HIDDEN MARKOV MODELS FOR ESTIMATING IMPORT DYNAMICS WITH PROBABILITY DISTRIBUTION ANALYSIS

Vyshnavi. M, Muthukumar. M

•

[1]Research Scholar, Department of Statistics, PSG College of Arts & Science, Coimbatore-641014
[2]Assistant Professor, Department of Statistics, PSG College of Arts & Science, Coimbatore-641014
E-mail: [1]vyshnavimp@gmail.com, [2]muthukumar@psgcas.ac.in

Abstract

This article examines the use of Hidden Markov Models and probability distributions to study agricultural import dynamics, with a focus on revealing hidden patterns in the data. The idea behind the study is to enhance our understanding of how transitions between different agricultural import states occur and to explore the most suitable probability distributions for modeling these hidden states. The purpose of the study is to identify the optimal distributional fit for hidden states by evaluating the Transition Probability Matrix, Emission Probability Matrix, and Initial Probability Vector (π) for each state. The research implements the Akaike Information Criterion and the Bayesian Information Criterion to select the best-fitting distribution for each scenario. Furthermore, the paper focuses on the practical implications of these discoveries, such as determining the most likely state sequence using the Viterbi path, which can help influence decision-making and forecasting. The analysis is carried out using R software, which provides information about the associations between probability distributions, stationarity tests, and the role of model selection criteria such as AIC and BIC. Graphical representations of AIC and BIC values over several probability distributions, and additionally a correlation matrix between selected distributions, help to highlight the findings. Overall, the paper enhances our understanding of probability distributions through HMM frameworks for agricultural import dynamics, providing recommendations for optimal model selection in various applications.

Keywords: Probability Distributions, Hidden Markov Model, Lomax Distribution, Viterbi Path, Correlation Matrix.

I. Introduction

Throughout the world, agriculture significantly impacts how countries develop in terms of their economies, societies, and environments. Indicators of a nation's reliance on foreign agricultural products and its ability to produce agricultural goods domestically include the percentage of agricultural imports to total national imports**.** Over the last few years, the agricultural industry has been using more sophisticated statistical techniques, such as HMM, to understand the complex dynamics present in agricultural systems and numerous variables that affect these systems, including market swings, weather conditions, and crop yields, cause them to analyze using conventional methods.

Hidden Markov Models and probability distributions are critical tools for studying and modelling agricultural import dynamics, providing advanced approaches for identifying

underlying trends and behaviours in import data. HMMs provide a strong foundation for capturing these dynamics by presenting import data as a series of hidden states and visible emissions. Hidden states in agricultural import contexts may represent specific import patterns or market conditions, whereas emissions relate to observed variables such as import amounts and prices. HMMs help to explain import dynamics, identify dominant patterns, and anticipate future import trends by estimating transition and emission probabilities. Finally, the investigation of probability distributions within the framework of Hidden Markov Models provides a reliable method for assessing agricultural import dynamics. This study improves our understanding of model selection and sheds light on optimal distribution choices by evaluating the suitability of various distributions using rigorous criteria such as AIC and BIC, thereby contributing to improved forecasting accuracy and decision-making in the agricultural sector. Forecast accuracy is greatly increased when agricultural import data is analyzed using Hidden Markov Models. By detecting underlying hidden states, HMMs can efficiently capture the dynamic elements that affect these imports, including market swings and seasonal variations. HMMs offer vital insights into the uncertainties that define agricultural markets by predicting the probability distributions connected to these states. This method provides a foundation for comparison with conventional forecasting techniques and enables in-depth analyses of changing import trends. This research paper was carried out using R software, which has robust statistical libraries and HMM programs. While R has several advantages, we ran into some computational issues during the investigation. Large datasets, in particular, increased calculation time, and memory management became an issue as the Baum-Welch algorithm was iterated several times. To address these issues, we used efficient coding techniques and R tools like depmixS4 and HMM to streamline calculations. We recommend that future users use parallel computing or data optimization approaches to improve the performance of their analysis, particularly with larger datasets or more sophisticated models.

Sebastian George and Ambily Jose [13] presented a Generalised Poisson Distribution (GPD) as a statistical tool for describing serial dependency in time series count data. The GP-HMM, paired with HMM, demonstrates good convergence properties in both simulated and actual data, indicating its superiority over the P-HMM. Joshni George and Seemon Thomas [5] presented a negative binomial hidden Markov model for Kerala AES situations, estimating parameters with the EM technique, getting hidden state sequences, and transition probabilities. Hao Zhang, Weidong Zhang, Ahmet Palazoglu, and Wei Sun [3] proposed HMM-Gamma model uses a Gamma distribution, pre-labeled monitoring days, and improved Expectation-Maximization approaches to forecast ozone exceedances in the Livermore Valley and Houston Metropolitan Area. Hui Zhang, Qing Ming Jonathan Wu, and Thanh Minh Nguyen [4] proposed a novel Student's t hidden Markov model (SHMM) that takes into account Markov states, latent components, and observations. Existing literature emphasizes the need to use proper probability distributions to improve model accuracy and forecasting capabilities. Several studies have shown that various HMM formulations and parameter estimation strategies are successful in capturing complicated patterns in dynamic datasets. The unique feature of this study is that it compares five different probability distributions in detail using Hidden Markov Models. The study aims to determine the best method for modeling sequential data by examining the effectiveness and suitability of each distribution. This paper also presents the creation of a Lomax HMM, which improves the modeling capabilities of conventional HMMs by incorporating the special characteristics of the Lomax distribution.

II. Methods

A Hidden Markov Model is a probabilistic model that is commonly used to represent time series data or observation sequences. It is made up of two major components: hidden states and observable states. Hidden States are unobservable or latent variables that represent the underlying states of the

system under consideration. Hidden states can change over time using a Markov process, which means that the likelihood of transitioning from one state to another is determined only by the present state and not by the previous states. Observable states are the data points or observations that can be directly observed and linked to each hidden state. The observations are derived from probability distributions that are conditional on the present concealed state. The type of probability distribution used depends on the nature of the data being modelled. The transition between concealed states and the creation of observations are probabilistic processes that are controlled by transition probabilities and emission probabilities, respectively. The parameters of these probability distributions are usually calculated from observed data using techniques like maximum likelihood estimation (MLE) or the Expectation-Maximization (EM) algorithm. The following are the distributions used for representing the observable outcomes.

