ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# STATISTICAL FORECASTING BASED ON CONDITIONAL DENSITY ESTIMATION

[1]RAVINDRA CHANGALA, [2]K.KIRAN KUMAR,[3] S. RENU DEEPTI,[4] V. PREETHI, [5]G. SUSHMA
[6]KIRAN KUMAR KAVETI, [7]NATHA DEEPTHI

[1]Associate Professor, Dept of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana.

[2]Assistant Professor,Dept of FE,Prasad V Potluri Siddhartha Institute of Technology, Vijayawada

[3]Assistant Professor, Dept of IT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad

[4]Asst.Professor ,Dept.of ECE,Aditya University, Surampalem, india

[5]Sr.Asst Professor, Dept of CSE,CVR College Of Engineering, Hyderabad ,Telangana

[6]Asst Professor,Dept of CSE,Vignan's Foundation for Science,Technology and Research, Guntur,AP

[7]Asst Professor,Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist.,

E-mail: changalaravindra@gmail.com

## ABSTRACT

Many real problems such as stock market prediction, weather forecasting etc has inherent randomness associated with them. Adopting a probabilistic framework for prediction can accommodate this uncertain relationship between past and future. Typically the interest is in the conditional probability density of the random variable involved. One approach for prediction is with time series and auto regression models. In this work, liner prediction method and approach for calculation of prediction coefficient are given and probability of error for different estimators is calculated. The existing techniques all require in some respect estimating a parameter of some assumed solution. So, an alternative approach is proposed. The alternative approach is to estimate the conditional density of the random variable involved. The approach proposed in this thesis involves estimating the (discretized) conditional density using a Markovian formulation when two random variables are statistically dependent, knowing the value of one of them lets us get a better estimate of the value of the other one. The conditional density is estimated as the ratio of the two dimensional joint density to the one-dimensional density of random variable whenever the later is positive. Markov models are used in the problems of making a sequence of decisions and problem that have an inherent temporality that is consisting of a process that unfolds in time in time. In the continuous time Markov chain models the time intervals between two consecutive transitions may also be a continuous random variable. The Markovian approach is particularly simple and fast for almost all classes of classes of problems requiring the estimation of conditional densities.

**Keywords:** *Statistical Prediction, Unbiased Ness, Sufficiency, Smoothing, Univariate Time Series, Autoregressive, Markov Chains.*

## 1. INTRODUCTION

**1.1Statistics:** Statistics are measurements, enumerations or estimates of natural phenomenon, usually systematically arranged, analyzed and presented as to exhibit important inter-relationships[1]

among them. Modern statistics [2] refers to a body of methods and principles that have been developed to handle the collection, description, summarization and analysis of numerical data.statistical theory, a "statistic" is a well-behaved function of the data. A statistic is *sufficient* [3] if it is just as informative as the full data. In many applications it is not unusual to have dozens or hundreds of parameters and thousands of training samples. A sufficient statistic is a function ' $s$ ' of the samples ' $D$ ' that contains all the information relevant to estimating some parameter '$\theta$'. A fundamental theorem concerning sufficient statistics is the Factorization theorem (8) which states that's' is sufficient for '$\theta$' if and only if $p(D/\theta)$ can be factored into the product of two functions: one depending only on ' $s$ 'and ' $\theta$ ', the other depending only on training samples.In applying statistics to a scientific, industrial, or societal problem, one begins with a process to be studied[4]. This might be a population of people in a country, of crystal grains in a rock, or of goods manufactured by a particular factory during a given period. It may instead be a process observed at various times what is called a time series. For practical reasons, rather than compiling data about an entire process, one usually instead studies a chosen subset of the process, called a sample [5]. Data are collected about the sample in an observational or experimental setting. The data are then subjected to statistical analysis, which serves two related purposes [6]: description and inference. Descriptive statistics can be used to summarize the data, either numerically or graphically, to describe the sample. Basic examples of numerical descriptors include the mean and standard deviation. Graphical summarizations include various kinds of charts and graphs. Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences [7]. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation), prediction of future observations, descriptions of association (correlation), or modeling of relationships (regression). Other modeling techniques include ANOVA, time series, and data mining [8]. A major problem lies in determining the extent to which the chosen sample is representative. Statistics offers methods to estimate and correct for randomness in the sample and in the data collection procedure, as well as methods for designing robust experiments in the first place[9].

**1.2 Statistical prediction:** A *prediction* or *forecast* is a statement or claim that a particular event will occur in the future. Usually, it depends on one of two prerequisites whether tests for quantitative trends are applied or not. First, the independent variable is quantitative, and second, the independent variable is quantitative and a particular quantitative trend hypothesis is to be tested. In the first case, the experimenter does not proceed from certain expectations; the experimenter just looks for the best functional description of the data. In the second case, however, the data are examined as to their compatibility with predictions derived from a certain theory. A null hypothesis ($H_0$) is any statistical hypothesis (6) which comprises one of the signs '=', '≤', or '≥' and which is testable by a given statistical test. Its opposite is an alternative hypothesis ($H_1$), which usually is complementary to the $H_0$ and against which the test is performed[10]. If the statistical prediction is not equivalent to a single testable $H_0$ or $H_1$, there are basically two options: either to perform a less well suited test and interpret the 'apparent' empirical relations among the sample statistics, or to apply more than one test[11]. The problem of parameter estimation is a classical one in statistics and it can be approached in several ways. The common approaches are maximum likelihood estimation [12] and Bayesian estimation.

**1.3 Parameter estimation:** Considering a random sample $x_1, x_2, ...., x_n$ of size '$n$' with probability function $f(x : \theta_1, \theta_2, ...., \theta_k)$ where $\theta_1, \theta_2, ....., \theta_k$ are the unknown parameters. Then, there will always be an infinite number of functions of sample values called statistics, which may be proposed as estimates of one or more of the parameters. Evidently, the best estimate would be the one that falls nearest to the true value of the parameter to be estimated i.e.; the statistic whose distribution concentrates as closely as possible near the true value of parameter is regarded the best estimate. The basic problem is to determine the functions of sample observations. The estimating functions are called estimators [13]. A good estimator needs to satisfy some characteristics[14]:

2. **Unbiased ness**: An estimator $T_n = T(x_1, x_2, ...., x_n)$ is said to be unbiased estimator of $y(\theta)$) if $E(T_n) = y(\theta)$, for all $\theta \in \Theta$, parameter space

3. **Consistency:** estimator $T_n = T(x_1, x_2, ...., x_n)$ , based on random sample of size $n$ , is said to be consistent estimator of $y(\theta), \theta \in \Theta$ , the parameter space, if $T_n$ converges to $y(\theta)$ in probability, i.e. If $T_n \xrightarrow{p} y(\theta)$ as $n \to \infty$

4. **Efficiency**: If in a class of consistent estimators for a parameter, there exists one whose sampling variance is less than that of any such estimator, it is called the most efficient estimator. Whenever such an estimator exists [15].

5. **Sufficiency**: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter [16].

**1.4 Non-parametric estimation:** In the parametric tests, the functional forms from which the samples are drawn is assumed to be known and are concerned with testing statistical hypothesis about the parameters of the function or estimating its parameters. On the other hand, a non-parametric estimation [17] does not depend on the particular functional form from which the samples are drawn i.e., no assumptions are made regarding the functional form [18].

