Maximal likelihood parameters for ARIMA(0,1,1) model

Baruch Youssin

January 8, 2022

ARIMA(0,1,1) model

This model for a time series Y_t , t = …, -1, 0, 1, …, is characterized by the requirement that the first differences of the time series are generated by the Moving Average model of order 1, MA(1).

The parameters of ARIMA(0,1,1) model are the initial value Y_0 of the series (or any other element of the series, chosen in advance) and the parameters of MA(1).

As the initial value is known, it remains to estimate the parameters of MA(1) from the first differences of the original time series.

We denote these first differences by $X_t = Y_{t+1} - Y_t$, t = …, -1, 0, 1, …

At this point we can forget our original time series Y_t and concentrate on its first differences X_t .

The MA(1) model is described by the following formula:

(1) $X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1}$

where ε_t , t = …, -1, 0, 1, …, are innovations (white noise), c and θ_1 (where $|\theta_1| < 1$) are parameters of the model, and X_t is the time series generated by the model; see Wikipedia.

To specify the model, we also need to specify the parameters of the innovations ε_t . Usually they are taken to be independently normally distributed with zero mean, and their standard deviation σ is another parameter of the model.

Thus, the parameters of the model are Y_0 , c, θ_1 (satisfying $|\theta_1| < 1$) and σ .

Maximal likelihood estimation of the parameters

Given a finite time series , $Y_0, Y_1, \ldots, Y_{n+1}$ we can ask, what are the values of the parameters Y_0 , c, θ_1 and σ that generate such series with the maximal likelihood; the likelihood is defined as the probability density of generating Y_t with the specified values of the parameters.

While the given series is finite, we are considering the probability density of it being generated as a part of an infinite series generated by this model.

Finite ARIMA(0,1,1) model

The time series $Y_0, Y_1, ..., Y_{n+1}$ is determined by the vector of the innovation values $\varepsilon = (\varepsilon_{-1}, \varepsilon_0, \varepsilon_1, ..., \varepsilon_n)$ by the following formulas.

The vector of the first differences $X = (X_0, X_1, ..., X_n)$ is found as

$$X^T = A\,\varepsilon^T + c\,\hat{1}$$

where

$A = \begin{pmatrix} \theta_1 & 1 & 0 & ... & ... \\ 0 & \theta_1 & 1 & 0 & ... \\ 0 & 0 & \theta_1 & 1 & ... \\ ... & ... & ... & ... & ... \end{pmatrix}$, the size of A is (n + 1) x (n + 2) (n + 1 rows and n + 2 columns),

$\hat{1} = (1, 1, ..., 1)$ and the superscript T indicates matrix transpose.

Y_t are found then as the cumulative sums of X_t : $Y_t = Y_0 + X_0 + ... + X_{t-1}$.

Note that the innovation values $\varepsilon_{-1}, \varepsilon_0, \varepsilon_1, ..., \varepsilon_n$ that define $Y_0, Y_1, ..., Y_{n+1}$, are not unique.

The covariance matrix

The model MA(1) implies that the vector $X = (X_0, X_1, ..., X_n)$ is normally distributed with the center $(c, c, ..., c) = c\,\hat{1}$ and the following covariance matrix:

$E((X_t - c)^2) = E((\varepsilon_t + \theta_1 \varepsilon_{t-1})^2) = (\theta_1^2 + 1)\,\sigma^2$,

$E((X_t - c)(X_{t-1} - c)) = E((X_{t-1} - c)(X_t - c)) = E((\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-1} + \theta_1 \varepsilon_{t-2})) = \theta_1 \sigma^2$,

$E((X_{t_1} - c)(X_{t_2} - c)) = 0$ in all other cases,

$Cov = \sigma^2 \begin{pmatrix} \theta_1^2 + 1 & \theta_1 & 0 & ... \\ \theta_1 & \theta_1^2 + 1 & \theta_1 & ... \\ 0 & \theta_1 & \theta_1^2 + 1 & ... \\ ... & ... & ... & ... \end{pmatrix}$.

The likelihood

The probability density of $X = (X_0, X_1, ..., X_n)$ is given by

$$NormDensity(X - c\,\hat{1}; Cov) = \frac{\exp\left(-\frac{1}{2}(X - c\,\hat{1})\,Cov^{-1}(X - c\,\hat{1})^T\right)}{\sqrt{(2\pi)^{n+1}\,det(Cov)}} \quad .$$

It follows that maximal likelihood is equivalent to minimization of the loss which is the log-likelihood multiplied by (-2), up to an additive constant:

$$loss = (X - c\,\hat{1})\,Cov^{-1}(X - c\,\hat{1})^T + \log det(Cov) \quad ,$$

$$NormDensity\left(X - c\,\hat{1}; Cov\right) = (2\pi)^{-(n+1)/2} \exp\left(-\frac{1}{2} loss\right)$$

This loss is minimized with respect to the parameters c, θ_1 (satisfying $|\theta_1| < 1$) and σ.

