Module 1.1
Data Mining and Time Series Analysis - TSA

S. Lakshmivarahan

School of Computer Science
University of Oklahoma
Norman, OK - 73019, USA

Data Mining and Time Series Analysis

- Time series analysis (TSA) is a part of a bigger and growing area of Data Mining (DM)
- Data Mining is a process of extracting the structure/patterns that are inherent in the data/observations.
- These inherent patterns offer clues relating to the data generating process
- The ultimate aim of DM is to quantify this data generating process using models of various kinds

Early Examples of DM - Astronomy

- Astronomy provides some of the spectacular examples of early practices in DM
  - Copernicus (1473-1543)
  - Galileo (1544-1642)
  - Kepler (1571-1630)
  - Newton (1643-1727)
- Discovery of physical/causal laws - Kepler's laws, Newton's laws, are examples of the fruits of DM
- Once laws/models are available, prediction becomes possible.
- Prediction of lunar/solar eclipses, ocean tides on full moon days, etc.

Abundance of Data - Bid Data

- Volume of data collected doubles in every 2–3 years
- Thanks to the advances in technology
  - Large and fast computers
  - Large scale storage devices
  - Communication technology
  - Sensor technology
- Data arises in various shapes and forms

Spatio-Temporal distribution of data

- Record of hourly temperature in major cities around the world
- Record of monthly employment across different industrial sectors and across each of the 50 states in the USA
- Record of annual rainfall across all parts of the globe

Cross-Sectional Data

- A slice of the spatio-temporal distribution at a given time is called cross-sectioned data
- Distribution of the number of employees by industrial sector across each state in the month of December, 2016
- Distribution of drought across the globe as of the first of the year 2017

Observations at a given location - Time Series

- Record of hourly temperature at the World Trade Center in New York City
- Record of daily (global) exchange rate between US dollars and British Pound
- Record of daily Microsoft stock prices.
- Record of hourly average wind speed at a given location

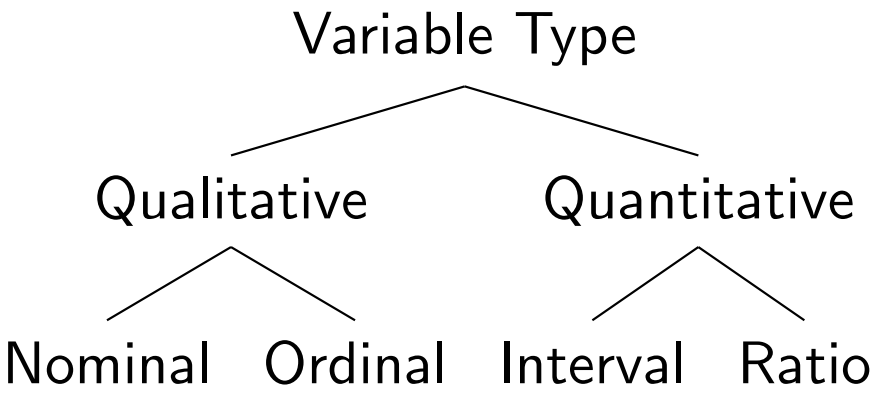
Data in Matrix forms

- A set of n objects represented on a common set of m attributes/variables in a matrix form:

$$D = \begin{array}{c} \\ 1 \\ 2 \\ i \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \cdots & j & \cdots & n \end{array} \\ \begin{pmatrix} & & & \vdots & & \\ & & & \vdots & & \\ \cdots & \cdots & \cdots & d_{ij} & \cdots & \cdots \\ & & & \vdots & & \\ & & & \vdots & & \end{pmatrix} \end{array} \begin{array}{ll} n: & \text{objects} \\ m: & \text{attributes/} \\ & \text{variables} \end{array} \tag{1}$$

- d_{ij} is the i^{th} attribute of the j^{th} object $1 \leq i \leq m$, $1 \leq j \leq n$
- Attributes: Height, weight, age, gender, education level, occupation, salary range, etc.
- Goal is to classify objects based on similarity/correlation between attributes

Scales for variables - A classification



Qualitative - Nominal Scale

- Nominal type variables take on a finite set of values
  - True/False - logical
  - Male/Female - gender
  - Colors of a rainbow - V I B G Y O R
- Allowed operations: Check for equality
  - $x_1 = x_2$
  - $x_1 \neq x_2$

Qualitative - Ordinal Scale

- Ordinal type variables take on a finite set of values
- Grades in a class: A, B, C, D, F
- Rating on a scale of 1-10, 10 being the best
- Besides equality, ordering is allowed
  - $x_1 = x_2$
  - $x_1 \neq x_2$
  - If $x_1 \neq x_2$, then $x_1 > x_2$ or $x_1 < x_2$
- Example:
  A is better than B,
  5 is not as good as 7