**1.5 Applications of Prediction:** In mathematical finance, deterministic mathematical models of stock market behavior are unreliable in predicting future behavior, because of various unknown factors that can affect the market trends. As an alternative, a statistical prediction problem can be formulated for the pertinent and classified commodities in the stock, and the required parameters involved in the functions of the prediction model or any non-parametric objects in the prediction model can be estimated from the data collected over long periods of time. In the least case, the trends can be predicted with reasonable confidence [19]. Quantum physics is an unusual field of science because it enables scientists to make predictions on the basis of probability In microprocessors, branch prediction permits to avoid pipeline emptying at microcode branching. Engineering is a field that involves predicting failure and avoiding it through component or system redundancy. Some fields of science are notorious for the difficulty of accurate prediction and forecasting, such as software reliability, natural disasters, pandemics, demography, population dynamics and meteorology [20].

6. **Other Fields** Statistical prediction is a major concern in areas such as machine learning, pattern recognition, neural networks, signal processing, computer vision and feature extraction [21]. It offers a flexible way to investigate the properties of a given data set and provides a solid basis for efficient data mining tools. It is crucial in unsupervised learning tasks and Bayesian inference and classification. It is often used in clustering of data in large databases [22].

**1.6 Paper Organization:** The rest of the paper is organized as: Chapter 2 gives details about some of the available time series models and auto regression models for statistical prediction. Chapter 3 gives brief introduction about Markov chains, their properties, Markov chains in discrete state space and continuous-time Markov chains. Hidden Markov models are also briefly discussed. Chapter 4 deals with linear prediction: Predictions of future sample, calculation of prediction coefficients in such a way to minimize prediction error are given. Chapter 5 mainly focuses on the Markov chain method for the estimation of conditional density. Chapter 6 shows the results: Actual samples and predicted samples are compared in linear prediction. Chapter 7 gives the conclusion of the paper and future work that can be done.

## 2.Time series and autoregressive models

**2.1 Time series prediction:** Most statistical forecasting methods are based on using historical data from a time series. A time series is a series of observations over time of some quantity of interest (a random variable).Thus, if $X_t$ is the random variable of interest at time $i$ , and if observations are taken at times $i = 1,2,...,t$ , then the observed values $\{X_1 = x_1, X_2 = x_2,......, X_t = x_t\}$ are a time series. The time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day[23]. A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information. The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series[24]

$$\hat{y}_{t+1} = f(y_t, y_{t-1},......, y_{t-M+1})$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem. Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult[25].

### 2.2 Input selection strategies

Input selection is an essential pre-processing stage to guarantee high accuracy, efficiency and scalability in problems such as machine learning, especially when the number of observations is relatively small compared to the number of inputs. It has been the subject in many application domains like pattern recognition, process identification, time series modeling and econometrics. Problems that occur due to poor selection of input variables are:

If the input dimensionality is too large, the 'curse of dimensionality' problem may happen. Moreover, the computational complexity and memory requirements of the learning model increase. Additional unrelated inputs lead to poor models (lack of generalization).Understanding complex models (too many inputs) is more difficult than simple models (less inputs), which can provide comparable good performances.

### 2.3 Forecasting methods for a constant level model

**2.3.1 Last- value forecasting method:** The last–value forecasting method sometimes is called the naïve method, because statisticians consider it naïve to use just a sample size of one when additional relevant data are available. By interpreting '$t$ 'as the current time, the last-value forecasting procedure uses the value of the time series observed at time $t$ $\left(x_t\right)$ as the forecast at time $t+1$ .Therefore, $F_{t+1} = x_t$ This forecasting procedure has the disadvantage of being precise i.e., its variance is large because it is based upon a sample of size one.

**2.3.2 Averaging Forecasting method:** Instead of using just a sample size of one, this method uses all the data points in the time series and simply averages these points. Thus, the forecast of what the next data point will turn out to be is

$F_{t+1} = \sum_{i=1}^{t} x_i / t$ this estimate is an excellent one if the process is entirely stable

**2.3.3 Moving average forecasting method:** Rather than using very old data that may no longer be relevant, this method averages the data for only the last '$n$' periods as the forecast for the next period i.e.

$F_{t+1} = \sum_{i=t-n+1}^{t} x_i / n$ This forecast is easily updated from period to period.

**2.3.4 Exponential smoothing forecasting method:**

This method overcomes the drawback of moving average method. This method uses the formula,

$$F_{t+1} = \alpha x_t + \left(1-\alpha\right)F_t$$

Where '$\alpha$' $\left(0 < \alpha < 1\right)$ is called the smoothing constant. Thus, the forecast is just a weighted sum of the last observation $x_t$ and the preceding forecast $F_t$ for the period just ended. Because of this recursive relationship between $F_{t+1}$ and $F_t$, alternatively $F_{t+1}$ can be expressed as

$$F_{t+1} = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \ldots\ldots$$

**2.4 Common approach to univarity Time series:** There are a number of approaches to modeling time series. One approach is to decompose the time series into a trend, seasonal, and residual component. Triple exponential smoothing is an example of this approach. When the data show trend and seasonality (sometimes called periodicity) then triple exponential smoothing is used.

The basic equations for the method are given by,

**Overall smoothing**:

$$S_t = \alpha Y_t / I_{t-L} + \left(1-\alpha\right)\left(S_{t-1} + b_{t-1}\right)$$

**Trend smoothing**:

$$b_t = \gamma\left(S_t - S_{t-1}\right) + \left(1-\gamma\right)b_{t-1}$$

**Seasonal smoothing**:

$$I_t = \beta Y_t / S_t + \left(1-\beta\right)I_{t-L}$$

**Forecast**:

$$F_{t+m} = \left(S_t + mb_t\right)I_{t-L+m}$$

Where, $y$ is the observation, Sis the smoothed observation , $B$ is the trend factor $I$ is the seasonal index, $F$ is the forecast at $m$ periods ahead , $T$ is an index denoting a time period , and $\alpha, \beta, \gamma$ are constants that must be estimated in such a way that the MSE of the error is minimized. **Initial values for the trend factor**: The general formula to estimate the initial trend is given by

$$b = 1/L\left(\left(\left(Y_{L+1} - Y_1\right)/L\right) + \left(\left(Y_{L+2} - Y_2\right)/L\right) + \ldots + \left(\left(Y_{L+L} - Y_L\right)/L\right)\right)$$

### 2.5 Autoregressive models for linear prediction:

The autoregressive model is one of a group of linear prediction formulas that attempt to predict an output $y[n]$ of a system based on the previous outputs $(y[n-1], y[n-2], .....)$ and inputs $(x[n], x[n-1], x[n-2], ......)$ .Deriving the linear prediction model involves determining the coefficients $a_1, a_2, ......$ and $b_1, b_2, ......$ in the equation: $y_e[n](estimated)$

$$= a_1 * y[n-1] + a_2 * y[n-2] + ...... + b_0 * x[n] + b_1 * x[n-1] + ......$$

**2.5.1 Autoregressive model:**The notation AR $(p)$ refers to the autoregressive model of order $p$ . The AR $(p)$ model is written as $X_t = c + \sum_{i=1}^{p} \theta_i X_{t-i} + \varepsilon_t$ Where $\theta_1, \theta_2, ...., \theta_p$ are the parameters of the model, '$c$' a constant and $\varepsilon_t$ is an error term.

