Conditional Density based Statistical Prediction

[1] **Dr J Rama Dévi [2]Dr. K. Koteswara Rao [3]Dr M Venkateswara Rao**

, [1]Sr Asst Professor [2]Associate Professor,
[12] Department of CSE, Prasad V Potluri Siddhartha Institute of Technology, VIjayawada,
[3]Professor, GIT, GITAM University, Visakhapatnam, India

Abstract
Numerous genuine issues, for example, financial exchange expectation, climate determining and so forth has inalienable arbitrariness related with them. Receiving a probabilistic system for forecast can oblige this dubious connection among past and future. Commonly the interest is in the contingent likelihood thickness of the arbitrary variable included. One methodology for expectation is with time arrangement and auto relapse models. In this work, liner expectation technique and approach for computation of forecast coefficient are given and likelihood of blunder for various assessors is determined. The current methods all need in some regard assessing a boundary of some accepted arrangement. In this way, an elective methodology is proposed. The elective methodology is to gauge the restrictive thickness of the irregular variable included. The methodology proposed in this theory includes assessing the (discretized) restrictive thickness utilizing a Markovian definition when two arbitrary factors are genuinely needy, knowing the estimation of one of them allows us to improve gauge of the estimation of the other one. The restrictive thickness is assessed as the proportion of the two dimensional joint thickness to the one-dimensional thickness of irregular variable at whatever point the later is positive. Markov models are utilized in the issues of settling on an arrangement of choices and issue that have an innate transience that comprises of an interaction that unfurls on schedule on schedule. In the nonstop time Markov chain models the time stretches between two successive changes may likewise be a ceaseless irregular variable. The Markovian methodology is especially basic and quick for practically all classes of classes of issues requiring the assessment of contingent densities.

Keywords:
Statistical prediction, unbiased ness, Sufficiency, smoothing, univariate time series, Autoregressive, Markov chains,

1. Introduction

1.1 Statistics:

Insights are estimations, identifications or assessments of common wonder, typically methodicallly masterminded, investigated and introduced as to display significant between connections among them. Present day insights (6) alludes to an assemblage of techniques and rules that have been created to deal with the assortment, depiction, outline and examination of mathematical information. In factual hypothesis, a "measurement" is a polite capacity of the information. A measurement is adequate (4) on the off chance that it is similarly just about as instructive as the full information. In numerous applications it isn't uncommon to have handfuls or many boundaries and a huge number of preparing tests. An adequate measurement is a capacity ' of the examples ' that contains all the data applicable to assessing some boundary ' '. An essential hypothesis concerning adequate insights is the Factorization hypothesis (8) which expresses that's' is adequate for ' if and just if can be figured into the result of two capacities: one relying just upon ' 'and ' ', the other relying just upon preparing samples.In applying measurements to a logical, mechanical, or cultural issue, one starts with a cycle to be contemplated.

This may be a populace of individuals in a country, of precious stone grains in a stone, or of products made by a specific manufacturing plant during a given period. It might rather be a cycle seen at different occasions what is known as a period arrangement. For functional reasons, as opposed to ordering information about a whole interaction, one for the most part rather examines a picked subset of the cycle, called an example (5). Information are gathered about the example in an observational or trial setting. The information are then exposed to measurable examination, which fills two related needs (6): portrayal and surmising. Elucidating measurements can be utilized to sum up the information, either mathematically or graphically, to depict the example. Essential instances of mathematical descriptors incorporate the mean and standard deviation. Graphical outlines incorporate different sorts of diagrams and charts. An inferential measurement is utilized to show designs in the information, representing arbitrariness and drawing inductions. These deductions may appear as answers to yes/no inquiries (speculation testing), evaluations of mathematical qualities (assessment), forecast of future perceptions, portrayals of affiliation (connection), or demonstrating of connections (relapse). Other demonstrating procedures incorporate ANOVA, time arrangement, and information mining. A significant issue

Manuscript received June 5, 2023
Manuscript revised June 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.6.16

lies in deciding the degree to which the picked test is agent. A measurement offers techniques to appraise and address for irregularity in the example and in the information assortment system, just as strategies for planning strong tests in any case.