Quantitative Type - Interval Scale

- Interval scale variables can take continuous values
- Besides equality and ordering, these allow the arithmetic operations - differencing
- If $x_1 > x_2$, then $x_1 - x_2$ is the difference between them
- If $x_1 = 30\,^{\circ}\mathrm{C}$, $x_2 = 10\,^{\circ}\mathrm{C}$, $x_1 - x_2 = 20\,^{\circ}\mathrm{C}$
- There is no fixed origin for this scale

Qualitative type - Ratio Scale

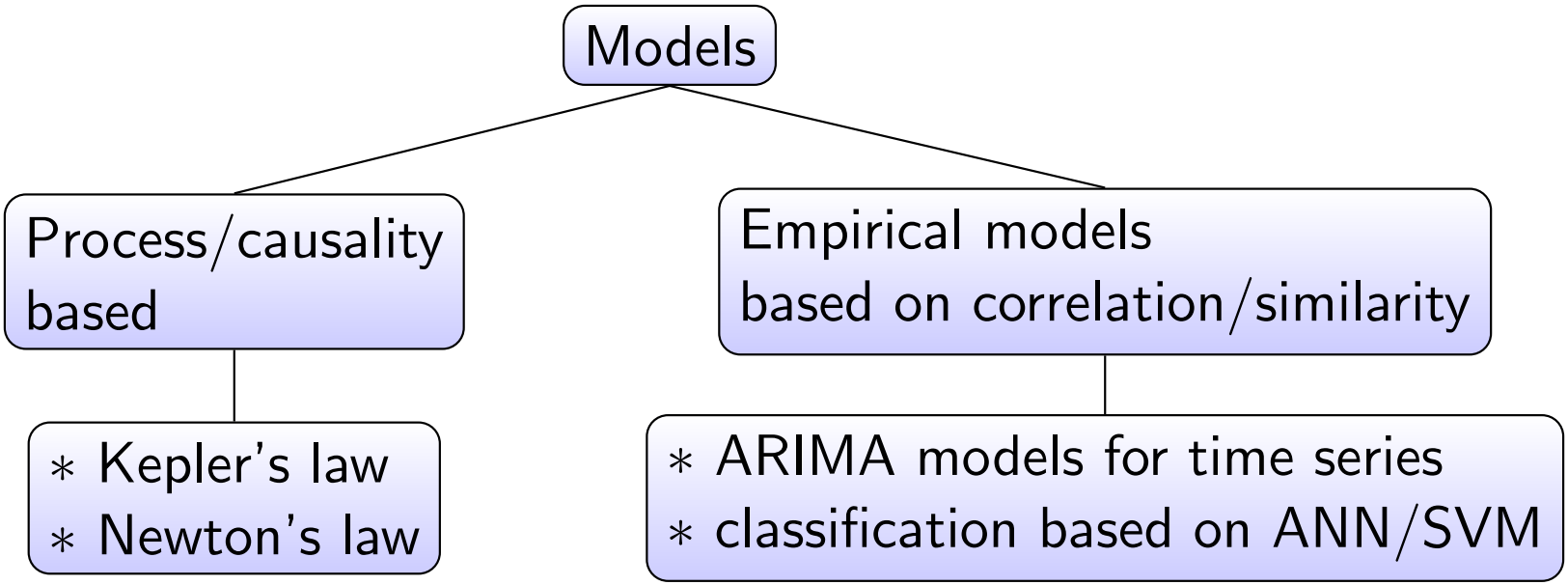
- It allows continuous values
- In addition to equality checking, ordering and differencing, it allows for taking ratios as well
- If $x_1 > x_2$, then $\frac{x_1}{x_2}$ has meaning
- Examples: salary, absolute temperature, pressure, wind speed, rainfall, price of a stock, etc.,

Modeling Time Series - Ratio Scale - Scope

- The technique for analysis of data differ widely with the type of scales
- In this course we will only study modeling of TS of data in ratio scale

Modeling in DM

- Ultimate goal of DM is prediction
- To predict we need models that capture the causality or correlation implied by the data



Model classification

- Irrespective of their origin, models can be classified along different dimensions
  - Deterministic or stochastic
  - Static or dynamic
  - Continuous or discrete time
  - Linear or nonlinear
- TSA deals with the development and analysis of stochastic, dynamic, discrete time linear and nonlinear models

Pathway to Modeling in TSA - Model Selection

- Compute and plot the inherent correlation structure in the given TS
- Compare the correlation structures corresponding to different types of models derived from the ARIMA family
- This step corresponds to pattern recognition and is subjective
- The output of this step is a small number of potential models whose theoretical correlation structures matche closely with that being observed in the given data set