***2.5.2 Moving average model: The** notation MA $(q)$ refers to the moving average model of order $q$ :*

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where the $\theta_1, \theta_2, ......, \theta_q$ are the parameters of the model and they $\varepsilon_t, \varepsilon_{t-1}, .....$ are the error terms. The moving average model is essentially a finite impulse response filter with some additional interpretation placed on it.

**2.5.3 Autoregressive moving average model***: The notation ARMA $(p, q)$ refers to the model with $p$ autoregressive terms and $q$ moving average terms. This model contains the AR$(p)$ and MA$(q)$ models,* $X_t = \varepsilon_t + \sum_{i=1}^{p} \theta_i X_{t-i}$

$$+ \sum_{i=1}^{q} \theta_i X_{t-i}$$

### 2.5.4 Calculation of the AR parameters:

The AR $(p)$ model is given by the equation $X_t = \sum_{i=1}^{p} \theta_i X_{t-i} + \varepsilon_t$ It is based on parameters $\theta_i$ where $i = 1, 2, ...., p$ . Those parameters may be calculated using Yule-Walker equations: $\gamma_m = \sum_{k=1}^{p} \theta_k \gamma_{m-k} + \sigma_\varepsilon^2 \delta_m$ Where, yielding $p+1$ equations. $\gamma_m$ is the autocorrelation function of $X$ , $\sigma_\varepsilon$ is the standard deviation of the input noise process, and $\delta_m$ is the Kronecker delta function. Because the last part of the equation is non-zero only if $m = 0$ , the equation is usually solved by

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ . \\ . \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_{-1} & \gamma_{-2} & \cdots \\ \gamma_1 & \gamma_0 & \gamma_{-1} & \cdots \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots \\ \dots & \dots & \dots & .... \\ \dots & \dots & \dots & .... \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ . \\ . \end{bmatrix}$$

Representing it as a matrix for $m > 0$, thus getting equation solving all $\theta$.For $m = 0$,

$\gamma_0 = \sum_{k=1}^{p} \theta_k \gamma_{-k} + \sigma_\varepsilon^2$ allows us to solve $\sigma_\varepsilon^2$ .

**2.5.5 Approaches for modeling univariate time series:** A common approach for modeling univariate time series is the autoregressive (AR) model. An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of $p$ is called the order of the AR model. AR models can be analyzed with one of various methods, including standard linear least squares techniques. They also have a straightforward interpretation

$$X_t = \delta + \theta_1 X_{t-1} + \theta_2 X_{t-2} + ...... + \theta_p X_{t-p} + A_t$$

Where $X_t$ is the time series, $A_t$ is white noise, and

$\delta = \left(1 - \sum_{i=1}^{p} \theta_i\right) \mu$ with $\mu$ denoting the process mean. Another common approach for modeling univariate time series models is the moving average (MA) model:

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots\ldots.. - \theta_q A_{t-q}$$

Where $X_t$ is the time series, $\mu$ is the mean of the series, $A_{t-i}$ are white noise, and $\theta_1, \theta_2, \ldots \theta_q$ are the parameters of the model. The value of $q$ is called the order of the MA model.

**2.5.6 Box Jenkins method:** The first step in developing a Box-Jenkins model is to determine if the series is stationary and if there is any significant seasonality that needs to be modeled. Seasonality (or periodicity) can usually be assessed from an autocorrelation plot, a seasonal sub series plot, or a spectral plot (9). Box and Jenkins recommend the differencing approach to achieve stationary. However, fitting a curve and subtracting the fitted values from the original data can also be used in the context of Box-Jenkins models.

## 3.Markov chain Models

**3.1 Markov chains:** In mathematics, a Markov chain is a discrete-time stochastic process with the Markov property named after Andrey Markov. In such a process, the previous states are irrelevant for predicting the subsequent states, given knowledge of the current state. A Markov chain describes at successive times the states of a system. At these times the system may have changed from the state it was in the moment before to another or stayed in the same state. The changes of state are called transitions. The Markov property means the system is memoryless, i.e. it does not "remember" the states it was in before, just "knows" its present state, and hence bases its "decision" to which future state it will transit purely on the present, not considering the past.N$^{th}$ order Markov chain: A Markov process moves from state to state depending only on the previous observations. In an nth order Markov model, the probability of observation depends on the previous n observations.

$0^{th}$ order $\quad \mathrm{P}(\mathrm{x_i})$

$1^{st}$ order $\quad \mathrm{P}(\mathrm{x_i}/x_{i-1})$

$2^{nd}$ order $\quad \mathrm{P}(\mathrm{x_i}/x_{i-1}x_{i-2})$

$n^{th}$ order $\quad \mathrm{P}(\mathrm{x_i}/x_{i-1}x_{i-2}\ldots x_{i-n})$

More generally, the Markov assumption for a n$^{th}$ order model is that $X_i$ depends only on $X_{i-1}X_{i-2}X_{i-3}$….$X_{i-n}$A Markov chain is a sequence $X_1, X_2, X_3, \ldots.$ of random variables with the property (Markov property): the conditional probability distribution (8) of the next future state $X_{n+1}$ given the present and past states is a function of the present state $X_n$ alone, i.e.:

$$\{\Pr(X_{n+1}) = x | X_0 = x_0, X_1 = x_1, \ldots. X_n = x_n\} = \{\Pr(X_{n+1} = x | X_n = x_n)\}$$

The range of the variables i.e., the set of their possible values, is called the *state space*, the value of $X_n$ being the state of the process at time $t$. There are also continuous-time Markov processes.

**3.2 Properties of Markov chains***: The probability of going from state* $i$ *to state* $j$ *in* $n$ *time steps is defined as*

$p_{ij}^{(n)} = \Pr(X_n = j | X_0 = i)$ and the single-step transition as

$p_{ij} = \Pr(X_1 = j | X_0 = i)$ The $n$-step transition satisfies the Chapman-Kolmogorov equation, that for any $0 < k < n$,

$p_{ij}^{(n)} = \sum_{r \in S} p_{ir}^{(k)} p_{rj}^{(n-k)}$ A Markov chain is characterized by the conditional distribution,

$\Pr(X_{n+1}) = x | X_n = y$ which is called the transition probability of the process. This is sometimes called the "one-step" transition probability. The probability of a transition in two, three, or more steps is derived from the one-step transition probability and the Markov property:

$$(\Pr(X_{n+2}) = x | X_n) = \int \Pr(X_{n+2} = x, X_{n+1} = y | X_n)dy$$

$$= \int \Pr(X_{n+2} = x | X_{n+1} = y).\Pr(X_{n+1} = y | X_n)dy$$

These formulas generalize to arbitrary future times $n+k$ by multiplying the transition probabilities and integrating $k-1$ times.