An HMM is designed to model a system as it changes over discrete time steps. At each time step, the system is in one of a finite number of hidden states, and it emits an observable symbol based on the probability distribution associated with the state. Transition probabilities stimulate hidden state transitions, while emission probabilities govern observable symbol emission. An HMM is defined formally by:

- A set of hidden states S = $\{s_1, s_2, \dots, s_v\}$
- A set of observable symbols O = $\{o_1, o_2, \dots, o_u\}$
- A transition probability matrix A = $[a_{kl}]$, where a_{kl} represents the probability of transitioning from state a_k to a_l.
- An emission probability matrix B = $[b_{mn}]$, where b_{mn} represents the probability of emitting a symbol s_n when in state s_m.
- An initial probability distribution $\pi = [\pi_k]$ represents the probability of starting in the states$_k$.

In a HMM, the system under consideration is considered a Markov process with unobservable (hidden) states. The set S includes all conceivable hidden states the system can be in at any one time. Each state s_i (where i ranges from 1 to v) has attributes that may influence the system's behavior. The hidden states are not immediately observable, but they impact the observable outputs (symbols), allowing the states to be inferred indirectly from the observation sequence. All of the observable outputs or symbols that the system is capable of producing are contained in the set O. Although we are unable to directly witness the concealed states, we can observe the symbols that they produce. We deduce the hidden states from the succession of these observable symbols. One important part of the HMM is matrix A, which shows the chances of changing from one hidden state to another. In particular, the chance of changing from state s_k to state s_l is represented by a_{kl}. The transition probability matrix is of size v × v, where v is the number of hidden states. To guarantee the total probability of changing from a particular state to any other state, all of the entries in each row of this matrix must add up to 1.

The Markov property, which postulates that the probability of changing to the next state depends only on the present state and not on any previous states, is contained in this matrix. The emission probability matrix B represents the correspondence between the observable symbols and the hidden states. The probability of observing the symbol o_n is represented by the entry b_{mn}, assuming that the system is in the hidden state s_m. The size of this matrix is v×u, where u is the number of observable symbols and v is the number of concealed states. The entries in every row of the emission probability matrix must add up to 1, just like in the transition probability matrix. The model can provide a series of observations based on the underlying hidden state sequence because the emission matrix essentially transfers the hidden states to the observable symbols. The initial probability distribution π defines the probability of the system starting in each hidden state. The

vector π = [$\pi_1, \pi_2, \ldots, \pi_v$] is of length v, where π_k indicates the chance that the system starts in the state s_k.This distribution is crucial in determining the likelihood of different sequences of concealed states. It serves as the starting point for generating sequences in the HMM, and like other probability distributions, the entries must add to one.

The model parameters in the HMM must be constantly updated and improved in real-world applications, especially in scenarios involving temporal data, such agricultural imports over time. This is necessary to describe specific time points adequately. This is especially crucial when using the model to forecast or examine long-term patterns in agricultural data. Estimating the parameters governing the state transitions and the emission of visible symbols is a major difficulty in this context.

One popular method for estimating an HMM's parameters is the Baum-Welch algorithm, a variation of the Expectation-Maximization (EM) algorithm. The program looks for the values of A, B, and π in the model that maximize the probability of the observed sequence of symbols. Improved estimates are produced by iteratively adjusting the model parameters in light of the observed data. A continuous process [8] of developing model parameters in the transition state to explain the specific time point in importing agricultural data. A HMM is typically denoted by, $\mu = (A, B, \pi)$. This model provides us with the state transition probability, the observational probability, and the probability of starting in a specific state. The Baum-Welch algorithm, commonly known as the EM algorithm, focuses on parameter estimation via direct numerical maximum likelihood estimation. To maximize and determine the posterior analysis of the hidden variables.

I. Transition Probability Estimation

The likelihood of a transition from one state to another, such as from state k to state l, is represented by the symbol a_{kl}. The following formula can be used to estimate it [14], [15],

$$a_{kl} = \frac{Expected\ number\ of\ transitions\ from\ state\ k\ to\ l}{Expected\ number\ of\ transitions\ from\ state\ k} \tag{1}$$

The ratio indicates how frequently the system moves from state k to state l throughout the observation time. It can be mathematically represented as follows:

$$a_{kl} = \frac{\sum_{t=1}^{T} p_t(k,l)}{\sum_{t=1}^{T} \gamma_k(t)} \tag{2}$$

where, $p_t(k, l)$ represents the likelihood of being in state k at time t and moving to state l at time t+1 and $\gamma_k(t)$ represents the chance of being in state k at time t. An alternative expression for a_{kl} that integrates the forward-backward variables α and β, signifying the forward and backward probabilities is:

$$a_{kl} = \frac{\sum_{t=1}^{T} \alpha_k(t) a_{kl} b_k(O_{t+1}) \beta_l(t+1)}{\sum_{t=1}^{T} \alpha_k(t) \beta_k(t)} \tag{3}$$

The combined likelihood [16] of being in state k at time t and changing to state l at time t+1 is as follows:

$$Pr(k, l) = Pr(S_t = k, S_{t+1} = l \ / \ O, \mu)$$

$$= \frac{Pr(S_t = k, S_{t+1} = l / O, \mu}{P(O/\mu)}$$

$$= \frac{\alpha_k(t)a_{kl}b_l(O_{t+1})\beta_l(t+1)}{\sum_{m=1}^{v}\sum_{n=1}^{v}\alpha_m(t)a_{mn}b_n(O_{t+1})\beta_n(t+1)} \quad (4)$$

Given the observation sequence O and the model μ, the probability of being in state k at time t is represented by the symbol $\gamma_k(t)$ is defined as:

$$\gamma_k(t) = \Pr(S_t = k/O,\mu) = \sum_{l=1}^{v} \Pr(S_{t,} = k, S_{t+1} = l/\ O, \mu)$$

$$= \sum_{l=1}^{v} Pr(k,l) \quad (5)$$

II. Initial State Probability

The initial state probability represents the likelihood of beginning in state k at time t = 1 π_k. It can be stated as:

$$\pi_k = \gamma_k(t) \quad (6)$$

III. Emission Probability Estimation

When moving from state k to state l, the probability of determining symbols n is represented by the emission probability, b_{kln}.

$$b_{kln} = \frac{Expected\ number\ of\ transitions\ from\ k\ to\ l\ with\ symbol\ n\ observed}{Expected\ number\ of\ transitions\ from\ k\ to\ l} \quad (7)$$

This can be expressed as:

$$b_{kln} = \frac{\sum_{t:O_t=n,1\le t\le v} Pr(k,l)}{\sum_{t=1}^{v} Pr(k,l)} \quad (8)$$

Given a collection of observations, the Viterbi algorithm is a dynamic programming technique determining the most likely order of hidden states in an HMM. The process starts with an initialization step in which the first observation's emission probabilities and initial distribution are used to compute the likelihood of beginning in each stage. The algorithm iterates through each time step in the recursion step, utilizing back pointers for retaining the best path and the transition and emission probabilities to determine the most likely transitions between states. After analysing all observations, the termination step finds the state with the highest probability at the last time step. Ultimately, the algorithm uses the backpointers to trace the series of hidden states in the backtracking step, generating the most likely path that explains the observed data. MATLAB software is used in this investigation to determine the Viterbi path.