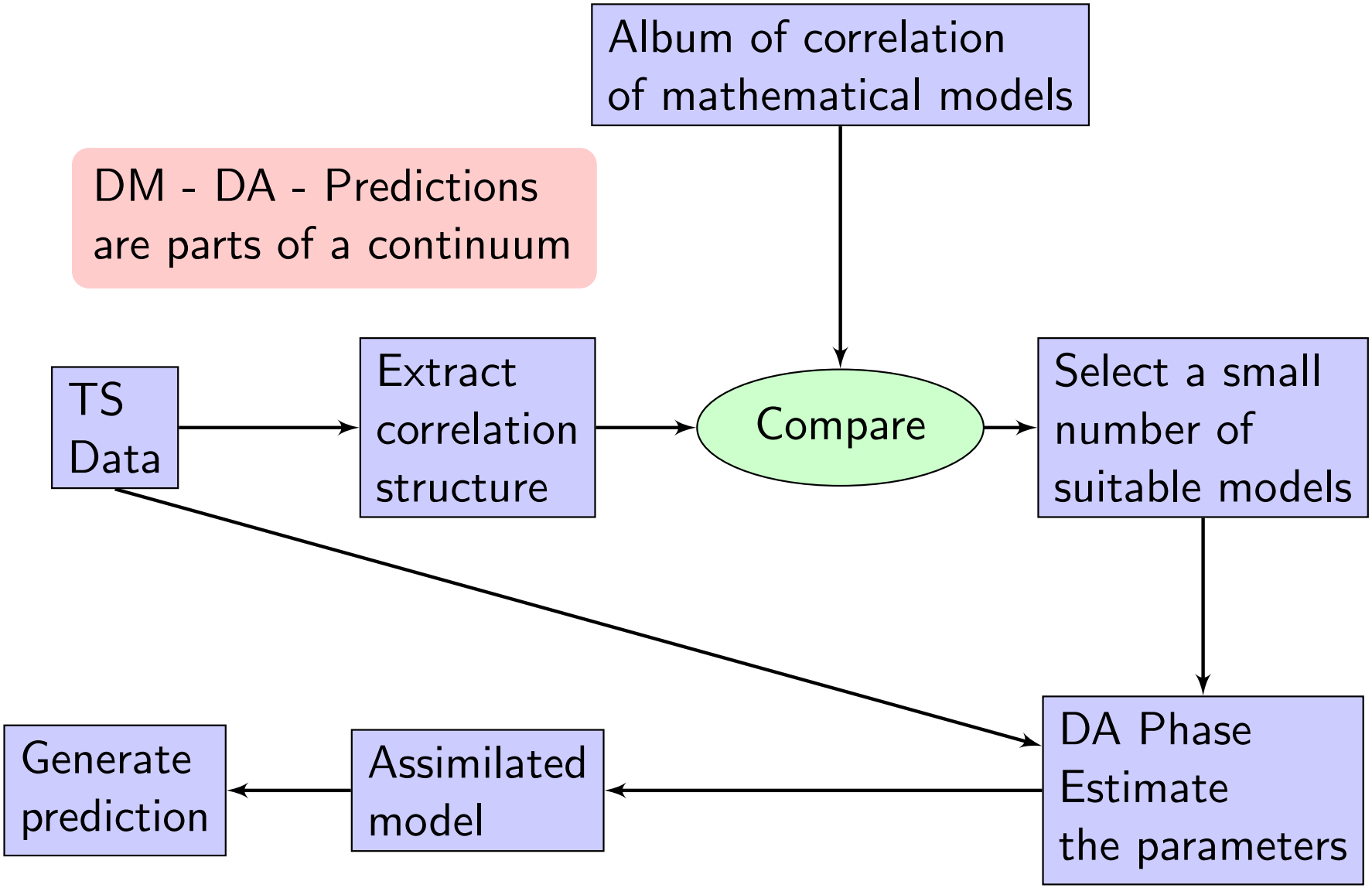
Pathway to Modeling in TSA - Data Assimilation

- Each of the selected model have a number of unknown parameters
- We then use the very same data that was used in the model selection, to estimate the unknown parameters by *fitting the models to data*
- This aspect of fusing data with model is called Data Assimilation (DA) step and uses well known statistical estimation techniques.

Pathway to Modeling - Prediction

- Once an assimilated model is made available, we are then ready to generate prediction
- Since the models are stochastic, perfect prediction is not possible
- In addition to predicting the level, also need to quantify the uncertainty in prediction as measured by the variance.

A Pictorial View



Prerequisites

- Probability Theory - Refer to Appendix
- Statistics - Estimation Theory, Hypothesis Testing - developed as needed
- Matrix Theory - introduced as needed
- Programming - MATLAB or R

References

- Brockwell, Peter J and Davis, Richard A "Time series: theory and methods" *Springer* (2013), MS/Ph.D level text
- Hamilton, James D. (1995),"Time series analysis." Princeton University. MS/Ph.D level text
- Fuller, Wayne A "Introduction to statistical time series",*John Wiley & Sons* (2009), (Second Edition), MS/Ph.D level text

Time Series Data Sources

- http://www.qlik.com/us/products/data-market provides examples of time series from various domains
- In addition you may consult Department of Labor statistics for more data