**Marginal distribution:** The marginal distribution $\Pr(X_n = x)$ is the distribution over states at time $n$. The initial distribution is $\Pr(X_0 = x)$. The evolution of the process through one time step is described by

$$\Pr(X_{n+1} = j) = \sum_{r \in S} p_{rj} \Pr(X_n = r) = \sum_{r \in S} p_{rj}^{(n)} \Pr(X_0 = r)$$

the superscript $(n)$ is intended to be an integer-valued label only; however, if the Markov chain is time-stationary, then this superscript can also be interpreted as a "raising to the power of".

Reducibility: A state $j$ is said to be accessible from state $i$ (written as $i \to j$) if, given that we are in state $i$, there is a non-zero probability that at some time in the future, we will be in state $j$. That is, that there exists an $n$ such that

$\Pr(X_n = j | X_0 = i) > 0$ A state $i$ is said to communicate (9) with state $j$ (written $i \to j$) if it is true that both $i$ is accessible from $j$ and that $j$ is accessible from $i$. A set of states $C$ is a communicating class if every pair of states in $C$ communicates with each other.

**Periodicity: A** state $i$ has period $k$ if any return to state $i$ must occur in some multiple of $k$ time steps. For example, if it is only possible to return to state $i$ in an even number of steps, then $i$ is periodic with period $2$. Formally, the period of a state is defined as

$$k = \gcd\{n : \Pr(X_n = i | X_0 = i) > 0\}$$

If $k = 1$, then the state is said to be **aperiodic**; otherwise $(k > 1)$, the state is said to be **periodic with period** $k$. An irreducible Markov chain is said to be **aperiodic**, if its states are aperiodic.

Recurrence: A state $i$ is said to be transient if, given that we start in state $i$, there is a non-zero probability that we will never return back to $i$. Formally, let the random variable $T_i$ be the next return time to state $i$ (the "hitting time"):

$$T_i = \min\{n : X_n = i | X_0 = i\}$$

Then, state $i$ is transient if $T_i$ is not finite with some probability: $\Pr(T_i < \infty) < 1$ If a state $i$ is not transient then it is said to be recurrent or persistent. Although the hitting time is finite, it need not have a finite average. Let $M_i$ be the expected (average) return time, $M_i = E[T_i]$ then, state $i$ is positive recurrent if $M_i$ is finite; otherwise, state $i$ is null recurrent. It can be shown that [cite reference] a state is recurrent if and only if

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$$

Ergodicity: A state $i$ is said to be ergodic if it is aperiodic and positive recurrent. If all states in a Markov are ergodic, the chain is said to be ergodic.

**3.3 Steady state analysis and limiting distributions:** If the Markov chain is a stationary Markov chain, so that the process is described by a single, time-independent matrix $p_{ij}$, then the vector $\prod$ is a stationary distribution if its entries $\prod_j$ sum to $1$ and satisfy

$\prod_j = \sum_{i \in S} \prod_i p_{ij}$ An irreducible chain has a stationary distribution if and only if all of its states are not null-recurrent. In that case, $\prod$ is unique and is related to the expected return time:

$\prod_j = 1/M_j$ Further, if the chain is both irreducible and aperiodic, then for any $i$ and $j$,

$\lim\limits_{n \to \infty} p_{ij}^{(n)} = 1/M_j$ There is no assumption on the starting distribution;

### 3.4 Markov chains in discrete state spaces

If the state space is finite, the transition probability distribution can be represented as a matrix, called the *transition matrix*, with the $(i, j)$ 'th element equal to $p_{ij} = p(X_{n+1} = j | X_n = i)$ For a discrete state space, the integrations in the $k$-step transition probability are summations, and can be computed as the $k$'th power of the transition matrix. That is, if $P$ is the one-step transition matrix, then $P^k$ is the transition matrix for the $k$-step transition. A Markov chain is reversible if there exists an initial distribution $\prod$ such that $\prod_i * p_{ij} = \prod_j * p_{ji}$ ..If the state space is finite, the transition probability distribution can be represented by a matrix, called the *transition matrix*, with the $(i, j)$'th element of $P$ equal to $p_{ij} = \Pr\left(X_{n+1} = j \middle| X_n = i\right)$ $P$ is a stochastic matrix.

**3.5 Continuous-time Markov process**

In probability theory, a continuous-time Markov process is a stochastic process $\{X(t): t \geq 0\}$ that satisfies the Markov property and takes values from amongst the elements of a discrete set called the state space. The Markov property states that at any times $s > t > 0$, the conditional probability distribution of the process at time *s* given the whole history of the process up to and including time $t$, depends only on the state of the process at time $t$. In effect, the state of the process at time $s$ is conditionally independent of the history of the process *before* time $t$, given the state of the process *at* time $t$. one can define a Markov process as follows. Let $X(t)$ be the random variable describing the state of the process at time $t$. Now prescribe that in some small increment of time from $t$ to $t+h$, the probability that the process makes a transition to some state $j$, given that it started in some state $i \neq j$ at time $t$, is given by $\Pr(X(t+h) = j | X(t) = i) = q_{ij} h + o(h),$ where $o(h)$ represents a quantity that goes to zero as $h$ goes to zero (see the article on order notation). Hence, over a sufficiently small interval of time, the probability of a particular transition is roughly proportional to the duration of that interval. Continuous-time Markov processes (8) are most easily defined by specifying the transition rates $q_{ij}$, and these are typically given as the $ij-th$ elements of the transition rate matrix, $Q$ (sometimes called a $Q$-matrix by convention). $Q$ is a finite matrix according to whether or not the state space of the process is finite (it may be countable infinite, for example in a Poisson process where the state space is the non-negative integers). The most intuitive continuous-time Markov processes have *Q*-matrices that are: conservative—the $i$-th diagonal element $q_{ii}$ of $Q$ is given by $q_{ii} = -q_i = -\sum_{j \neq i} q_{ij}$ stable—for any given state $i$, all elements $q_{ij}$ (and $q_{ii}$) are finite.(However, that a $Q$-matrix may be non-conservative, unstable or both.) When the $Q$-matrix is both stable and conservative, the probability that no transition happens in some time $r$ is

$$\Pr(X(t+r) = i | X(s) = i \quad \forall \mathrm{s} \in (\mathrm{t}, \mathrm{t}+\mathrm{r})) = e^{-q_{ir}}$$

**3.5.1 Related processes:** Given that a process that started in state $i$ has experienced a transition out of state $i$, the conditional probability that the transition is into state $j$ is

$q_{ij} / \sum_{k \neq i} q_{ik} = q_{ij} / q_i$ Using these probabilities, the sequence of states visited by the process (the so-called jump process) can be described by a (discrete-time) Markov chain. The *transition matrix* $P$ of the jump chain has elements $p_{ij} = q_{ij} / q_i, i \neq j, p_{ii} = 0$. Another discrete-time process that may be derived from a continuous-time Markov chain is a $\delta$-skeleton—the (discrete-time) Markov chain formed by observing $X(t)$ at intervals of $\delta$ units of time. The random variables $X(0), X(\delta), X(2\delta), \ldots\ldots\ldots$ give the sequence of states visited by the $\delta$-skeleton.

