

Autoregressive integrated moving average

From Wikipedia, the free encyclopedia

In statistics and econometrics, and in particular in time series analysis, an **autoregressive integrated moving average (ARIMA)** model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.^[1]

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

Non-seasonal ARIMA models are generally denoted $ARIMA(p,d,q)$ where parameters p , d , and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. Seasonal ARIMA models are usually denoted $ARIMA(p,d,q)(P,D,Q)_m$, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.^[2]^[3]

When two out of the three terms are zeros, the model may be referred to based on the non-zero parameter, dropping "AR", "I" or "MA" from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

ARIMA models can be estimated following the Box–Jenkins approach.

Contents

- 1 Definition
- 2 Other special forms
- 3 Differencing
- 4 Examples
- 5 Choosing the order
- 6 Estimation of coefficients
- 7 Forecasts using ARIMA models
 - 7.1 Forecast intervals
- 8 Variations and extensions
- 9 Software implementations
- 10 See also
- 11 References
- 12 Further reading
- 13 External links

Definition

Given a time series of data X_t where t is an integer index and the X_t are real numbers, an $ARMA(p,q)$ model is given by

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

or equivalently by

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_t are error terms. The error terms ε_t are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

Assume now that the polynomial $\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right)$ has a unit root (a factor $(1 - L)$) of multiplicity d . Then it can be rewritten as:

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) = \left(1 - \sum_{i=1}^{p'-d} \phi_i L^i\right) (1 - L)^d.$$

An ARIMA(p, d, q) process expresses this polynomial factorisation property with $p=p'-d$, and is given by:

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

and thus can be thought as a particular case of an ARMA($p+d, q$) process having the autoregressive polynomial with d unit roots. (For this reason, no ARIMA model with $d > 0$ is wide sense stationary.)

The above can be generalized as follows.

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \delta + \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t.$$

This defines an ARIMA(p, d, q) process with **drift** $\delta/(1 - \sum \phi_i)$.

Other special forms

The explicit identification of the factorisation of the autoregression polynomial into factors as above, can be extended to other cases, firstly to apply to the moving average polynomial and secondly to include other special factors. For example, having a factor $(1 - L^s)$ in a model is one way of including a non-stationary seasonality of period s into the model; this factor has the effect of re-expressing the data as changes from s periods ago. Another example is the factor $(1 - \sqrt{3}L + L^2)$, which includes a (non-stationary) seasonality of period 2. The effect of the first type of factor is to allow each season's value to drift separately over time, whereas with the second type values for adjacent seasons move together.

Identification and specification of appropriate factors in an ARIMA model can be an important step in modelling as it can allow a reduction in the overall number of parameters to be estimated, while allowing the imposition on the model of types of behaviour that logic and experience suggest should be there.

Differencing

Differencing in statistics is a transformation applied to time-series data in order to make it stationary. A stationary time series' properties do not depend on the time at which the series is observed.

In order to difference the data, the difference between consecutive observations is computed. Mathematically, this is shown as

$$y'_t = y_t - y_{t-1}$$

Differencing removes the changes in the level of a time series, eliminating trend and seasonality and consequently stabilizing the mean of the time series.

Sometimes it may be necessary to difference the data a second time to obtain a stationary time series, which is referred to as **second order differencing**:

$$\begin{aligned} y_t^* &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

Another method of differencing data is **seasonal differencing**, which involves computing the difference between an observation and the corresponding observation in the previous year. This is shown as:

$$y'_t = y_t - y_{t-m} \quad \text{where } m = \text{number of seasons.}$$

The differenced data is then used for the estimation of an ARMA model.

Examples

Some well-known special cases arise naturally or are mathematically equivalent to other popular forecasting models. For example:

- An ARIMA(0,1,0) model (or I(1) model) is given by $X_t = X_{t-1} + \varepsilon_t$ — which is simply a random walk.
- An ARIMA(0,1,0) with a constant, given by $X_t = c + X_{t-1} + \varepsilon_t$ — which is a random walk with drift.
- An ARIMA(0,0,0) model is a white noise model.
- An ARIMA(0,1,2) model is a Damped Holt's model.
- An ARIMA(0,1,1) model without constant is a basic exponential smoothing model.^[4]
- An ARIMA(0,2,2) model is given by $X_t = 2X_{t-1} - X_{t-2} + (\alpha + \beta - 2)\varepsilon_{t-1} + (1 - \alpha)\varepsilon_{t-2} + \varepsilon_t$ — which is equivalent to Holt's linear method with additive errors, or double exponential smoothing.^[4]

Choosing the order

To determine the order of a non-seasonal ARIMA model, a useful criterion is the Akaike information criterion (AIC) . It is written as

$$AIC = -2 \log(L) + 2(p + q + k + 1),$$

where L is the likelihood of the data, p is the order of the autoregressive part and q is the order of the moving average part. The parameter k in this criterion is defined as the number of parameters in the model being fitted to the data. For AIC, if $k = 1$ then $c \neq 0$ and if $k = 0$ then $c = 0$.

