA(q) model because the roots of its MA polynomial are $\lambda_1^{-1}, \dots, \lambda_s^{-1}$ which have modulus greater than 1 as well as $\lambda_{s+1}, \dots, \lambda_q$ which also have modulus greater than 1. It turns out that $\tilde{Y}_t$ and Y_t have the same mean (μ) as well as the same AutoCovariance Function. For a proof of this, see Section 3.7 of the book *Time Series Analysis* by James Douglas Hamilton.

If there are some roots with modulus exactly equal to one, then we cannot represent the process with an equivalent invertible MA process. We shall usually rule out MA processes having any root with magnitude exactly equal to one.

From now on, while fitting $MA(q)$ models, we shall assume invertibility i.e., that all roots of the MA polynomial have magnitude strictly larger than 1.

5.2 Parameter Estimation for $MA(q)$

One can either directly write down the likelihood and attempt to maximize over $\mu, \theta_1, \dots, \theta_q, \sigma$. The likelihood is given by

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

where y is the $n \times 1$ vector with components $y_1, \dots, y_n$, m is the $n \times 1$ vector $m := (\mu, \dots, \mu)$, and Σ is the covariance matrix given by

$$\Sigma(i,j) = \text{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2 \sum_{l=0}^{q-|i-j|} \theta_l \theta_{l+|i-j|} & \text{for } 0 \le |i-j| \le q \\ 0 & \text{for } h > q \end{cases}$$

As in the last lecture where we looked at the case $q = 1$, we can maximize the likelihood by first fixing $\theta_1, \dots, \theta_q, \mu$, and then solving the one-dimensional optimization problem over σ in closed form. Then fix $\theta_1, \dots, \theta_q$ and solve the one dimensional optimization problem over μ in closed form. Finally, use a numerical routine to maximize over $\theta_1, \dots, \theta_q$.

The conditional sum of squares method works as following. First initialize $Z_0 = 0, Z_{-1} = 0, \dots, Z_{1-q} = 0$. Then we recursively write $Z_1, \dots, Z_n$ in terms of $Y_1, \dots, Y_n$ using:

$$Z_t = Y_t - \mu - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - \dots - \theta_q Z_{t-q}$$

for $t = 1, 2, \dots, n$. Plug in the observed values $y_1, \dots, y_n$ for $Y_1, \dots, Y_n$ to obtain the values $\hat{Z}_1, \dots, \hat{Z}_n$ of $Z_1, \dots, Z_n$. Note that $\hat{Z}_1, \dots, \hat{Z}_n$ depend on the parameters $\theta_1, \dots, \theta_q, \mu$. The approximate (conditional) likelihood is then given by

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{S(\mu, \theta_1, \dots, \theta_q)}{2\sigma^2}\right) \qquad \text{where } S(\mu, \theta_1, \dots, \theta_q) := \sum_{t=1}^{n} \hat{Z}_t(\mu, \theta_1, \dots, \theta_q)^2.$$

$S(\mu, \theta_1, \dots, \theta_q)$ is referred to as "Conditional Sum of Squares". Minimizing it leads to estimates of μ and θ; these estimates should be quite close to the Maximum Likelihood Estimates. Similar to the $q = 1$ case, we can prove that

$$(\mu, \theta_1, \dots, \theta_q) \mid \text{data} \sim t_{n-q-1,q+1}\left((\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q), \frac{S(\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q)}{n-q-1}\left(\frac{1}{2}HS(\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q)\right)^{-1}\right).$$

Using this result, we can obtain standard errors corresponding to the estimates $\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q$ by taking the square roots of the diagonal entries of the matrix:

$$\frac{S(\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q)}{n-q-1}\left(\frac{1}{2}HS(\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q)\right)^{-1}.$$

The in-build function `arima` in `R` can also be used to fit MA (and AR) models. The method described above leads to answers which are basically the same as the ones output by this function.

6 Recommended Reading for Today

1. The ACF diagnostic for finding the MA order is described in Section 3.3 of the Shumway-Stoffer book.
2. More details on invertibility of MA models can be found in Section 3.3 of the Shumway-Stoffer book or Section 3.7 of the Hamilton book.
3. For more on Forecasting, read Section 3.4 of the Shumway-Stoffer book.
4. The initialization procedure for MA(1) can be found in Example 3.29 of the Shumway-Stoffer book.
5. For more on parameter estimation in MA models, read Section 3.5 of Shumway-Stoffer.