Lecture 20

ARIMA stands for **AutoRegressive Integrated Moving Average**, and it is a popular method for modeling time series data.

ARIMA models are based on the idea that past values of a time series can help predict future values. The model includes three key components:

Autoregression (AR) component: This component models the relationship between an observation and a linear combination of its past values. The "p" parameter in ARIMA(p,d,q) refers to the number of past values used in the autoregression component.

An AR only model is arima(1,0,0) with one lag. It's arima(2,0,0) for two lags.

Integrated (I) component: This component models the non-stationarity of the time series by including the differences between consecutive observations. The "d" parameter in ARIMA(p,d,q) refers to the number of times the time series needs to be differenced in order to make it stationary.

Moving Average (MA) component: This component models the relationship between an observation and a linear combination of past errors or residuals. The "q" parameter in ARIMA(p,d,q) refers to the number of past errors used in the moving average component.

An MA only model is arima(0,0,1).

Together, these three components make up an ARIMA model.

To use an ARIMA model, you need to determine the appropriate values for p, d, and q. Once the ARIMA model is fitted to the time series data, it can be used to make forecasts of future values. The accuracy of the forecasts can be evaluated using various measures, such as mean squared error (MSE) or mean absolute error (MAE).

ARIMA models are widely used in various fields, including finance, economics, engineering, and social sciences, to analyze and forecast time series data.

The components of an ARIMA model (p, d, q) can be determined by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series. Here are the steps to follow:

Stationarity Check: Check whether the time series is stationary or not. If the time series is non-stationary, then it needs to be differenced to make it stationary.

Determine the value of d: The number of times the time series needs to be differenced to make it stationary is the value of d. If the time series is already stationary, then d=0.

Determine the value of q: Look at the ACF plot and identify the lag at which the autocorrelation is significant. The value of q corresponds to the number of lags that are significant in the ACF plot after accounting for the effect of differencing.

Determine the value of p: Look at the PACF plot and identify the lag at which the partial autocorrelation is significant. The value of p corresponds to the number of lags that are significant in the PACF plot after accounting for the effect of differencing.

Finalize the values of p, d, and q: Use the identified values of p, d, and q to build the ARIMA model.

Model Selection: It is recommended to try different combinations of p, d, and q to find the best model that fits the data. You can use model selection criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to compare the performance of different models.

Once the components of the ARIMA model are identified, the model can be fitted to the time series data, and forecasts can be generated.

You can fit one, or more than one components of the ARIMA model. For example, to fit a moving average (MA) only model to a time series, we need to follow these steps:

Check Stationarity: Ensure that the time series is stationary or can be made stationary by differencing. Determine the value of q: Analyze the autocorrelation function (ACF) of the differenced time series to identify the number of significant lags. The value of q corresponds to the number of lags that are significant in the ACF plot.

Choose the order of the MA model: The order of the MA model is determined by the value of q.

Estimate the model coefficients: Use maximum likelihood estimation (MLE) to estimate the model coefficients. The MLE method involves finding the set of parameters that maximize the likelihood of observing the data.

Model Selection: It is recommended to try different orders of the MA model to find the best model that fits the data. You can use model selection criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to compare the performance of different models. While you have a q value to start with, you may need to try other nearby values to confirm that this is the best value.

Model Diagnostic Check: Check the residuals of the fitted model for any patterns or trends. If there are any significant patterns in the residuals, then it indicates that the model is not able to capture all the information in the data.

Once the MA model is fitted, we can use it to make forecasts of future values of the time series. We can also use it to analyze the impact of different factors on the time series by including them as exogenous variables in the model.

There are several variations of ARIMA models that can be used to model time series data. Some of the common variations are:

Seasonal ARIMA (SARIMA): SARIMA is used for time series data with seasonal patterns. It includes seasonal differences in addition to the non-seasonal differences used in ARIMA models.

Vector Autoregression (VAR): VAR models are used when multiple time series variables influence each other. In this approach, each variable is modeled as a linear combination of its own lags and the lags of other variables.

Bayesian Structural Time Series (BSTS): BSTS is a Bayesian approach for modeling time series data that includes trend, seasonality, and other relevant factors. It is a flexible and powerful modeling framework that can capture complex patterns in the data.

Exponential Smoothing (ETS): ETS models are used for time series data that exhibit trend and seasonality. These models include a smoothing parameter that determines how much weight is given to past observations.

Auto-Regressive Conditional Heteroskedasticity (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH): ARCH and GARCH models are used to model volatility clustering, which is a common feature in financial time series. These models allow for the conditional variance of the time series to vary over time.

State Space Models (SSM): SSM is a general modeling framework that can be used to model a wide range of time series data. It is a flexible and powerful approach that can incorporate multiple sources of uncertainty and can handle missing or irregularly spaced data.

