STAT 153 - Introduction to Time Series Lecture Eighteen

Fall 2022, UC Berkeley

Aditya Guntuboyina

October 29, 2022

1 Moving Average (MA) models

The $MA(q)$ (Moving Average Model of order q) model is given by

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

where $Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. This model has $q+2$ parameters: μ, $\theta_1, \ldots, \theta_q$ and σ^2. This model has been called the "Summation of Random Causes" by its inventor Slutzky in the original paper titled "The summation of random causes as the source of cyclic processes" published in Econometrica in 1937. Basically the Z_t's can be treated as random causes which are assumed to be independently and identically distributed. The actual observations Y_t's are consequences of these causes. The consequence for time t depends on the cause for time t as well as the causes for times $t-1, \ldots, t-q$. These different causes affect the consequence at time t differently depending on the values of $\psi_1, \ldots, \psi_q$. Note that successive observations Y_t share some common causes leading to dependence between the successive values of Y_t. To determine the exact form of the dependence between different Y_t's, we can calculate the covariance function as:

$$\text{Cov}\left(Y_t, Y_{t+h}\right) = \begin{cases} \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & \text{for } 0 \le h \le q \\ 0 & \text{for } h > q \end{cases}$$

where $\theta_0 = 1$. Note that the covariance function equals zero for $h > q$. This is often used as a diagnostic tool for deciding whether to fit MA models and to decide the value of q: one plots the sample autocorrelation function of the data and checks if it is negligible after a certain lag q.

2 Parameter Estimation for $MA(q)$ models

Given data $y_1, \ldots, y_n$, how do we estimate the parameters of the $MA(q)$ model?

2.1 Maximum Likelihood Estimation

Here we write down the likelihood and then maximize it to obtain the maximum likelihood estimates. 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

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

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 $y_1, \ldots, y_n$. We shall now discuss the problem of optimizing this with respect to the parameters $\mu, \theta_1, \ldots, \theta_q, \sigma$.

Today we shall consider the simple case $q = 1$: $Y_t = \mu + Z_t + \theta Z_{t-1}$. The case of higher values of q will be looked at next week. When $q = 1$, Σ equals the $n \times n$ matrix whose $(i,j)^{th}$ entry is given by

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

Let us write $\Sigma = \sigma^2\Gamma$ where Γ equals the $n \times n$ matrix whose $(i,j)^{th}$ entry is given by

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

Note that Γ depends on the parameter θ (sometimes we shall write $\Gamma(\theta)$ to emphasize the dependence on θ). The likelihood for $MA(1)$ is then:

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

The log-likelihood is

$$\ell(\theta, \mu, \sigma) := -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\log(\det\Gamma) - \frac{(y-m)'\Gamma^{-1}(y-m)}{2\sigma^2}.$$

The maximizer of the above function over θ, σ, μ cannot be written in closed form. Because

$$\max_{\theta,\mu,\sigma} \ell(\theta,\mu,\sigma) = \max_{\theta,\mu}\left[\max_{\sigma}\ell(\theta,\mu,\sigma)\right],$$

we can maximize $\ell(\theta,\mu,\sigma)$ by first maximizing over σ for fixed θ and μ. Differentiating $\ell(\theta,\mu,\sigma)$ with respect to σ, we get (below $\Gamma^{-1}(i,j)$ denotes the $(i,j)^{th}$ entry of Γ^{-1})

$$\frac{\partial \ell(\theta,\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{(y-m)'\Gamma^{-1}(y-m)}{\sigma^3}$$

Setting this partial derivative to zero, we get

$$\sigma = \sigma(\theta,\mu) = \sqrt{\frac{(y-m)'\Gamma^{-1}(y-m)}{n}} \tag{1}$$

Note that $m = (\mu, \ldots, \mu)^T$ and Γ depends on θ so the right hand above depends on μ and θ. Thus

$$\max_{\theta,\mu,\sigma}\ell(\theta,\mu,\sigma) = \max_{\theta,\mu} F(\theta,\mu)$$

where

$$\begin{aligned} F(\theta,\mu) &:= \ell(\theta,\mu,\sigma(\theta,\mu)) \\ &= -\frac{n}{2}\log(2\pi) - n\log\left(\sigma(\theta,\mu)\right) - \frac{1}{2}\log\left(\det\Gamma\right) - \frac{(y-m)'\Gamma^{-1}(y-m)}{2\left(\sigma(\theta,\mu)\right)^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{(y-m)'\Gamma^{-1}(y-m)}{n} - \frac{1}{2}\log\left(\det\Gamma\right) - \frac{n}{2}. \end{aligned}$$

In order to maximize $F(\theta,\mu)$ over θ and μ, we shall first maximize over μ for fixed θ because

$$\max_{\theta,\mu} F(\theta,\mu) = \max_{\theta}\left(\max_{\mu} F(\theta,\mu)\right)$$