IV. HMM Models with Different Distributions

To determine whether different probability distributions are appropriate for representing the dynamics of agricultural imports, they are incorporated into the HMM framework. These distributions offer a variety of approaches for simulating the data's underlying patterns and state transitions. These are as follows:

Normal - HMM

"Normal-HMM" is a Hidden Markov Model (HMM) in which the emission probabilities are represented using a normal (Gaussian) distribution. In the case of HMM, each hidden state generates observations with a normal distribution. This means the chance of witnessing a specific value given a hidden state is determined using normal distribution characteristics, specifically the mean and

variance. To examine agricultural import dynamics in India, we built a Hidden Markov Model with a normal distribution with μ = 4.575 and σ = 1.53. We used the normal distribution's probability density function to model the emission probabilities within the HMM framework. The pdf of the Normal-HMM is expressed as follows:

$$P(x_t/y_t = k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-(\frac{(x_t-\mu_k)^2}{2\sigma_k{}^2})} \quad (9)$$

N (μ, σ^2) as the Normal Distribution with mean μ and variance σ^2, x_t as the observed data at time t. y_t as the hidden state at time t. By incorporating this PDF into the HMM framework, we can successfully simulate and evaluate patterns and transitions in agricultural import dynamics in India, allowing for more informed decision-making and policy development in the agriculture sector. Given a sequence of observed import values $x_1, x_2, \dots, x_N$ and a sequence of hidden states $y_1, y_2, \dots, y_N$, the log-likelihood can be calculated as:

$$\text{Log-likelihood} = \sum_{t=1}^{N} log(P(x_n/y_n; \mu_{y_n}, \sigma_{y_n})) + \sum_{t=2}^{N} log(P(y_n/y_{n-1}; A)) \quad (10)$$

where, $P(x_n/y_n; \mu_{y_n}, \sigma_{y_n})$ is the probability of observing x_n given the hidden state y_n. $P(y_n/y_{n-1}; A)$ is the transition probability from state y_{n-1} to state y_n.

Gamma – HMM

In a Gamma-HMM, each hidden state generates observations with a Gamma distribution. The Gamma distribution, a continuous probability distribution, is commonly used to model positive-valued random variables. It's defined by two parameters: shape (α) and scale (β). The shape parameter indicates the shape of the distribution, whereas the scale parameter governs its spread. When we integrate a Gamma distribution within a Hidden Markov Model (HMM) to interpret agricultural import dynamics in India, with parameters α = 8.6192 and β = 1.8836, we use the Gamma distribution's probability density function to characterize emission probabilities within the HMM framework. The pdf of the Gamma-HMM is expressed as follows:

$$P(x_t/y_t = k) = \frac{\beta_k{}^{\alpha_k}}{\Gamma(\alpha_k)} x_t{}^{\alpha_k - 1} e^{-\beta_k x_k} \quad (11)$$

Γ(.) for the Gamma function. During the training phase of the Gamma HMM, the parameters α and β are commonly calculated using approaches such as maximum likelihood estimation (MLE) or EM algorithm, which are similar to other HMM versions. These algorithms seek to optimize the model parameters to maximize the likelihood of the observed data within the model. Overall, the Gamma HMM is a versatile framework for modelling sequential data with positive values, with emissions expected to follow a Gamma distribution. The forward algorithm, which is typically used to calculate the likelihood in HMMs, must be changed accordingly. Given a sequence of observations $X = \{x_1, x_2, \dots, x_N\}$ and a Gamma HMM with parameters ʎ = (A, B, π, α, β). The steps listed below are used to determine the log-likelihood [3]:

- Initialization

$$\alpha_1(k) = \pi_k P(x_1/y_{1=k}) \quad (12)$$

$$\alpha_1(k) = \pi_k \times \frac{\beta_k{}^{\alpha_k}}{\Gamma(\alpha_k)} x_1{}^{\alpha_k - 1} e^{-\beta_k x_1} \quad (13)$$

- Recursion

$$\alpha_{t+1}(l) = (\sum_{k=1}^{N} \alpha_t(k)\ A_{kl})\ P(\ x_{t+1}/y_{t+1=l}) \quad (14)$$

$$\alpha_{t+1}(l) = (\ \sum_{k=1}^{N} \alpha_t(k)\ A_{kl})\ \frac{\beta_l{}^{\alpha_l}}{\Gamma(\alpha_l)} x_{t+1}{}^{\alpha_l - 1} e^{-\beta_l x_{k+1}}$$

- Termination

$$log\ P\ (x/\Lambda) = log(\sum_{k=1}^{N} \alpha_T(k)) \quad (15)$$

Weibull – HMM

A Weibull Hidden Markov Model (Weibull HMM) is a variation of the classic Hidden Markov Model in which the emission distribution for each hidden state follows a Weibull distribution. Using a Weibull distribution with parameters α = 3.257 and β= 5.113. The emission probability for state k emitting observation x_t is given by the probability density function of the Weibull distribution:

$$P\ (x_t\ /y_t = k) = \frac{\alpha_k}{\beta_k} (\frac{x_t}{\beta_k})^{\alpha_k - 1} e^{-(\frac{x_t}{\beta_k})^{\alpha_k}} \quad (16)$$

where x_t is the observed value at time step t, α_k is the shape parameter and β_k is the scale parameter. We must modify the computation to take into account the emission probabilities modelled by Weibull distributions to determine the log-likelihood of a series of observations given Weibull HMM. We will apply the forward algorithm, which is comparable to the normal HMM log-likelihood computation, but using the probability density function for the Weibull distribution.

