Notes on ARIMA

Jonas Pichat

10 November 2019

1 ARMA

Given a centred time series $y = y_1, \dots, y_n$ of finite rank, ARIMA(p, d=0, q) equation is:

$$y_{t+1} = \mu(=0) + \epsilon_{t+1} + \Sigma_{i=1}^{q}\theta_i\epsilon_{t-i+1} + \Sigma_{i=1}^{p}\alpha_i y_{t-i+1} \tag{1}$$

where $\epsilon^{(h)}$ is the series of residuals referring to the h-th target[1], y_{t+h} (only 1 step ahead is used here, so the superscript was dropped for clarity).

The first part (in ϵ) is the MA(q) model and the last sum (in y) corresponds to the AR(p) model. As a side note, one could verify that the residuals are not correlated and iid variables sampled from a normal distribution with 0 mean.

It is also worth mentioning that we restricted ourselves to a (prediction) gap, g=0 here and used a non-differentiated series[2] (d=0) without any additional (exogenous) data; in such case, given k exogenous series $x^{(k)}$ (y still being our endogenous, unique target series) and a non-null gap, eq. (1) just becomes:

$$\begin{split} y_{t+g+1} = \epsilon_{t+g+1} + f(\epsilon_t, \dots, \epsilon_{t-q+1}, y_t, \dots, y_{t-p+1}, x_t^{(1)}, \dots, x_{t-p+1}^{(1)}, \dots, \\ x_t^{(k)}, \dots, x_{t-p+1}^{(k)}, \theta, \alpha) \end{split} \tag{2}$$

where f is a polynomial function of degree 1 with more variables and coefficient sets θ and α (referring to MA and AR respectively).

1.1 Difference with the classical ARIMA

When $q > 0$, we do not compute the series of residuals **exactly**, like in the classical MA model (i.e. via backward recursion). The reason is that in order to compute a residual, one needs a target (so $q + h$ observations)...but this is not

[1] here, "target" is to be taken in the scope of a prediction horizon: we do consider a unique "target time series", y but one may be interested in predicting multiple (h) steps ahead: $y_{t+1}, y_{t+2}, \dots, y_{t+h}$, or more generally $y_{t+g+1}, \dots, y_{t+g+h+1}, \forall g \geq 0, h > 0$.

[2] when using ARIMA(p, d>0, q), y and x are replaced with their differentiated versions in (2). One could also encounter the lag operator $\mathscr{L}$ in the literature, which, when applied to y simply gives $\mathscr{L}^i y_t = y_{t-i}$ (it just avoids carrying all the $t-i+1$ indices along); besides, note that $(1-\mathscr{L}^1)y_t = y_t - y_{t-1}$. For $d > 0$, $g = 0$, (1) can therefore be rewritten as: $\big(1 - \sum_{i=1}^{p}\alpha_i\mathscr{L}^i\big)(1-\mathscr{L}^1)^d y_{t+1} = \big(1 - \sum_{i=1}^{q}\theta_i\mathscr{L}^i\big)\epsilon_{t+1}$; see Section **??** for more details.

always possible (e.g. at the end of the series). So instead, we rather estimate a residual series $\hat{\epsilon}$, at once, from the set of all q consecutive observations available via a linear predictor (see `ResidualPredictor`), and use it in eq. (1). The ARMA equation thus becomes:

$$y_{t+1} - \hat{\epsilon}_{t+1} = \underbrace{\Sigma_{i=1}^{q}\theta_i\hat{\epsilon}_{t-i+1} + \Sigma_{i=1}^{p}\alpha_i y_{t-i+1}}_{\widetilde{y}_{t+1}}$$

Therefore, we seek α and θ that minimise:

$$\text{argmin}_{\alpha,\Theta}\Sigma_{t>\max(p,2q)}\Big[(y_{t+1} - \hat{\epsilon}_{t+1}) - \widetilde{y}_{t+1}\Big]^2$$