Maximizing $F(\theta,\mu)$ over μ for fixed θ is equivalent to minimizing $(y-m)'\Gamma^{-1}(y-m)$ over μ. The derivative of this with respect to μ equals:

$$\begin{aligned} \frac{\partial}{\partial\mu}(y-m)'\Gamma^{-1}(y-m) &= \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i-\mu)\Gamma^{-1}(i,j)(y_j-\mu)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}(2\mu - y_i - y_j)\Gamma^{-1}(i,j) \\ &= 2\mu\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma^{-1}(i,j) - \sum_{i=1}^{n}\sum_{j=1}^{n}y_i\Gamma^{-1}(i,j) - \sum_{i=1}^{n}\sum_{j=1}^{n}y_j\Gamma^{-1}(i,j) \\ &= 2\mu\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma^{-1}(i,j) - 2\sum_{i=1}^{n}\sum_{j=1}^{n}y_i\Gamma^{-1}(i,j). \end{aligned}$$

Setting this derivative to zero, we get

$$\mu = \mu(\theta) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}y_i\Gamma^{-1}(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma^{-1}(i,j)} \tag{2}$$

which is the best value for μ for fixed θ (note the right hand side above depends on θ). Thus

$$\max_{\theta,\mu,\sigma}\ell(\theta,\mu,\sigma) = \max_{\theta}G(\theta)$$

where

$$G(\theta) := -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{(y-m(\theta))'\Gamma^{-1}(y-m(\theta))}{n} - \frac{1}{2}\log\left(\det\Gamma\right) - \frac{n}{2}$$

where $m(\theta)$ is the $n\times 1$ vector with each entry equal to $\mu(\theta)$. Therefore our strategy to obtain the maximizer $\hat{\theta}, \hat{mu}, \hat{\sigma}$ of $\ell(\theta,\mu,\sigma)$ is following:

1. First obtain the maximizer $\hat{\theta}$ of $G(\theta)$ over θ. This can be done via grid maximization (note that for $MA(1)$, the parameter θ is one-dimensional) or by in-built optimization functions such as `optim` in R.
2. Calculate $\hat{\mu}$ as $\mu(\hat{\theta})$ using the formula (2).
3. Calculate $\hat{\sigma}$ as $\sigma(\hat{\theta},\hat{\mu})$ using the formula (1).

2.2 Parameter Nonidentifiability in $MA(1)$

The covariance matrix Σ for the MA(1) model depends on θ and σ so we denote it by $\Sigma(\theta, \sigma)$. It is now easy to see that

$$\Sigma(\theta, \sigma) = \Sigma(\tilde{\theta}, \tilde{\sigma})$$

where

$$\tilde{\theta} := \frac{1}{\theta} \quad \text{and} \quad \tilde{\sigma} = |\theta|\sigma$$

This means that the likelihood for (θ, μ, σ) and $(\tilde{\theta}, \mu, \tilde{\sigma})$ will be identical. A consequence is that the Maximum Likelihood Esimates for MA(1) are not unique because if $(\hat{\theta}, \hat{\mu}, \hat{\sigma})$ maximizes likelihood, then $(1/\hat{\theta}, \hat{\mu}, |\hat{\theta}|\hat{\sigma})$ also maximizes likelihood.

Because two different parametrizations give rise to identical likelihoods, there is nonidentifiability in the $MA(1)$ model. To fix this, we can restrict attention to $|\theta| \leq 1$. This will rule out nonidentifiability because one of $|\theta|$ and $|\tilde{\theta}|$ will have to be strictly larger than 1.

From now on, we shall assume that $|\theta| \leq 1$ while dealing with the $MA(1)$ model.

2.3 An alternative estimation method and uncertainty quantification for $MA(1)$

Two problems with the Maximum Likelihood Estimation method described in Subsection 2.1 are:

1. It involves inverting the $n \times n$ matrix $\Gamma(\theta)$ for various values of θ and this can be cumbersome and expensive when n is large.
2. It is not clear how to do uncertainty quantification.

The method described in this section fixes both these issues. Instead of writing the likelihood as the joint density of $Y_1, \ldots, Y_n$ in one go using the multivariate normal density, we shall decompose the joint density as

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{Y_1}(y_1) f_{Y_2|Y_1=y_1}(y_2) f_{Y_3|Y_1=y_1,Y_2=y_2}(y_3) \cdots f_{Y_n|Y_1=y_1,\ldots,Y_{n-1}=y_{n-1}}(y_n)$$

and attempt to write down each conditional density on the right hand side explicitly. In order to write these conditional densities, the following formulae will be useful. The model equation $Y_t = \mu + Z_t + \theta Z_{t-1}$ is equivalent to

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

Using this equation for $t-1$ in place of Z_{t-1}, we get

$$Z_t = -\mu - \theta\left(-\mu - \theta Z_{t-2} + Y_{t-1}\right) + Y_t = -\mu(1-\theta) + Y_t - \theta Y_{t-1} + \theta^2 Z_{t-2}.$$