- Initialization: $\alpha_1(k) = \pi_k \frac{\alpha_k}{\beta_k} (\frac{x_t}{\beta_k})^{\alpha_k - 1} e^{-(\frac{x_t}{\beta_k})^{\alpha_k}}$ (17)
- Recursion: $\alpha_{t+1}(l) = (\ \sum_{k=1}^{N} \alpha_t(k)\ A_{kl})\ \frac{\alpha_k}{\beta_k} (\frac{x_t}{\beta_k})^{\alpha_k - 1} e^{-(\frac{x_t}{\beta_k})^{\alpha_k}}$ (18)
- Termination: $log\ P\ (x/\Lambda) = log(\sum_{k=1}^{N} \alpha_T(k))$ (19)

Lomax – HMM

A continuous probability distribution with two parameters, shape (α) and scale (λ), is the Lomax distribution. Using parameters α = 0.3520 and Λ = 0.0792, we utilize the probability density function of the Lomax distribution to model the emission probabilities within the HMM to peruse agricultural import dynamics in India. The Lomax HMM can be examined by,

$$P\ (x_t\ /y_t = k)\ = \frac{\alpha_k}{\Lambda_k} (1 + \frac{x_t}{\Lambda_k})^{-(\alpha_k + 1)} \quad (20)$$

The Lomax distribution, also known as the Pareto Type II distribution or the generalized Pareto distribution, is a probability distribution commonly used to simulate extreme value events. A Lomax-HMM is thus an HMM in which the emission distributions for each hidden state follow the Lomax distribution. This means that at each time step in the sequence, the observable data is created by a Lomax distribution whose parameters may differ based on the underlying hidden state. The log-likelihood can be calculated using the following modified algorithm:

- Initialization: $\alpha_1(k) = \pi_k\ \frac{\alpha_k}{\Lambda_k} (1 + \frac{x_t}{\Lambda_k})^{-(\alpha_k + 1)}$ (21)
- Recursion: $\alpha_{t+1}(l) = (\ \sum_{k=1}^{N} \alpha_t(k)\ A_{kl})\ \frac{\alpha_k}{\Lambda_k} (1 + \frac{x_t}{\Lambda_k})^{-(\alpha_k + 1)}$ (22)

- Termination: $\log P(x / \lambda) = \log(\sum_{k=1}^{N} \alpha_T(k))$ (23)

log P(x/λ) calculates the log-likelihood of the observed sequence based on Lomax HMM parameters. During the model's training or parameter estimation phase, it is typical to aim to maximize this log-likelihood about the model parameters.

Log-Normal – HMM

A Lognormal-HMM is a probabilistic model that is widely employed in time-series research, particularly when working with data with a log-normal distribution. HMMs are a sort of probabilistic graphical model that is used to represent sequences of observations. They are made up of a series of hidden states that are not directly observable, as well as a series of seen emissions that are affected by the hidden states. With parameters mean of the logarithm (mean log) = 1.4616 and standard deviation of the logarithm (standard deviation log) = 0.3503. The following is the pdf,

$$P(x_t / y_t = k) = \frac{1}{x_k \sigma_k \sqrt{2\pi}} e^{-\frac{(ln\, x_t - \mu_k)^2}{2\sigma_k^2}} \quad (24)$$

Log-normal HMM training involves estimating parameters μ_k and σ_k^2 using approaches such as MLE, EM, or Baum-Welch algorithm, as with other HMM versions. These algorithms seek to optimize the model parameters to maximize the likelihood of the training data given the model. To determine the log-likelihood of a sequence of observations in a Log-Normal Hidden Markov Model, we must modify the method to account for the emission probabilities represented by Log-Normal distributions. We will compute the likelihood using the forward technique:

- Initialization: $\alpha_1(k) = \pi_k \frac{1}{x_k \sigma_k \sqrt{2\pi}} e^{-\frac{(\ln x_t - \mu_k)^2}{2\sigma_k^2}}$ (25)
- Recursion: $\alpha_{t+1}(l) = (\sum_{k=1}^{N} \alpha_t(k)\ A_{kl}) \frac{1}{x_k \sigma_k \sqrt{2\pi}} e^{-\frac{(\ln x_t - \mu_k)^2}{2\sigma_k^2}}$ (26)
- Termination: $\log P(x / \lambda) = \log(\sum_{k=1}^{N} \alpha_T(k))$ (27)

V. Correlation Matrix

A correlation matrix is a square matrix that shows the correlation coefficients between pairs of variables. In statistics, correlation assesses the degree and direction of a linear relationship between two variables. The correlation coefficient, abbreviated ϱ, varies from -1 to 1. To create a correlation matrix with various probability distributions such as Normal, Gamma, Weibull, Lomax, and Lognormal, the correlation between pairs of variables must be specified. Because these distributions reflect continuous random variables, we can use the correlation coefficient between any two variables to determine the strength and direction of their linear link. Here, Statistical R software is used to construct the correlation coefficient between variables X_i and X_j. The correlation matrix serves as a valuable tool for understanding the relationships between different probability distributions, offering insights into their interdependencies and guiding the selection of the most appropriate distributions for modelling agricultural import dynamics within the HMM framework.

VI. Algorithm

The algorithm for the study involves several key steps:

1. Collecting import data for agricultural products from reliable sources covering specific years.
2. Select five statistically appropriate probability distributions for modelling the data.
3. Using R software to determine the parameters of each selected probability distribution.
4. Reconstructing the import data using the values of the parameters.
5. Determine the correlation matrix among the chosen probability distributions to investigate their connections.
6. Use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to conduct stationarity tests to determine whether the data is stationary.
7. Use Hidden Markov Models to model the data. Determine the HMM's parameters, such as the initial probability vector, the emission probability matrix, and the transition probability matrix.
8. When evaluating various models, choose the best-fitting HMM using information metrics such as the Akaike Information Criterion and the Bayesian Information Criterion.
9. In a Hidden Markov Model, figuring out the Viterbi path provides the best estimate of the hidden states that produced the observed data.

These steps are crucial for analyzing agricultural import dynamics and selecting the most appropriate probability distributions within the Hidden Markov Model framework. They provide a systematic approach to understanding the relationships between distributions and their suitability for modelling complex systems.

VII. Data Source

The "Economics & Statistics Division within the Ministry of Agriculture & Farmers Welfare" website, may be accessed at http://desagri.gov.in.

III. Results and Discussion

Table 1: Correlation Coefficient for Different Probability Distributions.