1.2 ARIMA(p, 0, q)

In general (though here, exogenous features were excluded for clarity), this means that the "feature-target" matrix, X (with h=2 targets) fed to the final linear model (within the `fit` method of the class `ARIMA`) has the form:

$$\left[\begin{array}{ccc|cccc|cccc|cc}
. & . & . & . & . & . & . & . & . & . & . & . & y_1 - \hat{\epsilon}_1^{(2)} \\
. & . & . & . & . & . & . & . & . & . & . & y_1 - \hat{\epsilon}_1^{(1)} & y_2 - \hat{\epsilon}_2^{(2)} \\
. & . & y_1 & . & . & . & \hat{\epsilon}_1^{(1)} & . & . & . & \hat{\epsilon}_1^{(2)} & y_2 - \hat{\epsilon}_2^{(1)} & y_3 - \hat{\epsilon}_3^{(2)} \\
. & y_1 & y_2 & . & . & \hat{\epsilon}_1^{(1)} & \hat{\epsilon}_2^{(1)} & . & . & \hat{\epsilon}_1^{(2)} & \hat{\epsilon}_2^{(2)} & y_3 - \hat{\epsilon}_3^{(1)} & y_4 - \hat{\epsilon}_4^{(2)} \\
y_1 & y_2 & y_3 & . & \hat{\epsilon}_1^{(1)} & \hat{\epsilon}_2^{(1)} & \hat{\epsilon}_3^{(1)} & . & \hat{\epsilon}_1^{(2)} & \hat{\epsilon}_2^{(2)} & \hat{\epsilon}_3^{(2)} & y_4 - \hat{\epsilon}_4^{(1)} & y_5 - \hat{\epsilon}_5^{(2)} \\
y_2 & y_3 & y_4 & \hat{\epsilon}_1^{(1)} & \hat{\epsilon}_2^{(1)} & \hat{\epsilon}_3^{(1)} & \hat{\epsilon}_4^{(1)} & \hat{\epsilon}_1^{(2)} & \hat{\epsilon}_2^{(2)} & \hat{\epsilon}_3^{(2)} & \hat{\epsilon}_4^{(2)} & y_5 - \hat{\epsilon}_5^{(1)} & y_6 - \hat{\epsilon}_6^{(2)} \\
y_3 & y_4 & y_5 & \hat{\epsilon}_2^{(1)} & \hat{\epsilon}_3^{(1)} & \hat{\epsilon}_4^{(1)} & \hat{\epsilon}_5^{(1)} & \hat{\epsilon}_2^{(2)} & \hat{\epsilon}_3^{(2)} & \hat{\epsilon}_4^{(2)} & \hat{\epsilon}_5^{(2)} & y_6 - \hat{\epsilon}_6^{(1)} & y_7 - \hat{\epsilon}_7^{(2)} \\
y_4 & y_5 & y_6 & \hat{\epsilon}_3^{(1)} & \hat{\epsilon}_4^{(1)} & \hat{\epsilon}_5^{(1)} & \hat{\epsilon}_6^{(1)} & \hat{\epsilon}_3^{(2)} & \hat{\epsilon}_4^{(2)} & \hat{\epsilon}_5^{(2)} & \hat{\epsilon}_6^{(2)} & y_7 - \hat{\epsilon}_7^{(1)} & y_8 - \hat{\epsilon}_8^{(2)} \\
y_5 & y_6 & y_7 & \hat{\epsilon}_4^{(1)} & \hat{\epsilon}_5^{(1)} & \hat{\epsilon}_6^{(1)} & \hat{\epsilon}_7^{(1)} & \hat{\epsilon}_4^{(2)} & \hat{\epsilon}_5^{(2)} & \hat{\epsilon}_6^{(2)} & \hat{\epsilon}_7^{(2)} & y_8 - \hat{\epsilon}_8^{(1)} & y_9 - \hat{\epsilon}_9^{(2)} \\
y_6 & y_7 & y_8 & \hat{\epsilon}_5^{(1)} & \hat{\epsilon}_6^{(1)} & \hat{\epsilon}_7^{(1)} & \hat{\epsilon}_8^{(1)} & \hat{\epsilon}_5^{(2)} & \hat{\epsilon}_6^{(2)} & \hat{\epsilon}_7^{(2)} & \hat{\epsilon}_8^{(2)} & y_9 - \hat{\epsilon}_9^{(1)} & y_{10} - \hat{\epsilon}_{10}^{(2)} \\
 & \vdots & & & & \vdots & & & & \vdots & & \vdots & \\
y_{n-(p+h-1)} & \cdots & y_{n-h} & \hat{\epsilon}_{n-(q+h-1)}^{(1)} & & \cdots & \hat{\epsilon}_{n-h}^{(1)} & \hat{\epsilon}_{n-(q+h-1)}^{(2)} & & \cdots & \hat{\epsilon}_{n-h}^{(2)} & y_{n-h+1} - \hat{\epsilon}_{n-h+1}^{(1)} & y_n - \hat{\epsilon}_n^{(2)}
\end{array}\right]$$

where the first $\max(p, 2q) + g + (h-1)$ (recall g=0 here) rows should be excluded from training, and $\Theta = [\theta_1, \theta_2, \theta_3]^T$ and $\alpha = [\alpha_1^{(1)}, \dots, \alpha_4^{(1)}, \alpha_1^{(2)}, \dots, \alpha_4^{(2)}]^T$ are the parameters to be optimised.