1.2 Statistical prediction:

A prediction or conjecture is an assertion or guarantee that a specific occasion will happen later on. Generally, it relies upon one of two requirements if tests for quantitative patterns are applied. In the first place, the autonomous variable is quantitative, and second, the free factor is quantitative and a specific quantitative pattern theory is to be tried. In the primary case, the experimenter doesn't continue from specific assumptions; the experimenter simply searches for the best useful portrayal of the information. In the subsequent case, notwithstanding, the information are inspected concerning their similarity with expectations got from a specific hypothesis. An invalid speculation (H0) is any measurable theory (6) which involves one of the signs '=', '≤', or '≥' and which is testable by a given factual test. Its inverse is an elective speculation (H1), which normally is reciprocal to the H0 and against which the test is performed. On the off chance that the measurable forecast isn't identical to a solitary testable H0 or H1, there are fundamentally two alternatives: either to play out a less appropriate test and decipher the 'evident' observational relations among the example measurements, or to apply more than one test. The issue of boundary assessment is an old style one in measurements and it very well may be drawn closer severally. The basic methodologies are most extreme probability assessment (8) and Bayesian assessment.

1.3 Parameter estimation:

Thinking about an irregular example of size ' with likelihood work where are the obscure boundaries. At that point, there will consistently be an endless number of elements of test esteems called measurements, which might be proposed as evaluations of at least one of the boundaries. Clearly, the best gauge would be the one that falls closest to the genuine estimation of the boundary to be assessed i.e.; the measurement whose conveyance thinks as intently as conceivable close to the genuine estimation of boundary is respected the best gauge. The essential issue is to decide the elements of test perceptions. The assessing capacities are called assessors (7). A decent assessor needs to fulfill a few attributes:

Unprejudiced ness: An assessor is supposed to be fair-minded assessor of) if , for all , boundary space

Consistency: assessor, in light of irregular example of size, is supposed to be steady assessor of , the boundary space, if combines to in likelihood, for example On the off chance that as

Effectiveness: If in a class of predictable assessors for a boundary, there exists one whose examining fluctuation is not exactly that of any such assessor, it is known as the most proficient assessor. At whatever point such an assessor exists.

1.4 Non-parametric estimation:

In the parametric tests, the practical structures from which the examples are attracted is thought to be known and are worried about testing measurable theory about the boundaries of the capacity or assessing its boundaries. Then again, a non-parametric assessment (8) doesn't rely upon the specific useful structure from which the examples are drawn i.e., no presumptions are made with respect to the utilitarian structure.

In numerical account, deterministic numerical models of securities exchange conduct are inconsistent in foreseeing future conduct, due to different obscure variables that can influence the market patterns. As another option, a measurable forecast issue can be defined for the appropriate and grouped products in the stock, and the necessary boundaries associated with the elements of the expectation model or any non-parametric articles in the expectation model can be assessed from the information gathered throughout significant stretches of time. At all case, the patterns can be anticipated with sensible certainty. Quantum physical science is a surprising field of science since it empowers researchers to make forecasts based on likelihood In microchips, branch expectation grants to stay away from pipeline discharging at microcode expanding. Designing is a field that includes foreseeing disappointment and keeping away from it through part or framework excess. A few fields of science are infamous for the trouble of exact expectation and gauging, like programming dependability, catastrophic events, pandemics, demography, populace elements and meteorology.

1.5 Organization of the Paper:

The remainder of the paper is coordinated as: segment 2 gives insights concerning a portion of the accessible time arrangement models and auto relapse models for factual expectation. Section 3 gives brief presentation about Markov chains, their properties, Markov chains in discrete state space and ceaseless time Markov chains. Secret Markov models are additionally momentarily examined. Area 4 arrangements with straight forecast:

Predictions of future example, computation of expectation coefficients in such an approach to limit expectation mistake are given. Segment 5 for the most part centers on the Markov chain technique for the assessment of contingent thickness. Segment 6 shows the outcomes: Actual examples and anticipated examples are analyzed in direct expectation. Segment 7 gives the finish of the paper and future work that should be possible.

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. 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

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

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.