Distribution	Normal	Gamma	Weibull	Lomax	Lognormal
Normal	1.00	-0.5157	0.9975	-0.2504	0.7913
Gamma	-0.5157	1.000	-0.5646	-0.6641	-0.7316
Weibull	0.9975	-0.5646	1.00	-0.2057	0.81337
Lomax	-0.25048	-0.6641	-0.2057	1.00	0.1569
Lognormal	0.7913	-0.7316	0.8133	0.1569	1.00

The correlation table shows notable relationships between probability distributions, such as strong negative correlations between Normal-Gamma and Gamma-Lomax distributions and a strong positive correlation between Weibull and Lognormal distributions, which allows for more informed model selection for dependent variables.

Table 2: Stationarity of KPSS Test Results in Different Distributions

Test For	Test Statistics	Lag Parameter	P-Value	Result
Normal	0.3096	3	0.1	Stationary
Gamma	0.1808	3	0.1	Stationary
Weibull	0.3096	3	0.1	Stationary
Lomax	0.1075	3	0.1	Stationary
Lognormal	0.2245	3	0.1	Stationary

According to the KPSS test results, all distributions exhibit stationarity at a significance level of 0.1. The KPSS test assesses if a time series is stationary around a mean or has a linear trend; in this case, all distributions pass the test, implying that they meet the stationarity assumption. As a result, the data is appropriate for researching HMM.

The parameter ranges for 2, 3, and 4 states are shown in the table below for various HMM. Each model is specifically designed to capture various data distributions, with parameter ranges reflecting import-level characteristics. By comparing these ranges, the table emphasizes the strengths and limitations of each HMM in predicting import dynamics, delivering useful insights into enhancing forecast accuracy.

Table 3: Parameter ranges for various hidden Markov models with different distributions and numbers of states.

States		Normal-HMM	Gamma-HMM	Weibull-HMM	Lomax-HMM	Log-Normal HMM
2	Low Export (S_1)	[0.016, 0.138]	[0.000021, 0.013]	[0.018, 0.132]	[0.008, 0.029]	[0.026, 0.153]
	High Export (S_2)	[0.138, 0.260]	[0.013, 0.027]	[0.132, 0.247]	[0.029, 0.05]	[0.153, 0.280]
3	Low Export (S_1)	[0.016, 0.097]	[0.000021, 0.009]	[0.018, 0.085]	[0.008, 0.021]	[0.026, 0.102]
	Medium Export (S_2)	[0.097, 0.179]	[0.009, 0.018]	[0.085, 0.151]	[0.021, 0.035]	[0.102, 0.177]
	High Export (S_3)	[0.179, 0.260]	[0.018, 0.027]	[0.151, 0.247]	[0.035, 0.050]	[0.177, 0.280]
4	Very low Export (S_1)	[0.016, 0.077]	[0.00002, 0.006]	[0.018, 0.075]	[0.008, 0.018]	[0.028, 0.090]
	Low Export (S_2)	[0.077, 0.138]	[0.006, 0.0125]	[0.075, 0.132]	[0.018, 0.029]	[0.090, 0.153]
	High Export (S_3)	[0.138, 0.199]	[0.012, 0.020]	[0.132, 0.189]	[0.029, 0.040]	[0.153, 0.217]
	Very High Export (S_4)	[0.199, 0.260]	[0.020, 0.027]	[0.189, 0.247]	[0.040, 0.050]	[0.217, 0.280]

This table highlights the delicate impact of distribution and state count variations on export-level modelling in HMM. From Normal to Log-Normal distributions, and from two to four states, each configuration provides unique probabilistic insights on export behaviour, underscoring the need for careful model selection in effectively representing export dynamics.

In a 2-State Hidden Markov Model, the framework is divided into two states: "Low" and "High" with observations showing whether import levels are "Decreasing (D)" or "Increasing (I)". This setup provides a broad view of trends. The 3-State Model adds a "Moderate" state, allowing for a more nuanced analysis where observations can be "Decreasing", "Stable (S)," or "Increasing" capturing a range of fluctuations between low and high import levels. The 4-State Model introduces even more detail with states like "Very Low", "Low," "High," and "Very High," and observations of "Very Decreasing (VD)", "Decreasing", "Increasing", and "Very Increasing (VI)" offering a comprehensive look at significant changes and trends in import levels.

TPM, EPM, and π for Normal-HMM (2,3 & 4 States)

$$\text{TPM} = \begin{matrix} & s_1 & s_2 \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.4 & 0.6 \\ 0.238 & 0.727 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & I \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.1 & 0.9 \\ 0.59 & 0.40 \end{bmatrix},\quad \pi_0 = [\,0.312\,, 0.687]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0.5 & 0 & 0.5 \\ 0.076 & 0.461 & 0.461 \\ 0.071 & 0.428 & 0.5 \end{bmatrix},\text{EPM} = \begin{matrix} & D & S & I \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0.25 & 0 & 0.75 \\ 0.15 & 0.07 & 0.76 \\ 0.2 & 0.53 & 0.26 \end{bmatrix},\ \pi_0 = [\,0.125, 0.406, 0.468]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 & s_4 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0.33 & 0 & 0 & 0.66 \\ 0.16 & 0.33 & 0.16 & 0.33 \\ 0.11 & 0.11 & 0.33 & 0.44 \\ 0 & 0.15 & 0.38 & 0.46 \end{bmatrix},\ \text{EPM} = \begin{matrix} & VD & D & I & VI \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0 & 0 & 0.33 & 0.66 \\ 0 & 0 & 0.66 & 0.33 \\ 0 & 0.2 & 0.55 & 0.22 \\ 0.5 & 0.2 & 0.28 & 0 \end{bmatrix},\ \pi_0 = [0.09,\ 0.18,\ 0.28,\ 0.43]$$

TPM, EPM, and π for Gamma-HMM (2,3 & 4 States)

$$\text{TPM} = \begin{matrix} & s_1 & s_2 \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.93 & 0.03 \\ 0.5 & 0.5 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & I \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.43 & 0.56 \\ 0.5 & 0.5 \end{bmatrix},\quad \pi_0 = [\,0.937\,, 0.062]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0.846 & 0.115 & 0.03 \\ 1 & 0 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & S & I \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0.18 & 0.18 & 0.62 \\ 0.66 & 0.33 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix},\ \pi_0 = [\,0.843, 0.093, 0.062]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 & s_4 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0.82 & 0.09 & 0.09 & 0 \\ 0.4 & 0.4 & 0 & 0.2 \\ 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix},\text{EPM} = \begin{matrix} & VD & D & I & VI \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0.17 & 0.13 & 0.35 & 0.35 \\ 0.2 & 0.6 & 0.2 & 0 \\ 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix},\ \pi_0 = [0.72,\ 0.16,\ 0.63,\ 0.06]$$