The first column block constitutes the AR(3) features, the second, the MA(4) features (residuals)[3] relative to the first target (see superscript "1"), the third, the MA(4) features relative the second target and the last block is the targets.

Note that one should see this final linear model (when $q > 0$) as a "corrected predictor", where the error has been predicted and is hereby accounted for.

Finally, including exogenous data comes down to adding k column blocks similar to the leftmost one but with $x^{(k)}$ series (and the h residual series, $\epsilon^{(h)}$ would now be estimated using both endogenous and exogenous data[4]).

[3] recall that $\hat{\epsilon}_{t+1} = f(y_t, \dots, y_{t-q+1}, \beta)$ where f is a polynomial function of degree 1 which coefficients are optimised using `ResidualPredictor`. Hence, residuals are only defined for $t > q$ (and conversely, $\hat{\epsilon}_1, \dots, \hat{\epsilon}_q$ are not defined).

[4] $\hat{\epsilon}_{t+h} = f(y_t, \dots, y_{t-q+1}, x_t^{(1)}, \dots, x_{t-q+1}^{(1)}, x_t^{(k)}, \dots, x_{t-q+1}^{(k)}, \beta)$

For the sake of completeness, we detail X for AR(p=3) (and h=2 targets) and MA(q=4) (and h=1 target) in the following.

1.3 ARIMA(3, 0, 0)

Considering h=2 targets, X has the form:

$$\left[\begin{array}{ccc|cc}
\cdot & \cdot & \cdot & \cdot & y_1 \\
\cdot & \cdot & \cdot & y_1 & y_2 \\
\cdot & \cdot & y_1 & y_2 & y_3 \\
\cdot & y_1 & y_2 & y_3 & y_4 \\
y_1 & y_2 & y_3 & y_4 & y_5 \\
y_2 & y_3 & y_4 & y_5 & y_6 \\
 & \vdots & & \vdots & \\
y_{n-(p+h-1)} & \cdots & y_{n-h} & y_{n-1} & y_n
\end{array}\right]$$

where the first p rows should be excluded from training.

1.4 ARIMA(0, 0, 4)

Considering h=1 target, X has the form:

$$\left[\begin{array}{cccc|c}
\cdot & \cdot & \cdot & \cdot & y_1 - \hat{\epsilon}_1 \\
\cdot & \cdot & \cdot & \hat{\epsilon}_1 & y_2 - \hat{\epsilon}_2 \\
\cdot & \cdot & \hat{\epsilon}_1 & \hat{\epsilon}_2 & y_3 - \hat{\epsilon}_3 \\
\cdot & \hat{\epsilon}_1 & \hat{\epsilon}_2 & \hat{\epsilon}_3 & y_4 - \hat{\epsilon}_4 \\
\hat{\epsilon}_1 & \hat{\epsilon}_2 & \hat{\epsilon}_3 & \hat{\epsilon}_4 & y_5 - \hat{\epsilon}_5 \\
\hat{\epsilon}_2 & \hat{\epsilon}_3 & \hat{\epsilon}_4 & \hat{\epsilon}_5 & y_6 - \hat{\epsilon}_6 \\
\hat{\epsilon}_3 & \hat{\epsilon}_4 & \hat{\epsilon}_5 & \hat{\epsilon}_6 & y_7 - \hat{\epsilon}_7 \\
\hat{\epsilon}_4 & \hat{\epsilon}_5 & \hat{\epsilon}_6 & \hat{\epsilon}_7 & y_8 - \hat{\epsilon}_8 \\
\hat{\epsilon}_5 & \hat{\epsilon}_6 & \hat{\epsilon}_7 & \hat{\epsilon}_8 & y_9 - \hat{\epsilon}_9 \\
 & & \vdots & & \vdots \\
\hat{\epsilon}_{n-q} & & \cdots & \hat{\epsilon}_{n-1} & y_n - \hat{\epsilon}_n
\end{array}\right]$$

where first $2q$ rows should be excluded from training.

1.5 Prediction function

The information needed from the model for the prediction function consists of: the parameters of the model p, d, q and the weights and biases W_1, W_2, b_1, b_2[5] such that:

[5] W_1 and b_1 come from the residual predictor that maps nq observations (n is the number of series under consideration: exogenous + (1) endogenous) with h residual(s). Note that Eq. **??** holds for $d > 0$ except all series are differentiated first—and the prediction must then be back-transformed (see Section **??**).