***3.5.2 Embedded Markov chain:***

One method of finding the stationary probability distribution, $\prod$, of an ergodic Continuous-time Markov process, $Q$, is by first finding its embedded Markov chain (EMC). Strictly speaking, the EMC is a regular discrete-time Markov chain. Each element of the one-step transition probability matrix of the EMC, $S$ is denoted by $s_{ij}$, such that $s_{ij} = q_{ij} / \sum_{k \neq i} q_{ik}$, if $i$ is not equal to $j$ and is $0$ otherwise. From this, $S$ may be written as

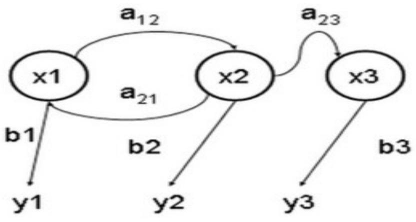
$S = 1 - D_Q^{-1} Q$ Where $D_Q = diag\{Q\}$ is the diagonal matrix of $Q$ To find the stationary probability distribution vector, we must next find $\varphi$ such that $\varphi(I - S) = 0$, with $\varphi$ being a row vector, such that all elements in $\varphi$ are greater than $0$ and $\|\varphi\|_1 = 1$ (the $1$ -norm, $\|x\|_1$ , is explained in Norm_(mathematics)), and the $0$ on the right side also being a row vector of $0's$ . From this, $\prod$ may be found as $\prod = -\varphi D_Q^{-1} / \|\varphi D_Q^{-1}\|$,

**3.6 Applications:** Markovian systems appear extensively in physics, particularly statistical mechanics, Markov chains can also be used to model various processes in queuing theory and statistics. Even without describing the full structure of the system perfectly, the signal models can make possible very effective data compression through entropy coding techniques such as arithmetic coding. They also allow effective state estimation and pattern recognition. The world's mobile telephone systems depend on the Viterbi algorithm for error-correction, while Hidden Markov models (where the Markov transition probabilities are initially unknown and must also be estimated from the data) are extensively used in speech recognition and also in bioinformatics, for instance for coding region/gene prediction. The Page Rank of a webpage as used by Google is defined by a Markov chain. It is the probability to be at page *i* in the stationary distribution on the following Markov chain on all (known) web pages. If $N$ is the number of known web pages, and a page $i$ has $k_i$ links then it has transition probability $(1-q)/k_i + q/N$ for all pages that are linked to and $q/N$ for all pages that are not linked to. The parameter $q$ is taken to be about 0.15.Markov chain methods have also become very important for generating sequences of random numbers to accurately reflect very complicated desired probability distributions - a process called Markov chain Monte Carlo or MCMC for short.. Markov chains also have many applications in biological modeling, particularly population processes, which are useful in modeling processes that are (at least) analogous to biological populations. A recent application of Markov chains is in geostatistics. That is, Markov chains are used in two to three dimensional stochastic simulations of discrete variables conditional on observed data. Such an application is called "Markov chain geostatistics", similar with kriging geostatistics. The Markov chain geostatistics method is still in development. Markov chains can be used to model many games of chance. The children's games

### 3.7 Hidden Markov Models (HMM)

An HMM consists of a signal modeled as a finite state Markov chain and an *observation* model that relates an observed process to the underlying Markov chain. Typically, the observation model consists of observing the state of the Markov chain perturbed by additive white noise. Such models have become increasingly popular over the last decade: application areas including speech processing, target tracking, digital communications, biomedical engineering, and finance. A major reason for this is the enormous flexibility and generality of the model and the fact that efficient state and parameter estimation algorithms exist and are well understood. In particular, the finite-state property means that finite dimensional state filters result even when the model is nonlinear. This makes the HMM formulation very attractive for approximating continuous state space nonlinear models for which finite-dimensional filters rarely exist.

3.7.1 Hidden Markov model



State transitions in a hidden markov model (example)x - hidden states, y – observable outputs.
a – transition probabilities, b – output probabilities

A **hidden Markov model** (**HMM**) is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. A HMM can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Hidden Markov models are especially known for their application in speech recognition and bioinformatics (e.g. HMMer).

***Probability of an observed sequence;***
The probability of observing a sequence $Y = y(0), y(1),...., y(L-1)$ of length $L$ is given by:

$$P(Y) = \sum_{x} P(Y/X)P(X)$$

Where the sum runs over all possible hidden node sequences $X = x(0), x(1),....., x(L-1)$. A brute force calculation of $P(Y)$ is intractable for realistic problems, as the number of possible hidden node sequences typically is extremely high. The calculation can however be speeded up enormously using a dynamic programming algorithm, called the forward algorithm.

***Using Hidden Markov Models:*** *There are 3 canonical problems associated with HMMs*

Given the parameters of the model, compute the probability of a particular output sequence. This problem is solved by the forward algorithm. Given the parameters of the model, find the most likely sequence of hidden states that could have generated a given output sequence. This problem is solved by the Viterbi algorithm. Given an output sequence or a set of such sequences, find the most likely set of state transition and output probabilities. In other words, train the parameters of the HMM given a dataset of sequences. This problem is solved by the Baum-Welch algorithm.

## Linear Prediction

**4.1 Linear prediction in time series:** One of the central problems in time series analysis is that of prediction i.e. given a series of sample values of a stationary discrete-time process, the future samples are to be predicted. Specifically, given $x(n-1), x(n-2),......., x(n-M)$, it is needed to predict the value of $x(n)$. The predicted value is expressed as a function of the given $M$ past samples. i.e.

$$\hat{x}(n|n-1,n-2,..n-M) = \psi(x(n-1), x(n-2),...x(n-M))$$

Now, if the function $\psi$ is a linear function of the variables $x(n-1), x(n-2),......., x(n-M)$, the prediction is linear. This is visualized in a $M$ - dimensional space spanned by $x(n-1), x(n-2),......., x(n-M)$.

$$\hat{x}(n|\, n-1, n-2,..., n-M) = \sum_{k=1}^{M} a_k x(n-k)$$

Where, $a_k$ are constant coefficients? The prediction error is defined as

$$f_M(n) = x(n) \text{-} \hat{x}(n|n\text{-}1,n\text{-}2,.......,n\text{-}M)$$

The subscript $M$ in $f_M(n)$ denotes the order of the prediction. i.e., the number of past samples that are used to predict the next sample. Hence, the problem of Linear Prediction (13) reduces to determining these coefficients subject to some condition. These coefficients are called linear prediction coefficients or predictor coefficients. The main challenge in linear prediction is estimation of predictor coefficients. Different algorithms and conditions on $a_k{}'s$ have been proposed and are used such as autocorrelation method, auto covariance method, Burg's method etc., (14)
A commonly used measure for this in probability theory is the RMS Error, i.e., Root Mean Square Error. RMS error is defined as $P_M = E(|f_M(n)|^2)$ The error can be minimized by finding the best, or optimal value of $a_k$. The error is minimized by differentiating $E$ w.r.t $a_k$ and setting the result equal to zero.