Continuing further, we can write Z_t in terms of $Y_t, \ldots, Y_1$ and Z_0 as follows:

$$Z_t = -\mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + Y_t - \theta Y_{t-1} + \theta^2 Y_{t-2} - \cdots + (-1)^{t-1}\theta^{t-1} Y_1 + (-1)^t \theta^t Z_0$$

for each $t = 1, 2, \ldots$. We can write the above equation with Y_t on the left hand side as

$$Y_t = Z_t + \mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + \theta Y_{t-1} - \theta^2 Y_{t-2} + \cdots - (-1)^{t-1}\theta^{t-1} Y_1 - (-1)^t \theta^t Z_0$$

Thus conditional on $Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}$, we get

$$Y_t = Z_t + \mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots - (-1)^{t-1}\theta^{t-1} y_1 - (-1)^t \theta^t Z_0$$

The conditional distribution of Y_t given $Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}$ is therefore normal with mean

$$\begin{aligned}&\mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots - (-1)^{t-1}\theta^{t-1} y_1 \\ &+ \mathbb{E}\left(Z_t \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}\right) - (-1)^t \theta^t \mathbb{E}\left(Z_0 \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}\right)\end{aligned}$$

and variance

$$\operatorname{var}\left(Z_t - (-1)^t \theta^t Z_0 \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}\right).$$

Because Z_t is independent of $Y_1 \ldots, Y_{t-1}$, the conditional mean above becomes:

$$\begin{aligned}&\mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots - (-1)^{t-1}\theta^{t-1} y_1 \\ &- (-1)^t \theta^t \mathbb{E}\left(Z_0 \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}\right).\end{aligned}$$

In order to calculate the above mean and variance, we need to understand the conditional distribution of Z_0 given $Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}$. This is somewhat tricky. One can avoid this by changing the model and assuming that $Z_0 = 0$. When $|\theta| < 1$, this will have little effect as the multiplier θ^t will be small for all but very small values of t. Under this assumption, the conditional distribution of Y_t given $Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}$ will be normal with mean

$$\mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots - (-1)^{t-1}\theta^{t-1} y_1$$

and variance

$$\operatorname{var}\left(Z_t \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}\right) = \sigma^2.$$

Therefore

$$\begin{aligned}&f_{Y_t \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}}(y_t) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_t - \mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) - \theta y_{t-1} + \theta^2 y_{t-2} - \cdots + (-1)^{t-1}\theta^{t-1} y_1\right)^2}{2\sigma^2}\right)\end{aligned}$$

for $t = 2, \ldots, n$. The likelihood is therefore obtained by taking the product of these for $t = 2, \ldots, n$ and multiplying by the density $N(\mu, \sigma^2)$ of Y_1. Thus the likelihood is

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

where

$$\begin{aligned}&S(\mu, \theta) = (y_1 - \mu)^2 \\ &+ \sum_{t=2}^{n} \left(y_t - \mu\left(1 - \theta + \theta^2 - \cdots + (-1)^{t-1}\theta^{t-1}\right) - \theta y_{t-1} + \theta^2 y_{t-2} - \cdots + (-1)^{t-1}\theta^{t-1} y_1\right)^2.\end{aligned}$$

For uncertainty quantification, we shall take a Bayesian approach and combine the likelihood with a prior on θ, μ, σ. 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. Note that we have restricted the range of θ to $(-1, 1)$ because we assumed that $|\theta| < 1$ for identifiability. The posterior is then

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

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 Two)':

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

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 let $\hat{\alpha}$ denote the minimizer of $S(\alpha)$ over α. This will be an approximation to the maximum likelihood estimator (approximation because this likelihood uses the assumption $Z_0 = 0$ which distorts the actual likelihood slightly). Taylor expansion for α near $\hat{\alpha}$ gives

$$S(\alpha) = S(\hat{\alpha}) + \langle \nabla S(\hat{\alpha}), \alpha - \hat{\alpha}\rangle + (\alpha - \hat{\alpha})^T \left(\frac{1}{2} H S(\hat{\alpha})\right)(\alpha - \hat{\alpha})$$
$$= S(\hat{\alpha}) + (\alpha - \hat{\alpha})^T \left(\frac{1}{2} H S(\hat{\alpha})\right)(\alpha - \hat{\alpha})$$

where we used $\nabla S(\hat{\alpha}) = 0$ because $\hat{\alpha}$ minimizes $S(\alpha)$. Therefore

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

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 $I\{-1 < \theta < 1, -C < \mu < C\}$)

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

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

3 Recommended Reading for Today

1. Parameter estimation and uncertainty quantification for $MA(1)$ are described in Example 3.32 and Example 3.33 of the book by Shumway and Stoffer titled *Time Series Analysis and its applications* (Fourth Edition).