These variations can provide greater flexibility in modeling different types of time series data, and can help capture specific patterns and features of the data that may not be well suited for the standard ARIMA model.

SARIMA, or Seasonal Autoregressive Integrated Moving Average, is a variation of the ARIMA model that incorporates seasonality in the time series.

The main difference between SARIMA and ARIMA is that SARIMA includes additional parameters for seasonal components. In ARIMA, we use the p, d, and q parameters to model the non-seasonal components of the time series, while in SARIMA, we use additional parameters for the seasonal components of the time series.

In SARIMA, the parameters are denoted by $(p, d, q) \times (P, D, Q)m$, where m is the number of time periods per season. The parameters (P, D, Q) represent the seasonal AR, differencing, and MA components, respectively.

In essence, SARIMA allows us to model the seasonal behavior of the time series, which can be particularly useful for forecasting and identifying patterns in data that vary periodically.

VAR, or Vector Autoregression, is a statistical model used in time series analysis to capture the relationship between multiple variables.

Unlike ARIMA models, which focus on modeling a single time series, VAR models consider the joint behavior of multiple time series variables. The idea behind VAR is to model each variable as a linear function of its past values and the past values of the other variables in the system. The model assumes that each variable in the system is affected by the past values of all variables in the system, and that the variables are mutually dependent.

In a VAR model, we start by specifying the number of lags to include for each variable in the system. We then estimate the coefficients of the model using maximum likelihood or another estimation method. The resulting model can be used to forecast the values of each variable in the system, as well as to analyze the causal relationships between the variables.

VAR models are useful for analyzing and predicting the behavior of complex systems with multiple interdependent variables, such as macroeconomic indicators, financial markets, and weather patterns. By capturing the dynamic relationships between variables over time, VAR models can provide insights into the underlying structure and causal relationships of the system, as well as help to identify key drivers of behavior and potential areas for intervention.

We can combine other independent variables with time series data to create a model that accounts for the effects of both time and other factors on the outcome variable of interest. This type of model is known as a dynamic regression model or an **ARIMAX** model (ARIMA with eXogenous variables).

The general form of a dynamic regression model is:

$$Y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \varepsilon_t$$

where Y_t is the value of the outcome variable at time t, $X_{1,t}, X_{2,t}, \ldots, X_{k,t}$ are the values of the k independent variables at time t, $\beta_1, \beta_2, \ldots, \beta_k$ are the corresponding coefficients, α is a constant term, and ε_t is the error term.

The dynamic regression model allows us to investigate the relationship between the outcome variable and the independent variables while accounting for the effects of past values of the outcome variable and other time-related factors captured by the ARIMA model.

Including other independent variables in a time series model can improve the accuracy of the forecasts by accounting for other factors that may affect the outcome variable. However, it is important to carefully select the independent variables and consider potential confounding factors to avoid bias and overfitting.

BSTS (Bayesian Structural Time Series) is a flexible and powerful time series modeling framework that allows for the inclusion of multiple seasonalities, trends, and other explanatory variables. The model is based on Bayesian statistics, which allows for the inclusion of prior knowledge about the parameters and hyperparameters of the model.

The basic structure of a BSTS model consists of three main components:

Local level component: This component captures the short-term fluctuations in the time series and is modeled using a random walk with Gaussian errors. The variance of the errors can be set to be time-varying, allowing the model to capture changes in volatility over time.

Seasonal component: This component captures the seasonal patterns in the time series and is modeled using Fourier terms. The number of terms can be adjusted to capture multiple seasonal patterns, such as weekly and annual cycles.

Trend component: This component captures the long-term trends in the time series and is modeled using a polynomial function of time. The degree of the polynomial can be adjusted to capture different types of trends, such as linear, quadratic, or cubic.

In addition to these main components, a BSTS model can also include additional explanatory variables, such as economic indicators or weather data, as well as intervention variables to capture the effects of specific events or policy changes.

The BSTS model is estimated using Bayesian methods, which involve specifying prior distributions for the parameters and hyperparameters of the model and updating these distributions based on the observed data using Markov Chain Monte Carlo (MCMC) methods. The posterior distributions of the parameters and hyperparameters are then used to generate probabilistic forecasts of future values of the time series.

Overall, the BSTS model is a powerful and flexible framework for modeling time series data, allowing for the incorporation of multiple seasonalities, trends, and other explanatory variables while accounting for the uncertainty in the model parameters and forecasts.

It is closely related to gaussian process regression.

ETS stands for **Exponential Smoothing State Space Model**, and it is a popular method used for time series forecasting.