TPM, EPM, and π for Weibull-HMM (2,3 & 4 States)

$$\text{TPM} = \begin{matrix} & s_1 & s_2 \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.33 & 0.66 \\ 0.16 & 0.84 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & I \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.16 & 0.83 \\ 0.5 & 0.5 \end{bmatrix},\quad \pi_0 = [\,0.187, 0.812]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0.33 & 0 & 0.66 \\ 0.14 & 0.28 & 0.57 \\ 0.04 & 0.19 & 0.76 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & S & I \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{bmatrix} 0 & 0.33 & 0.66 \\ 0 & 0 & 1 \\ 0.41 & 0.18 & 0.41 \end{bmatrix},\ \pi_0 = [\,0.09, 0.218, 0.68]$$

$$\text{TPM} = \begin{matrix} & s_1 & s_2 & s_3 & s_4 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0.25 & 0.25 & 0.5 \\ 0 & 0.18 & 0.36 & 0.45 \\ 0.07 & 0.07 & 0.35 & 0.5 \end{bmatrix},\text{EPM} = \begin{matrix} & VD & D & I & VI \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{bmatrix} 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0.18 & 0.45 & 0.36 \\ 0.46 & 0.26 & 0.26 & 0 \end{bmatrix},\ \pi_0 = [0.06, 0.13,\ 0.34,\ 0.46]$$

TPM, EPM, and π for Lomax-HMM (2,3 & 4 States)

$$\text{TPM} = \begin{matrix} & s_1 & s_2 \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.92 & 0.08 \\ 0.37 & 0.62 \end{bmatrix},\ \text{EPM} = \begin{matrix} & D & I \\ s_1 \\ s_2 \end{matrix}\begin{bmatrix} 0.45 & 0.54 \\ 0.87 & 0.13 \end{bmatrix},\quad \pi_0 = [\,0.75, 0.25]$$

$$\text{TPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{matrix} \begin{matrix} s_1 & s_2 & s_3 \end{matrix} \\ \begin{bmatrix} 0.78 & 0.21 & 0 \\ 0.4 & 0.5 & 0.1 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{matrix}, \text{EPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{matrix} \begin{matrix} \text{D} & \text{S} & \text{I} \end{matrix} \\ \begin{bmatrix} 0.05 & 0.35 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.5 & 0.5 & 0 \end{bmatrix} \end{matrix}, \pi_0 = [\,0.63, 0.315, 0.062]$$

$$\text{TPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{matrix} \begin{matrix} s_1 & s_2 & s_3 & s_4 \end{matrix} \\ \begin{bmatrix} 0.81 & 0.13 & 0.06 & 0 \\ 0.43 & 0.43 & 0.14 & 0 \\ 0.17 & 0.17 & 0.5 & 0.17 \\ 0 & 0.5 & 0 & 0.5 \end{bmatrix} \end{matrix}, \text{EPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{matrix} \begin{matrix} \text{VD} & \text{D} & \text{I} & \text{VI} \end{matrix} \\ \begin{bmatrix} 0.05 & 0.35 & 0.41 & 0.17 \\ 0.14 & 0.28 & 0.57 & 0 \\ 0.17 & 0.66 & 0.17 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{bmatrix} \end{matrix}, \pi_0 = [0.53, 0.22,\ 0.18,\ 0.06]$$

TPM, EPM, and π for Lognormal-HMM (2,3 & 4 States)

$$\text{TPM} = \begin{matrix} \\ s_1 \\ s_2 \end{matrix}\begin{matrix} \begin{matrix} s_1 & s_2 \end{matrix} \\ \begin{bmatrix} 0.54 & 0.45 \\ 0.25 & 0.75 \end{bmatrix} \end{matrix}, \text{EPM} = \begin{matrix} \\ s_1 \\ s_2 \end{matrix}\begin{matrix} \begin{matrix} \text{D} & \text{I} \end{matrix} \\ \begin{bmatrix} 0.18 & 0.82 \\ 0.52 & 0.47 \end{bmatrix} \end{matrix}, \quad \pi_0 = [\,0.343, 0.656]$$

$$\text{TPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{matrix} \begin{matrix} s_1 & s_2 & s_3 \end{matrix} \\ \begin{bmatrix} 0.28 & 0.14 & 0.57 \\ 0.5 & 0.25 & 0.25 \\ 0.15 & 0.1 & 0.75 \end{bmatrix} \end{matrix}, \text{EPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \end{matrix}\begin{matrix} \begin{matrix} \text{D} & \text{S} & \text{I} \end{matrix} \\ \begin{bmatrix} 0 & 0.14 & 0.85 \\ 0.25 & 0 & 0.75 \\ 0.38 & 0.14 & 0.47 \end{bmatrix} \end{matrix}, \pi_0 = [\,0.22, 0.12, 0.65]$$

$$\text{TPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{matrix} \begin{matrix} s_1 & s_2 & s_3 & s_4 \end{matrix} \\ \begin{bmatrix} 0.33 & 0.16 & 0.16 & 0.33 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.33 & 0 & 0.16 & 0.5 \\ 0.07 & 0.14 & 0.14 & 0.64 \end{bmatrix} \end{matrix}, \text{EPM} = \begin{matrix} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{matrix} \begin{matrix} \text{VD} & \text{D} & \text{I} & \text{VI} \end{matrix} \\ \begin{bmatrix} 0 & 0.16 & 0.16 & 0.66 \\ 0 & 0.2 & 0.2 & 0.6 \\ 0.33 & 0 & 0.5 & 0.16 \\ 0.4 & 0.26 & 0.33 & 0 \end{bmatrix} \end{matrix}, \pi_0 = [0.18, 0.15,\ 0.18,\ 0.46]$$

Table 4: Model Comparison for Various Distributions Based on States and Information Criteria.

Distribution	States	Log-Likelihood	AIC	BIC
Normal	2	-21.621	63.243	77.902
	3	-32.522	107.04	137.82
	4	-46.865	165.73	218.49
Gamma	2	-26.886	73.772	88.429
	3	-43.243	128.48	159.26
	4	-43.828	159.65	212.42
Weibull	2	-19.714	59.428	74.085
	3	-29.521	101.04	131.82
	4	-45.444	162.88	215.65
Lomax	**2**	**-19.136**	**58.272**	**72.930**
	3	-43.226	128.45	159.23
	4	-43.240	158.48	211.24
Lognormal	2	-20.448	60.896	75.553
	3	-27.879	97.759	128.54
	4	-42.204	156.40	209.17

Based on the AIC and BIC values provided, the Lomax distribution appears to be the best fit among the distributions listed. This conclusion is based on the consistently lower AIC and BIC values across a variety of states when compared to other distributions.