### 4.2 Autocorrelation method

Minimizing the prediction RMS error $(P_M)$, the Weiner-Hopf equations are obtained.

$$Ra = b$$

Where,

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_M \end{bmatrix}$$

$$R = \lfloor E\lfloor x_i x_j \rfloor\rfloor_{i,j=1,2,...,M}$$

$$b = [E[x_i x_{M+1}]]_{i=1,2,.., M}$$

Here, $R(k)$ denotes the autocorrelation function $E(x(n)x(n-k))$ of the sequence $x(n)$ for a lag $k$. $R_{xx}(-k) = R_{xx}(k)$, since the process is assumed to be stationary..

In order to solve for the coefficient $a_k$ ,

First, determine the autocorrelation function up to order $M$ for the input process $x(n)$.Then, solve the equation, $Ra = b$

$a = R^{-1}b$

### 4.3 Calculation of the Autocorrelation coefficients

The autocorrelation function of the input process may not be known apriori. Hence it is to estimate it based on the input process itself.

$$R_{xx}[i][j] = e[abs\ (i - j)]$$

$$e[k] = x(n) * x(n + k)$$

$$b[i] = e[M - i]$$

Segregating the wet wastes is done first and then metal and iron particles are separated with the use of magnets. There are also methods that utilize water jets for classifications. But some wastes are still segregated by workers manually. Even though there are safety precautions adopted, it is still highly risky and dangerous for the manual labour. If this process is completely automated, then the segregation process can continue without human intervention. There are some robotic processes for this purpose, but installing them is tedious and expensive [6]. But an AI based solution can reduce the machinery cost and size and also make the segregation process easier. The goal here is to process the image and categorize it into **4.4 Algorithm**

**Step 1:** Generate the random values $x_0, x_1, x_2,....., x_M$ where $M > 1$ is large. The random values are chosen by using the 'rand' function. The random values are chosen uniformly such that they fall in the interval $[0,1]$.

$x_{i+1} = \rho x_i + (1-\rho)r$ Where, $\rho \in [0,1]$ is a fixed constant. $r$ is randomly chosen from $[0,1]$ uniformly.

This estimation of the autocorrelation function assumes the apriori knowledge of the entire process.

Let $f_{x_a}$ be the true probability density of the random variable $X$ . $X_e$ is the estimated value. It is the function of previous samples $x_{-1}, x_{-2},....., x_{-M}$ . Assuming all the samples $x_{-1}, x_{-2},....., x_{-M}$ are all independent,

$$X_e = g(x_{-1}, x_{-2},...., \ x_{-M})$$

$$f_{x_e} = g(x_{-1}, x_{-2},...., x_{-M}).f_{x_a}(x_{-1}).f_{x_a}(x_{-2})...f_{x_a}(x_{-M})$$

$$f_{x_e} = g(x_{-1}, x_{-2},..., x_{-M})\prod_{i=1}^{M} f_{x_a}(x_{-i})$$

$$g(x_{-1}, x_{-2},...., x_{-M}) = \sum_{i=1}^{M} a_i x_i$$ for some $M > 1$

$a_{-i}{}'s$ can be chosen to adapt to the particular dataset. But the functional form of the estimator is seriously restrictive.

**Step 2**: After obtaining the random values, they are to be normalized. Given $x(0), x(1),...., x(M-1)$, normalizing of values is done by:

Let max = maximum of $x(i)$ for $i = 0,1,...., M-1$ and min = minimum of $x(i)$ for $i = 0,1,.....M-1$ The normalized values are obtained by, $y(i) = (x(i) - \min)/(\max - \min)$ Then $y(i)$ is in the interval $[0,1]$

**Step 3**: Now compute $x_i^e = \hat{x}(x_{i-1}, x_{i-2},...., x_{i-k})$ using the previous $k$ actual samples.

**Step 4**: Then compare $x_i$ and $x_i^e$ to get the probability of error by using the condition $|x_i - x_i^e| < \varepsilon$ , where $\varepsilon$ is a constant.

### 5.Markov chain method for prediction:

In linear prediction, the functional form is to be chosen and the parameters for the data set are to be estimated. But it is very critical to choose the best estimator for prediction. So, an alternative approach is to be used for prediction. In this paper work, an approach based on Markov chains is proposed.

### 5.1 Approach via Markov chains

In this approach, first the transformation is to be done i.e., discrediting the state space and digitizing the functional values. Discrimination concerns the process of transferring continuous models and equations into discrete counterparts. This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers. In order to be processed on a digital computer another process named quantization is essential. Discrete values are intervals in a continuous system of values. While the number of continuous values for an attribute can be infinitely many, the number of discrete values is often few or finite. There are many other advantages of using discrete values over continuous ones. Discrete features are closer to a knowledge-level representation than continuous ones. Data can also be reduced and simplified through discretization. For both users and experts, discrete features are easier to understand, use, and explain. The transformation is done by discretizing the state space $R^{n+1}$ to $Q^{n+1}$ where $Q$ is finite set.

Let

$$A=(a_1,b_1)\times(a_2,b_2)\times........\times(a_n,b_n) \rightarrow (q_1,q_2,q_3,.....,q_n,q_{n+1}) \in Q\text{, a finite set}$$

$$\forall\ (x_1,x_2,.....,x_n,x_{n+1}) \in A$$

$$\rightarrow q \in Q$$

Digitize the functional values

$$f(x_n|x_{n-1},x_{n-2},....,x_0) \rightarrow g(q_0|q_{n-1,}q_{n-2},.....,q_0)$$

$$P_{R^{n+1}}(A)=P_Q(q)$$

$$P_Q(q_{n+1}|q_1,q_2,\ldots q_n)=$$

$$P\{x_{n+1}\in(a_{n+1},b_{n+1})|(x_1,x_2,\ldots.x_n)\in(a_1,b_1)\times(a_2,b_2)\times\ldots\times(a_n,b_n)\}$$

### 5.2 Markov Chain Method:

The time series analysis is developed to model a set of observations developing in time i.e., the fundamental starting point for time series and for more general Markov models is virtually identical. A Markov model immediately assumes a short-term dependence structure on the variables at each time point, time series theory concentrates rather on the parametric form of dependence between the variables.

A Markov chain is a sequence of random variables $S=\{x_n : n\in T\}$ , where $T$ is a countable time-set. $T$ is written as $Z_+ := \{0,1,....\}$ . The critical aspect of a Markov model is that it is forgetful of all but its most immediate past i.e., the future of the process is independent of the past given only its present value. For a process $\Phi$, evolving on a state space $X$ and governed by an overall probability law $P$ , to be a time-homogenous Markov chain, there must be a set of transition probabilities $\{P^n(x,A), x\in X, A\subset X\}$, for appropriate states $A$ such that for times $n,m$ in $Z_+$