On estimating the constant c

A recent reference Umberto Triacca suggests *in the context of maximal likelihood* to estimate c as the expectation of X: $\frac{1}{n+1}\sum_{t=0}^{n} X_t = \frac{X.\hat{1}^T}{n+1}$ where $X.\hat{1}^T$ is the (dot) product of X with the vector $\hat{1}$.

The logic behind this suggestion is clear: c is the expectation of the process that yields X, and we can estimate it in this way.

However, this estimation may be different from the estimation by the maximal likelihood principle (by minimizing the loss as above), as follows.

Minimizing the above loss with respect to only c yields the equation

$$-\hat{1} Cov^{-1}\left(X - c\,\hat{1}\right)^T - \left(X - c\,\hat{1}\right) Cov^{-1} \hat{1}^T = 0 \quad ,$$

$$c = \frac{X\, Cov^{-1} \hat{1}^T}{\hat{1}\, Cov^{-1} \hat{1}^T}$$

One can see that for large n (large being understood in comparison with $\frac{\theta_1 + 1}{\theta_1}$) this value of c is close to $\frac{X.\hat{1}^T}{n+1}$.

Indeed, the matrix Cov is Toeplitz; it is the matrix of a discrete convolution with a kernel of size 3. Hence, its inverse is also Toeplitz, the matrix of a discrete convolution with a kernel K of large size whose elements decrease away from its center. The denominator $\hat{1} Cov^{-1} \hat{1}^T$ is approximately (n + 1) times the sum of the elements K while the numerator $X\, Cov^{-1} \hat{1}^T$ is obtained by first applying the convolution with K to X – replacing each X_t with a linear combination of nearby elements – and then taking the sum of all elements of the result. If n is much larger than the size of the essential support of K (the region where the elements of K are not negligible) then $X\, Cov^{-1} \hat{1}^T$ is approximately the product of the sum of the elements of X (this is $X.\hat{1}^T$) and the sum of the elements of the kernel, and the ratio $c = \frac{X\, Cov^{-1} \hat{1}^T}{\hat{1}\, Cov^{-1} \hat{1}^T}$ is close to $\frac{X.\hat{1}^T}{n+1}$.

However, when n is not so large, there could be a difference.

Probability of observing a time series

The question

We may ask, what is the probability of observing a time series $Y_{n_1}, Y_{n_1+1}, \ldots, Y_{n_2+1}$ under an ARIMA(0,1,1) model with specific values of the parameters Y_0, c, θ_1 and σ.

Specifically, we may fit our model to a time series $Y_0, Y_1, \ldots, Y_{n+1}$ and ask, what is the probability of observing a time series $Y_{n+2}, \ldots, Y_{m+1}$.

Note that one of the parameters of the fitted model is Y_0 (it can be possibly replaced by any other element of $Y_0, Y_1, \ldots, Y_{n+1}$).

Thus, our question amounts to asking, what is the probability of observing Y_0 at time 0 and $Y_{n+2}, \ldots, Y_{m+1}$ at the times n+2,…,m+1 under the parameters c, θ_1 and σ .

This question involves observing multiple time blocks; this question is more difficult and we shall discuss it later.

We replace this question by a different one which is simpler: what is the probability of observing $Y_{n+2}, \ldots, Y_{m+1}$ *on condition* that before that we have observed $Y_0, Y_1, \ldots, Y_{n+1}$ under the same model.

This is equivalent to asking, what is the probability of observing $Y_{n+1}, Y_{n+2}, \ldots, Y_{m+1}$ under the parameters c, θ_1 and σ .

To simplify the notation, we ask, what is the probability of observing $Y_0, Y_1, \ldots, Y_{n+1}$ under the parameters c, θ_1 and σ .

The answers

The precise answer to this question is 0, since the probability distribution is continuous.

We can modify the question and ask for the value of the *probability distribution*; it is given by the above formulas.

However, the original question on the *probability* has informal answers, as follows.

The probability is a measure how probable or improbable certain values of the random variables are.

For the univariate normal distribution centered at zero, the probable values of x are those close to zero, and the improbable ones satisfy $|x| \gg \sigma$ where σ is the standard deviation. The measure of this for a given value x_0 is

$$P\left(x : |x| \geq |x_0|\right) 1 - \left(\Phi\left(|x_0|\right) - \Phi\left(-|x_0|\right)\right) = 1 - 2\,\Phi\left(-|x_0|\right)$$

where $\Phi(r)$ is the Cumulative Normal Distribution function.