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 (x_t) 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,

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

Where 'α' $(0<\alpha<1)$ 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

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

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:

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

Trend smoothing:

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

Seasonal smoothing:

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

Forecast:

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

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 α,β,γ 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

$$b=1/L(((Y_{L+1}-Y_1)/L)+((Y_{L+2}-Y_2)/L)+.....+((Y_{L+L}-Y_L)/L))$$

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 $X_t=c+\sum_{i=1}^{p}\theta_iX_{t-i}+\varepsilon_t$ Where $\theta_1,\theta_2,....,\theta_p$ are the parameters of the model, 'c' a constant and ε_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,

$$X_t=\varepsilon_t+\sum_{i=1}^{p}\theta_iX_{t-i}+\sum_{i=1}^{q}\theta_iX_{t-i}$$

2.5.4 Calculation of the AR parameters:

The AR(p) model is given by the equation $X_t=\sum_{i=1}^{p}\theta_iX_{t-i}+\varepsilon_t$ It is based on parameters θ_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. γ_m is the autocorrelation function of X , σ_ε is the standard deviation of the input noise process, and δ_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

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

Representing it as a matrix for $m > 0$, thus getting equation solving all θ.For $m = 0$, $\gamma_0 = \sum_{k=1}^{p} \theta_k \gamma_{-k} + \sigma_\varepsilon^2$ allows us to solve σ_ε^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

$$X_t = \delta + \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots\cdots$$

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

$$\delta = \left(1 - \sum_{i=1}^{p} \theta_i\right)\mu \text{ with } \mu \text{ denoting the process mean.}$$

Another common approach for modeling univariate time series models is the moving average (MA) model:

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

Where X_t is the time series, μ 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.Nth 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 $\quad P(x_i)$

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

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

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

More generally, the Markov assumption for a nth order model is that X_i depends only on $X_{i-1}X_{i-2}X_{i-3}\ldots.X_{i-n}$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 n. 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:

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

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

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

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

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

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"):

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

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,

$\lim_{n \to \infty} p_{ij}^{(n)} = 1/M_j$ 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 $p_{ij} = \Pr(X_{n+1} = j | X_n = i)$ 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

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

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 δ -skeleton—the (discrete-time) Markov chain formed by observing $X(t)$ at intervals of δ units of time. The random variables $X(0), X(\delta), X(2\delta), \ldots\ldots$ give the sequence of states visited by the δ -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 φ such that $\varphi(I - S) = 0,$ with φ being a row vector, such that all elements in φ 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 $\prod = -\varphi D_Q^{-1} / \|\varphi D_Q^{-1}\|,$

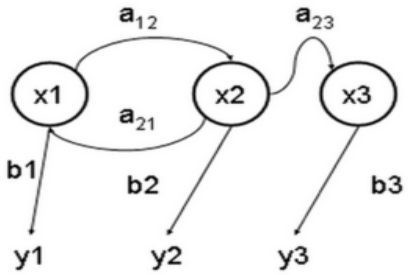
3.6 Applications:

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.

4. 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 ψ 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

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

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 $P_M = E\left(|f_M(n)|^2\right)$ 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,

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

$$R = \left[E\left[x_i x_j\right]\right]_{i,j=1,2,...,M}$$

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

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,

$$Ra = b \qquad a = R^{-1}b$$

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.

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

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

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

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,

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

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

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

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

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

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.

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

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

Digitize the functional values

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

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

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

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

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 Φ, 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 Φ_n in a state space X. A discrete time stochastic process Φ on a state space is, a collection $\Phi=\{\Phi_{0,}\Phi_1,....\}$ of random variables, with each Φ_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 $P^n_{x_0}(\Phi_1=x_1,....\Phi_n=x_n):=P_{x_0}(\Phi_1=x_1,.....\Phi_n=x_n)$, defined for any sequence $\{x_0,.....,x_n\}\in X^{n+1}$ and $x_0\in X$, and the initial probability distribution μ 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 $\{x_0,.....,x_n\}$,

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

The probability μ is called the initial distribution of the chain. The process Φ 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^∞ 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 Φ,

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

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,

$$P\{x_{n+1} | x_n x_{n-1}\} \neq P\{x_{n+1} | x_n\}$$

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

In general,

$$P\{x_{n+1} | x_n, Z\} \neq P\{x_{n+1} | x_n, Z'\}$$

for $Z \neq Z'$ But for Markov chain of order one,

$$P\{x_{n+1} | x_n, Z\} = P\{x_{n+1} | x_n, Z'\}$$

$\forall$ Z and Z'

$$P\{x_{i+N+1} | x_{i+1} x_{i+2} ... x_{i+N}\} \approx P\{x_{i+N+1} | x_1, x_2, ..., x_N\}$$

For sufficiently large N (≤ 10)

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

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

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

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., $y_1, y_2, ..., y_N$ 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,

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

$\hat{f}_{x/y}(x / y) = 0$, 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
$F(x_n)=x_{n-1}$	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.

6. 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.

7. Bibliography:

[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.