$$P\{\Phi_{n+m}\in A|\Phi_i, i\le m; \Phi_m = x\}=P^n(x,A)$$

that is, $P^n(x,A)$ denotes the probability that a chain at $x$ will be in the state $A$ after $n$ steps or transitions. The independence of $P^n$ on the value of $\Phi_i, i\le m$ , is the Monrovian property, and the independence of $P^n$ and $m$ is the time-homogeneity property. A Markov chain $\Phi=\{\Phi_{0,}\Phi_1,....\}$ is a particular type of stochastic process, at times $n \in Z_+$ , taking values $\Phi_n$ in a state space $X$ . A discrete time stochastic process $\Phi$ on a state space is, a collection $\Phi=\{\Phi_{0,}\Phi_1,....\}$ of random variables, with each $\Phi_i$ taking values in $X$ the defining characteristic of a Markov chain is that its future trajectories depend on its present and its past only through the current value. The random variables $\{\Phi_0,\Phi_1,....,\Phi_n\}$ , as a sequence take on values in the space $X^{n+1}=X_0\times X_1\times......\times X_n$ , the $(n+1)$ copies $X_i$ of the countable space $X$ , equipped with the product field $B(X^{n+1})$ which consists again of all subsets of $X^{n+1}$ .The conditional Probability $P^n_{x_0}$ $(\Phi_1=x_1,\ldots.\Phi_n=x_n):=P_{x_0}(\Phi_1=x_1,\ldots..\Phi_n=x_n)$ , defined for any sequence $\{x_0, ........, x_n\} \in X^{n+1}$ and $x_0 \in X$, and the initial probability distribution $\mu$ on $B(X)$ completely determine the distributions of $\{\Phi_0, ........, \Phi_n\}$.

**Countable space Markov chain**: The process $\Phi = \{\Phi_0, \Phi_1, .........\}$ taking values in the state space is a Markov chain if for every $n$, and any sequence of states $\{x_0, ........, x_n\}$,

$$P_\mu(\Phi_0 = x_o, \Phi_1 = x_1, ..... \Phi_n = x_n) = \mu(x_0) P_{x_0}(\Phi_1 = x_1) P_{x_1}(\Phi_1 = x_2) ..... P_{x_{n-1}}(\Phi_1 = x_n).$$

The probability $\mu$ is called the initial distribution of the chain. The process $\Phi$ is a time-homogenous Markov chain if the probabilities $P_{x_j}(\Phi_1 = x_{j+1})$ depend only on the values of $x_j, x_{j+1}$ and are independent of the time points $j$ .By extending this in the obvious way from events in $X^n$ to events in $X^\infty$ the initial distribution, followed by the probabilities of transitions from one step to the next are obtained to completely define the probabilistic motion of the chain.

If ø is a time-homogenous Markov chain,

$$P(x, y) := P_x(\Phi_1 = y);$$

Then the definition can be written as

$$P_\mu(\Phi_0 = x_o, \Phi_1 = x_1, ......., \Phi_n = x_n) = \mu(x_0) P(x_0, x_1) P(x_1, x_2) ...... P(x_{n-1}, x_n),$$

or equivalently, in terms of the conditional probabilities of the process $\Phi$,

$$P_\mu(\Phi_{n+1} = x_{n+1} | \Phi_n = x_n, ..........., \Phi_0 = x_0) = P(x_n, x_{n+1})$$

This equation incorporates both the 'loss of memory' of Markov chains and the 'time-homogeneity For a given model, probability $P_{x_0}$ for a fixed $x_0$ is defined by defining the one-step transition probabilities for the process, and building the overall distribution using Markov transition matrix.

Transition Probability Matrix: The matrix $P = \{P(x, y), x, y \in X\}$ is called a Monrovian transition matrix if

$$P(x, y) \geq 0, \sum_{Z \varepsilon X} P(x, y) \geq 0, \sum_{Z \in X} P(x, z) = 1, x, y \in X$$

The usual matrix iterates $P^n = \{P^n(x, y), x, y \in X\}$ by setting $P^0 = I$, the identity matrix and then taking inductively

$$P^n(x, z) = \sum_{y \in X} P(x, y) P^{n-1}(y, z)$$

. $P^n$ is called the $n$ -step transition matrix. For $A \subset X$,

$$P^n(x, A) := \sum_{y \in A} P^n(x, y)$$

To define a Markov chain from a transition function the laws governing a trajectory of fixed length $n \geq 1$ . The random variables $\{\Phi_0, \Phi_1, .............., \Phi_n\}$, thought of as sequence, take values in the space $X^{n+1} = X_0 \times ........ \times X_n$, equipped with $B(X^{n+1})$ which consists of all subsets of $X^{n+1}$

For a general time series,

$$P\{x_{n+1} | \mathrm{x}_n x_{n-1}\} \neq P\{x_{n+1} | \mathrm{x}_n\}$$

$$P\{x_{n+1} | \mathrm{x}_n\} = \sum_Z P\{x_{n+1} | \mathrm{x}_n, Z\} P(Z)$$

In general, $P\{x_{n+1} | \mathrm{x}_n, Z\} \neq P\{x_{n+1} | \mathrm{x}_n, Z'\}$ for $Z \neq Z'$ But for Markov chain of order one,

$$P\{x_{n+1} | \mathrm{x}_n, Z\} = P\{x_{n+1} | \mathrm{x}_n, Z'\}$$

$\forall$ $Z$ and $Z'$

$$P\{x_{i+N+1} | \mathrm{x}_{i+1} x_{i+2} .. x_{i+N}\} \approx P\{x_{i+N+1} | \mathrm{x}_1, x_2, ...., x_N\}$$

For sufficiently large $N$ $(\leq 10)$

$$P\{X_{n+2} = x | \mathrm{X}_{n+1} = y, X_n = z\} = P\{X_{n+2} = x, X_{n+1} = y, X_n = z\} / P\{X_{n+1} = y, X_n = z\}$$

$$P\{X_{n+2} = x | \mathrm{X}_{n+1} = y\} = P\{X_{n+2} = x, X_{n+1} = y\} / P\{X_{n+1} = y\}$$

$$\sum_Z P\{X_{n+2} = x, X_{n+1} = y | \mathrm{X}_n = z\} P\{X_n = z\} / \sum_Z P\{X_{n+1} = y | \mathrm{X}_n = z\} P\{X_n = z\}$$

**5.3 Alternative method to estimate conditional density:** When two random variables are statistically dependent, knowing the value of one of them lets experimenter get a better estimate of the value of the other one. Given the set of random variables $\{(x, y)\}$ in which $x$ is statistically related to the other random variable $y$ whose value can be observed. Now the objective is to estimate the conditional density of $x$ given $y$ . To estimate the conditional density $\hat{f}(x/y)$ , the two dimensional joint density $\hat{f}(x, y)$ for each pair of random variables formed in a cyclic fashion of estimated values i.e., $y_1, y_2, \ldots, y_N$ and the one dimensional density $\hat{f}_y(y)$ are to be known. Then the conditional density is estimated as the ratio of the two dimensional joint density to the one dimensional density of random variable multiplied by constant correction factor. Suppose $f_x$ and $f_y$ are the densities of the random variables $x$ and $y$ respectively and $f_{x,y}$ be the two dimensional joint density of $x, y$ . For some fixed $\varepsilon > 0$ , when $\hat{f}_y(y) \geq \varepsilon$ then,

$$\hat{f}_{x/y}(x/y) = \hat{f}_{x,y}(x, y)/H_{\varepsilon} * \hat{f}_y(\varepsilon)$$

$$\hat{f}_{x/y}(x/y) = 0, \text{ Otherwise.}$$

**6.Results**

The probability of error is calculated for different estimators by considering the past samples. The threshold value that is taken into account for calculating the probability of error is also critical. In this work, the threshold values that are considered for calculating the probability of error are T=0.05 and T=0.005. The probability of error is calculated considering the past 100 samples.