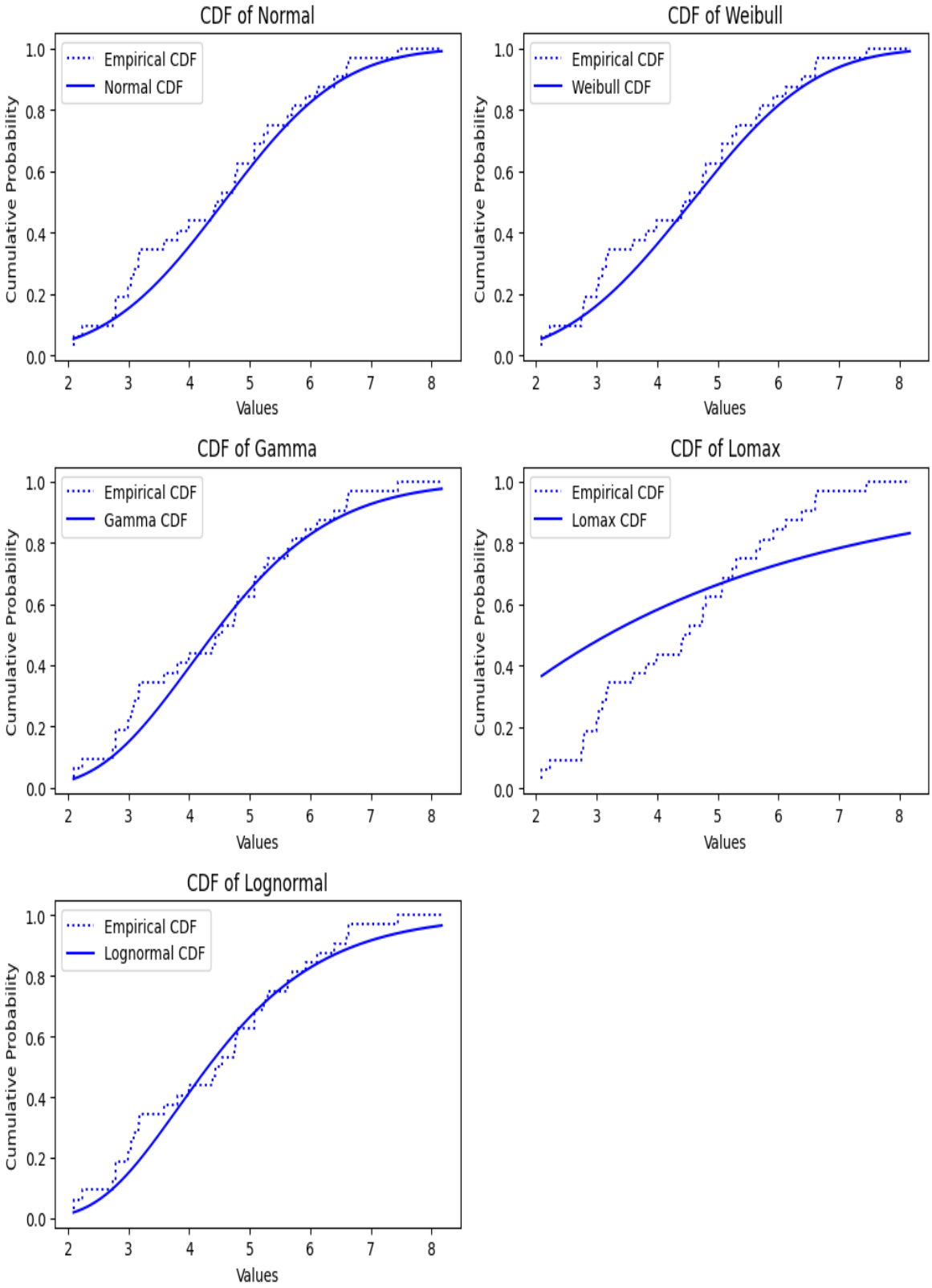

Figure 1: The graphs show the comparative evaluation of CDFs using HMM in different distributions.

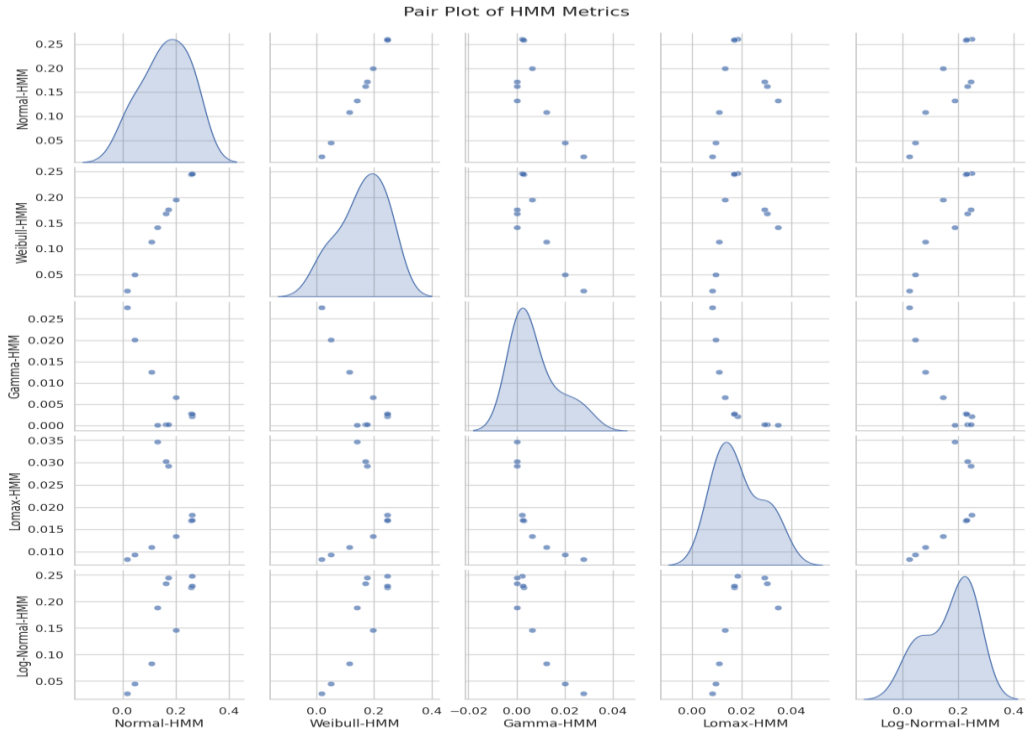

Figure 2: This graph shows the pair plot of different HMM distributions for comparison

The pair plot is a helpful visualization that shows the relationships between the performance metrics of various Hidden Markov Models. The scatter plots reveal how different models are correlated, while the diagonal plots indicate the distribution of each model's performance. This enables a comparison of trends, correlations, and distribution patterns, assisting in identifying the model that performs the best.