ETS models use a state space framework to represent a time series as the sum of different components, such as level, trend, and seasonality. The model updates the estimates of these components using exponential smoothing techniques, which weigh recent observations more heavily than past observations.

There are several variations of ETS models, each of which incorporates different components and smoothing parameters. For example, the "$ETS(A, Ad, N)$" model includes additive error, additive trend, and no seasonality components, while the "$ETS(M, M, M)$" model includes multiplicative error, multiplicative trend, and multiplicative seasonality components.

Once an ETS model is fit to a time series, it can be used to forecast future values of the series. The accuracy of the forecast depends on the specific ETS model chosen, as well as the quality and predictability of the data being forecasted.

ARCH (Autoregressive Conditional Heteroskedasticity) and **GARCH (Generalized Autoregressive Conditional Heteroskedasticity)** models are used to model and forecast time series data with changing volatility over time.

ARCH models assume that the variance of the error term in a time series is a function of the past error terms. Specifically, it assumes that the variance of the error term in the current time period depends on the square of the error term from the previous time period. In other words, if the previous error was large, the current error is expected to be large as well.

GARCH models extend ARCH models by allowing for the conditional variance to depend not only on past error terms but also on past variances. This allows for more flexibility in modeling volatility patterns in time series data.

To estimate ARCH and GARCH models, one can use maximum likelihood estimation. Once the models are estimated, they can be used to forecast future values of the time series, taking into account the changing volatility patterns.

ARCH and GARCH models are commonly used in finance and economics to model the volatility of asset returns, such as stock prices or exchange rates. They can also be used in other fields to model time series data with changing volatility over time, such as in climate science or engineering.

State space modeling (SSM) is a statistical approach used for analyzing time series data. It models a system that is not directly observable but can be observed through a set of measurements. The underlying system is represented by a set of unobservable states, and a set of equations defines how the system evolves over time. The observations are related to the states by a set of observation equations.

In SSM, the goal is to estimate the unobservable states and parameters of the model that best explain the observed data. This is done by using a Bayesian framework that combines the information from the prior distribution, which represents our knowledge about the states and parameters before seeing the data, and the likelihood function, which represents the probability of observing the data given the states and parameters.

The estimation of the model parameters and the unobservable states is performed using the Kalman filter and the Kalman smoother algorithms. The Kalman filter is used to compute the posterior distribution of the states given the observed data, while the Kalman smoother is used to compute the posterior distribution of the states and the parameters given the observed data.

SSM is a powerful framework for time series modeling because it can handle a wide range of time series data, including those that exhibit nonlinear and non-Gaussian behavior. It is used in many fields, including finance, economics, engineering, and physics, to model and forecast time series data.

The Kalman filter is a mathematical algorithm used to estimate the state of a system, based on a sequence of noisy measurements. It is commonly used in time series modeling, control systems, robotics, and signal processing.

The Kalman filter works by recursively updating an estimate of the state of a system based on measurements of the system, taking into account both the uncertainty in the measurements and the dynamics of the system. The filter is able to incorporate information from previous measurements and predictions to produce a more accurate estimate of the current state.

The Kalman filter is based on the principle of Bayesian inference, which allows for the incorporation of prior knowledge and uncertainty into the estimation process. The filter updates the estimate of the state of the system and the associated uncertainty based on each new measurement, using a set of equations that take into account the dynamics of the system, the measurement noise, and the uncertainty in the previous estimate.

One of the key advantages of the Kalman filter is that it is able to provide an estimate of the state of the system even when the measurements are noisy or incomplete. It can also handle situations where the underlying system is nonlinear, by using an extended version of the filter known as the extended Kalman filter. However, the Kalman filter does assume that the system is linear and that the noise in the measurements is Gaussian and has a constant variance.

Resources:

1. https://www.educba.com/arima-model-in-r/
2. https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/
3. https://otexts.com/fpp2/arima-r.html
4. https://jtr13.github.io/cc19/time-series-modeling-with-arima-in-r.html
5. https://neptune.ai/blog/arima-vs-prophet-vs-lstm
6. https://www.geeksforgeeks.org/exponential-smoothing-in-r-programming/
7. https://medium.com/analytics-vidhya/a-complete-introduction-to-time-series-analysis-with-r-sarima-models-ff86d526d1d7
8. https://www.analyticsvidhya.com/blog/2021/11/basic-understanding-of-time-series-modelling-with-auto-arimax/
9. https://cran.r-project.org/web/packages/bsts/bsts.pdf
10. https://talksonmarkets.files.wordpress.com/2012/09/time-series-analysis-with-arima-e28093-arch013.pdf
11. https://mfe.baruch.cuny.edu/wp-content/uploads/2014/12/TS_Lecture5_2019.pdf