The corrected AIC for ARIMA models can be written as

$$AIC_c = AIC + (2(p + q + k + 1)(p + q + k + 2))/(T - p - q - k - 2).$$

The Bayesian Information Criterion can be written as

$$BIC = AIC + (\log(T) - 2)(p + q + k + 1).$$

The objective is to minimize the AIC, AICc or BIC values for a good model. The lower the value of one of these criteria for a range of models being investigated, the better the model will suit the data. It should be noted however that the AIC and the BIC are used for two completely different purposes. Whilst the AIC tries to approximate models towards the reality of the situation, the BIC attempts to find the perfect fit. The BIC approach is often criticized as there never is a perfect fit to real-life complex data; however, it is still a useful method for selection as it penalizes models more heavily for having more parameters than the AIC would.

AICc can only be used to compare ARIMA models with the same orders of differencing. For ARIMAs with different orders of differencing, RMSE can be used for model comparison.

Estimation of coefficients

Forecasts using ARIMA models

The ARIMA model can be viewed as a "cascade" of two models. The first is non-stationary:

$$Y_t = (1 - L)^d X_t$$

while the second is wide-sense stationary:

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) Y_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t.$$

Now forecasts can be made for the process Y_t , using a generalization of the method of autoregressive forecasting.

Forecast intervals

The forecast intervals (confidence intervals for forecasts) for ARIMA models are based on assumptions that the residuals are uncorrelated and normally distributed. If either of these assumptions does not hold, then the forecast intervals may be incorrect. For this reason, researchers plot the ACF and histogram of the residuals to check the assumptions before producing forecast intervals.

95% forecast interval: $\hat{y}_{T+h|T} \pm 1.96 \sqrt{v_{T+h|T}}$, where $v_{T+h|T}$ is the variance of $y_{T+h} | y_1, \dots, y_T$.

For $h = 1$, $v_{T+h|T} = \hat{\sigma}^2$ for all ARIMA models regardless of parameters and orders.

For ARIMA(0,0,q), $y_t = e_t + \sum_{i=1}^q \theta_i e_{t-i}$.

$$v_{T+h|T} = \hat{\sigma}^2 \left[1 + \sum_{i=1}^{h-1} \theta_i e_{t-i} \right], \text{ for } h = 2, 3, \dots$$

In general, forecast intervals from ARIMA models will increase as the forecast horizon increases.

Variations and extensions

A number of variations on the ARIMA model are commonly employed. If multiple time series are used then the X_t can be thought of as vectors and a VARIMA model may be appropriate. Sometimes a seasonal effect is suspected in the model; in that case, it is generally better to use a SARIMA (seasonal ARIMA) model than to

increase the order of the AR or MA parts of the model. If the time-series is suspected to exhibit long-range dependence, then the d parameter may be allowed to have non-integer values in an autoregressive fractionally integrated moving average model, which is also called a Fractional ARIMA (FARIMA or ARFIMA) model.

Software implementations

Various packages that apply methodology like Box–Jenkins parameter optimization are available to find the right parameters for the ARIMA model.