| Estimator | Pe for T=0.005 | Pe for T=0.05 |
|---|---|---|
| F(xn)=xn-1 | 0.50 | 0.30 |
| $F(x_{n+1}) = \sum_{i=1}^{n} x_i/n$ | 0.48 | 0.30 |
| $F(x_{n+1}) = \sum_{i=n-K+1}^{n} x_i/k$ | 0.16 | 0.06 |

In the third case where the moving average method is considered, K refers to the number of past samples that are considered to predict the future one. K is taken as 10 i.e., the history of only 10 samples is considered.

$$x_{n+1} = \sum_{i=1}^{n} a_i x_i$$

Where ai are the coefficients that are to be calculated using the autocorrelation method? It is very critical to choose the functional form (estimator) for linear prediction. Also it is not easy to choose the parameters that best fit the linear predictor, minimizing the error in the process. Increasing the number of parameters will not always lead to better results. Thus, Linear Prediction method has several restrictions. So, an alternative method based on Markov chains is proposed in which the conditional density is estimated.

**7.Conclusion:**

In signal processing applications, the estimation of the predictor is a common problem. There are two different types of estimating the predictor. One way is considering the total history of the predictor and the other way is considering the length of the predictor. It is always critical how the predictor is valid to about what region. In non-deterministic methods, the range of values that the predictor can have is very large. For these the nature of the predictor is obtained by using fuzzy systems. With linear prediction, the future values can be predicted using the past values. In order to get the best prediction results, the linear prediction coefficients are to be calculated in order to best fit for the predictor. But the time series models and autoregressive models for linear prediction need a functional form to be chosen in advance based on data set which is very critical. In addition the parameters are to be chosen in such a way to minimize the RMS error. To avoid such problems, an alternative approach based on Markov chains is proposed in which the conditional density is estimated.

**8.Future Work:** Future work in statistical forecasting based on conditional density estimation should focus on developing scalable, data-efficient models that perform reliably under distributional shifts. Integrating hybrid approaches that combine classical statistical methods with deep learning, such as normalizing flows and Bayesian neural networks, can enhance interpretability and uncertainty quantification. Advancements in online and adaptive conditional density estimation will enable real-time forecasting for non-stationary environments. Additionally, incorporating domain knowledge, causal constraints, and explainable AI techniques can improve trust and decision support. Extending conditional density models to handle high-dimensional, multimodal, and spatiotemporal data remains a critical research direction.

## REFERENCES

[1] P.M.T.Broersen and S.de Waele, Selection of order and type of time series models estimated from reduced statistics, , IEEE , Instrumentation and measurement, May 2002.

[2]Michalis K.Titsias and Aristidis C.Likas, Shared kernel models for class conditional density estimation, IEEE Transactions on neural networks, vol-12, September 2001.

[3]Jan S.Erkelens and Piet M.T.Broersen, Bias propagation in the autocorrelation method of linear prediction, IEEE Transactions on speech and audio processing, vol 5, March 1997.

[4]Robert V. Hogg, Allen T. Craig, Introduction to Mathematical Statistics, Pearson Edition, Fifth Edition, 2004.

[5]Athanasios Papoulis, S. Unnikrishna pillai, Probability, Random variables and Stochastic Processes, Tata McGraw-Hill Edition, Fourth Edition, 2004.

[6]S.C.Gupta and V.K.Kapoor, Fundamentals of Mathematical Statistics, Sultan chand and sons, Eleventh Edition, 2002.

[7] Richard A.Johnson, Miller & Freund's Probability and Statistics for Engineers, Prentice V Hall, Sixth Edition, 2003.

[8] Richard O.Duda, Peter E.Harl, David G.Stork, Pattern Classification, Wiley Interscience Publications, Second Edition, 2004.

[9] Hillier & Lieberman, Introduction to Operations Research, Tata McGraw Hill, Seventh Edition, 2005.

[10]Gao, Zhuang, and Trevor Hastie. "LinCDE: Conditional Density Estimation via Lindsey's Method." *Journal of Machine Learning Research*, vol. 23, 2022, pp. 1–55.

[11] Wen, Haoran, et al. "Continuous and Distribution-Free Probabilistic Wind Power Forecasting: A Conditional Normalizing Flow Approach." *arXiv*, 2022, arXiv:2206.02433.

[12] Arpogaus, Matthias, et al. "Short-Term Density Forecasting of Low-Voltage Load Using Bernstein-Polynomial Normalizing Flows." *arXiv*, 2022, arXiv:2204.13939.

[13]Dey, Bishal, et al. "Towards Instance-Wise Calibration: Local Amortized Diagnostics and Reshaping of Conditional Densities (LADaR)." *arXiv*, 2022, arXiv:2205.14568.

[14]Rittler, Nicholas, et al. "A Deep Learning Approach to Probabilistic Forecasting of Weather." *arXiv*, 2022, arXiv:2203.12529.

[15]Jamgochian, Alex, et al. "Conditional Approximate Normalizing Flows for Joint Multi-Step Probabilistic Forecasting with Application to Electricity Demand." *arXiv*, 2022, arXiv:2201.02753.

[16]De Gooijer, Jan G., et al. "Kernel-Based Hidden Markov Conditional Densities." *Computational Statistics & Data Analysis*, vol. 169, 2022, Article 107431.

[17]Wang, Yifan, et al. "Short-Term Probability Density Function Forecasting of Industrial Loads Based on ConvLSTM-MDN." *Frontiers in Energy Research*, vol. 10, 2022, Article 891680.

[18]Montes-Galdón, Carlos, Juan Paredes, and Elias Wolf. "Conditional Density Forecasting: A Tempered Importance Sampling Approach." *ECB Working Paper Series*, no. 2754, 2022.

[19]Gündüz, Nurdan, and Şeref Karakoç. "Probability Density Forecasting of Wind Speed Based on Quantile Regression and Kernel Density Estimation." *Chaos, Solitons & Fractals*, 2022, Article 112416.

[20]Pinson, Pierre, et al. "Probabilistic Forecasting: Theory and Practice." *Annual Review of Statistics and Its Application*, vol. 9, 2022, pp. 1–25.

[21]Izbicki, Rafael, and Ann B. Lee. "Conformal Prediction for Conditional Density Estimation." *Journal of Computational and Graphical Statistics*, vol. 31, no. 2, 2022, pp. 349–360.

[22]Gasthaus, Jan, et al. "Probabilistic Forecasting with Spline Quantile Function RNNs." *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022, pp. 1–15.

[23]Salinas, David, et al. "Deep Probabilistic Forecasting with Autoregressive Normalizing Flows." *International Journal of Forecasting*, vol. 38, no. 4, 2022, pp. 1480–1495.

[24] Rangapuram, Syama Sundar, et al. "End-to-End Learning of Conditional Probability Distributions for Time-Series Forecasting." *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 1–14.

[25]Laio, Francesco, and Stefano Tamea. "Probabilistic Forecasting in Hydrology: A Review." *Water Resources Research*, vol. 58, no. 6, 2022, Article e2021WR031956.