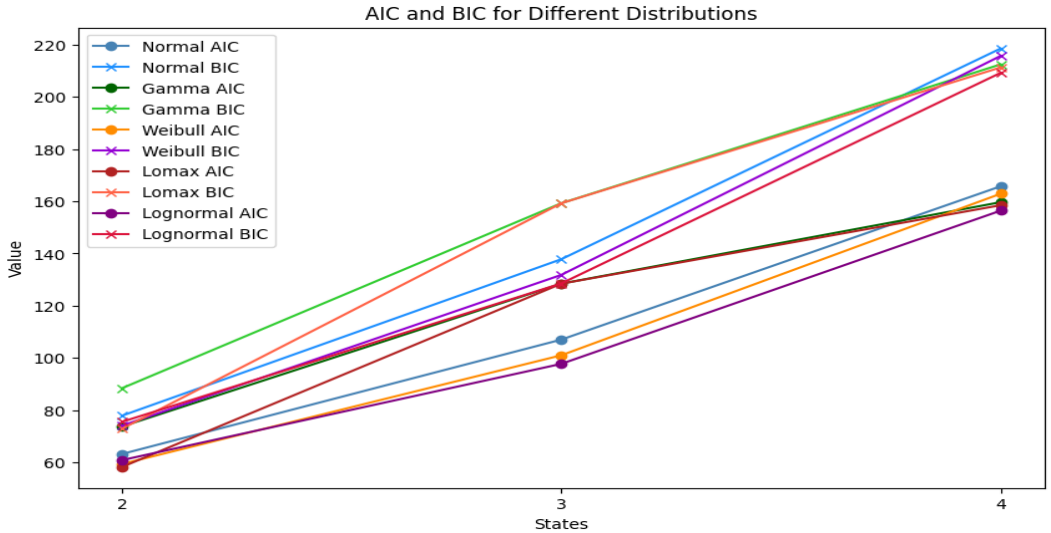

Figure 4: Graphical representation of comparative analysis of AIC and BIC values for different probability distributions.

IV. Conclusion

In conclusion, our findings shed light on the selection of probability distributions for Hidden Markov Models. We recognized the Lomax distribution as extremely beneficial, particularly within

the HMM's state 2. This conclusion emphasizes the need to consider multiple distributions when modelling complex systems. Furthermore, our investigation of the Viterbi path demonstrates the practical relevance of our findings in determining the most likely state sequence. The Viterbi path is 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2. Overall, this study helps advance our understanding of probability distributions in HMMs.

References

[1] C.D. Mitchell and L.H. Jamieson. (1993). Modeling duration in a hidden Markov model with the exponential family. IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 1993, pp. 331-334 vol.2.

[2] Cota, Napat, Teerasit Kasetkasem, Preesan Rakwatin, Thitiporn Chanwimaluang, and Itsuo Kumazawa. (2015). Rice phenology estimation based on statistical models for time-series SAR data. In 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pp. 1-6. IEEE.

[3] Hao Zhang, Weidong Zhang, Ahmet Palazoglu and Wei Sun. (2012). Prediction of ozone levels using a Hidden Markov Model (HMM) with Gamma distribution. Atmospheric Environment, Volume 62, Pages 64-73, ISSN 1352-2310.

[4] Hui Zhang, Qing Ming Jonathan Wu and Thanh Minh Nguyen. (2013). Modified student's t-hidden Markov model for pattern recognition and classification. IET Signal Processing, Vol. 7, Iss. 3, pp. 219–227.

[5] Joshni George and Seemon Thomas. (2019). Negative Binomial Hidden Markov Model for AES count data. Think India Journal, Vol-22- Issue-14, ISSN:0971-1260.

[6] Lethanh, Nam, Kiyoyuki Kaito, and Kiyoshi Kobayashi. (2015). Infrastructure deterioration prediction with a Poisson hidden Markov model on time series data. Journal of Infrastructure Systems 21.3, 04014051.

[7] MacDonald I. L, and Bhamani F. (2018). A time-series model for underdispersed or overdispersed counts. The American Statistician, Taylor & Francis.

[8] Mahalakshmi Rajendran, Sentamarai Kannan Kaliyaperumal and Balasubramaniam Ramakrishnan. (2021). Hidden Markov Model of Evaluation of Break-Even Point of HIV patients: A Simulation Study. International Journal of Medical Sciences and Nursing Research,1(2):19-22.

[9] Marisa, Ukur Arianto Sembiring and Helena Margaretha. (2019). Earthquake Probability Prediction in Sumatra Island Using Poisson Hidden Markov Model (HMM). AIP Conf. Proc. 2192, 090006-1–090006-12.

[10] Mingyuan Zhou and Lawrence Carin. (2013). Negative binomial process count and mixture modelling. IEEE Transactions on Pattern Analysis and Machine Intelligence ,37.2, 307-320.

[11] Sarvi F, Nadali A, Khodadost M, Moghaddam M. K, and Sadeghifar M. (2017). Application of Poisson Hidden Markov model to predict the number of PM2. 5 exceedance days in Tehran during 2016-2017. Avicenna Journal of Environmental Health Engineering, 4(1), 58031-58031.

[12] Satu Helske and Jouni Helske. (2019). Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R, Journal of Statistical Software, 88.

[13] Sebastian George and Ambily Jose. (2020). Generalized Poisson Hidden Markov Model for Overdispersed or Underdispersed Count Data. Revista Colombiana de Estadística, Volume 43, Issue 1, pp. 71 to 82.

[14] Sebastian T, Jeyaseelan V, Jeyaseelan L, Anandan S, George S, and Bangdiwala S.I. [2019]. Decoding and modelling of time series count data using Poisson hidden Markov model and Markov ordinal logistic regression models. Statistical Methods in Medical Research, 28(5):1552-1563.

[15] Takashi Kaburagi, and Takashi Matsumoto. (2008). A generalized hidden Markov model approach to transmembrane region prediction with Poisson distribution as state duration probabilities. IPSJ Digital Courier, 4: 193-206.