This is related to the notion of *confidence interval*.

For the multivariate normal distribution that defines ARIMA(0,1,1), the probable values are those which can be generated by probable values of innovations, and the improbable ones are those that can be generated *only* by improbable values of innovations. A simple way to decide which values are more probable or less probable, is to compare their probability densities.

Thus, for the *informal* probability of the time series $Y = \left(Y_0, Y_1, \ldots, Y_{n+1}\right)$ we can take the probability

$$P\left(Y' : NormDensity\left(X' - c\,\hat{1}\,; Cov\right) \leq NormDensity\left(X - c\,\hat{1}\,; Cov\right)\right)$$

$$= P\left(X' : (X' - c\hat{1}) Cov^{-1} (X' - c\hat{1})^T \geq (X - c\hat{1}) Cov^{-1} (X - c\hat{1})^T \right)$$

Here the probability is taken with respect to the distribution of X' which is the multivariate normal distribution with the covariance matrix Cov; making a linear coordinate change from X' to variables U whose covariance matrix is unity, shows that this probability is equal to

$$P(U : U \geq r)$$

where $r = \sqrt{(X - c\hat{1}) Cov^{-1} (X - c\hat{1})^T}$.

The probability $P(U : U \geq r)$ is the probability that a point lies outside the ball of radius r under the standard multivariate normal distribution of dimension n + 1. This probability can be found explicitly but the formula is complicated.

Since we need an informal estimate, we replace the ball by the cube $[-r, r]^{n+1}$; the probability that U is outside this cube, is

$$1 - (\Phi(r) - \Phi(-r))^{n+1} = 1 - (1 - 2\Phi(-r))^{n+1}$$.

For large r $\Phi(-r)$ is very small and the above formula may evaluate to 0 due to precision loss; if r > 4, we replace it by $2(n+1)\Phi(-r)$.

Note that these measures are quite different: I have observed a case when the probability density was small, ~ 1e-10, while the probability estimate was close to 1. This was due to the fact that learning did not progress sufficiently and σ was still large, ~ 14, and this made r small. In other words, a large value of σ made the values of X' probable.

Multiple pieces of a time series

Suppose we are given a few contiguous time blocks of a time series, say,
$Y_0, Y_1, \ldots, Y_{n_1+1}, Y_{n_2}, \ldots, Y_{n_3+1}, Y_{n_4}, \ldots, Y_{n_m+1}$.

As in case of one time block, we want to estimate the parameters of an ARIMA(0,1,1) model that generates the series containing these blocks.

These parameters consist of the parameters of the MA(1) model, c, θ_1 (satisfying $|\theta_1| < 1$) and σ , and any one element Y_t of the time series, chosen in advance.

When we take the differences of this series, we need to take the differences within the blocks and the differences between the blocks:

$X_0 = Y_1 - Y_0$, $X_1 = Y_2 - Y_1$, …, $X_{n_1} = Y_{n_1+1} - Y_{n_1}$, $X_{n_1+1 \to n_2-1} = Y_{n_2} - Y_{n_1+1}$,
$X_{n_2} = Y_{n_2+1} - Y_{n_2}$, …, $X_{n_m} = Y_{n_m+1} - Y_{n_m}$

We have

$$X_{n_1+1 \to n_2-1} = X_{n_1+1} + X_{n_1+2} + \ldots + X_{n_2-1}$$.

It follows that the vector $(X_0, X_1, \ldots, X_{n_1}, X_{n_1+1 \to n_2-1}, X_{n_2}, \ldots, X_{n_m})$ is normally distributed.

Its center is determined by the expectations

$E(X_t) = c$ for all t, and

$E(X_{n_1+1\to n_2-1})=(n_2-n_1-1)c$ (similarly for the other gaps),

Its covariance matrix can be found from the following equalities:

$E((X_t-c)^2)=(\theta_1^2+1)\sigma^2$ and $E((X_t-c)(X_{t-1}-c))=\theta_1\sigma^2$ as before,

$E((X_{n_1+1\to n_2-1}-E(X_{n_1+1\to n_2-1}))^2)=\left((n_2-n_1-1)(\theta_1^2+1)+2(n_2-n_1-2)\theta_1\right)\sigma^2$ and similarly for the other gaps,

$E((X_{n_1+1\to n_2-1}-E(X_{n_1+1\to n_2-1}))(X_{n_1}-c))=\theta_1\sigma^2$ and similarly for the other gaps,

$E((X_{n_2}-c)(X_{n_1+1\to n_2-1}-E(X_{n_1+1\to n_2-1})))=\theta_1\sigma^2$ and similarly for the other gaps, with all the other covariances being zero.

If we denote this covariance matrix by Cov, the above formulas for the probability density and for the loss remain valid.