- EViews: has extensive ARIMA and SARIMA capabilities.
- Julia: contains an ARIMA implementation in the TimeModels package^[5]
- Mathematica: includes ARIMAProcess function.
- MATLAB: the Econometrics Toolbox includes ARIMA models and regression with ARIMA errors
- NCSS: includes several procedures for ARIMA fitting and forecasting.^{[6][7][8]}
- Python: the "statsmodels" package includes models for time series analysis – univariate time series analysis: AR, ARIMA – vector autoregressive models, VAR and structural VAR – descriptive statistics and process models for time series analysis.
- R: the standard R *stats* package includes an *arima* function, which is documented in "ARIMA Modelling of Time Series". Besides the *ARIMA(p,d,q)* part, the function also includes seasonal factors, an intercept term, and exogenous variables (*xreg*, called "external regressors"). The CRAN task view on Time Series is the reference with many more links. The "forecast" package in R can automatically select an ARIMA model for a given time series with the *auto.arima()* function. The package can also simulate seasonal and non-seasonal ARIMA models with its *simulate.Arima()* function. It also has a function *Arima()*, which is a wrapper for the *arima* from the "stats" package.^[9]
- Ruby: the "statsample-timeseries" gem is used for time series analysis, including ARIMA models and Kalman Filtering.
- SAFE TOOLBOXES: includes ARIMA modelling and regression with ARIMA errors.
- SAS: includes extensive ARIMA processing in its Econometric and Time Series Analysis system: SAS/ETS.
- IBM SPSS: includes ARIMA modeling in its Statistics and Modeler statistical packages. The default Expert Modeler feature evaluates a range of seasonal and non-seasonal autoregressive (p), integrated (d), and moving average (q) settings and seven exponential smoothing models. The Expert Modeler can also transform the target time-series data into its square root or natural log. The user also has the option to restrict the Expert Modeler to ARIMA models, or to manually enter ARIMA nonseasonal and seasonal p , d , and q settings without Expert Modeler. Automatic outlier detection is available for seven types of outliers, and the detected outliers will be accommodated in the time-series model if this feature is selected.
- SAP: the APO-FCS package^[10] in SAP ERP from SAP allows creation and fitting of ARIMA models using the Box–Jenkins methodology.
- SQL Server Analysis Services: from Microsoft includes ARIMA as a Data Mining algorithm.
- Stata includes ARIMA modelling (using its *arima* command) as of Stata 9.
- TOL (Time Oriented Language) is designed to model ARIMA models (including SARIMA, ARIMAX and DSARIMAX variants) [1].
- Scala: spark-timeseries library contains ARIMA implementation for Scala, Java and Python. Implementation is designed to run on Apache Spark.

See also

- Autocorrelation
- ARMA
- X-12-ARIMA
- Partial autocorrelation

References

1. For further information on Stationarity and Differencing see <https://www.otexts.org/fpp/8/1>
2. "Notation for ARIMA Models" (https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tffordet_sect016.htm). *Time Series Forecasting System*. SAS Institute. Retrieved 19 May 2015.
3. Hyndman, Rob J; Athanasopoulos, George. "8.9 Seasonal ARIMA models" (<https://www.otexts.org/fpp/8/9>). *Forecasting: principles and practice*. oTexts. Retrieved 19 May 2015.
4. "Introduction to ARIMA models" (<http://people.duke.edu/~rnau/411arim.htm#arima010>). *people.duke.edu*. Retrieved 2016-06-05.
5. TimeModels.jl (<https://github.com/JuliaStats/TimeModels.jl>) www.github.com
6. ARIMA in NCSS (<http://ncss.wpengengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/ARIMA-Box-Jenkins.pdf>),
7. Automatic ARMA in NCSS (http://ncss.wpengengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Automatic_ARMA.pdf),
8. Autocorrelations and Partial Autocorrelations in NCSS (<http://ncss.wpengengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Autocorrelations.pdf>)
9. "8.7 ARIMA modelling in R | OTexts" (<https://www.otexts.org/fpp/8/7>). *www.otexts.org*. Retrieved 2016-05-12.
10. "Box Jenkins model" (http://help.sap.com/saphelp_45b/helpdata/en/35/8a524b52060634e10000009b38f9b9/content.htm). SAP. Retrieved 8 March 2013.

Further reading

- Asteriou, Dimitros; Hall, Stephen G. (2011). "ARIMA Models and the Box–Jenkins Methodology". *Applied Econometrics* (Second ed.). Palgrave MacMillan. pp. 265–286. ISBN 978-0-230-27182-1.
- Mills, Terence C. (1990). *Time Series Techniques for Economists*. Cambridge University Press. ISBN 0-521-34339-9.
- Percival, Donald B.; Walden, Andrew T. (1993). *Spectral Analysis for Physical Applications*. Cambridge University Press. ISBN 0-521-35532-X.

External links

- The US Census Bureau uses ARIMA for "seasonally adjusted" data (programs, docs, and papers here)
- Lecture notes on ARIMA models (Robert Nau, Duke University)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Autoregressive_integrated_moving_average&oldid=787780468"

-
- This page was last edited on 27 June 2017, at 14:04.
 - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.