International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 6, No. 1 (May-2019)

http://dx.doi.org/10.12785/ijcts/060103

Modeling of Agricultural Price Data Using Hidden Markov Model

Joshni George[1] and Seemon Thomas[2]

[1,2] Department of Statistics, St.Thomas College Palai, Kottayam, Kerala , India.

Received May 3, 2018, Revised September 15, 2018, Accepted November 28,2018 Published May 1, 2019

Abstract: In this paper, we explore the application of hidden Markov model (HMM) in the modeling of agricultural price data. Normal hidden Markov model is fitted and compared with univariate autoregressive moving average (ARMA) model. The parameters of the model are estimated using EM algorithm and the sequence of hidden states are obtained based on the best fitted model.

Keywords: Markov Process, Transition Probability Matrix, Stationary Distribution, Decoding

1. INTRODUCTION

Agricultural prices are highly volatile as they are largely influenced by so many unpredictable and irregular factors that are random in nature. Presently the agricultural prices are determined by the domestic and international market forces. Therefore price forecasting is a challenging and difficult task due to the stochastic and irregular form of the price.

The two data sets considered in this study are the global monthly average price of banana during the period January 2011 to December 2016 and local price of banana during January 2013 to December 2017. During the past few decades, one of the most important and widely used time series model for price forecasting is the ARMA model which is based on normal distribution. The time series data of global and local banana prices have kurtosis less than that of a normal distribution. In this situation, we explore the possibility of applying hidden Markov model (HMM) in the modeling of banana price and compare it with ARMA model, the most commonly used model for the analysis of price data.

2. HIDDEN MARKOV MODEL

An HMM is a stochastic model with an underlying Markov process that is not directly observable. But the process can be observable through another stochastic process that depends on hidden states constitute a sequence of observations. The theory of HMM was introduced by [1] as an extension to the first order Markov process. In the late 1980s and early 1990s, HMMs were introduced to computational sequence analysis [3] and applied HMM in DNA sequence analysis. HMMs have been widely used in modern continuous speech recognition systems [5], biological sequence analysis [4], speech recognition [8], computational finance [7] , gene prediction [6] etc. A description of parameter estimation of HMMs can be found in [2]. An elaborate discussion of the theory, application and computation of the HMMs are available in [10].

In an HMM, the states are not directly visible, but the output depends on hidden states is observable. Let $\{S_t : t = 1,2,...\}$ represents the unobserved parameter process satisfying Markov property. Let $\{Z_t : t = 1,2,...\}$ is the state-dependent process such that, when S_t is known, the distribution of Z_t depends only on current state S_t and not on previous states.

Let $p_i(z)$ be the probability density function of Z_t if the Markov chain is in the current state i at time t.

E-mail address: joshni.tony@gmail.com, seemonpala@rediffmail.com

That is

$$p_i(z) = P(Z_t = z \mid S_t = i).$$

Now define $u_i(t) = P(S_t = i)$ for $t = 1, \ldots, n$. Then

$$\begin{aligned} P(Z_t = z) &= \sum_{i=1}^{m} P(S_t = i) P(Z_t = z \mid S_t = i) \\ &= \sum_{i=1}^{m} u_i(t) p_i(z) \\ &= (u_1(t), \ldots, u_m(t)) \begin{pmatrix} p_1(z) & & 0 \\ & \ddots & \\ 0 & & p_m(z) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \mathbf{u}(t)\mathbf{P}(z)\mathbf{1}'. \end{aligned}$$

where $\mathbf{u}(t) = (u_1(t), \ldots, u_m(t))$ is the initial distribution of the Markov chain, $\mathbf{P}(z)$ is the diagonal matrix with i^{th} diagonal element $p_i(z)$ and $\mathbf{1} = (1 \ldots 1)$. If $\mathbf{u}(t)$ is the initial distribution and $\mathbf{\Gamma} = (\gamma_{ij})$ is the transition probability matrix of the Markov chain such that $\sum_{j=1}^{m} \gamma_{ij} = 1$ and $\gamma_{ij} \geq 0$. Then the distribution at time $t+1$ can be given as follows:

$$\mathbf{u}(t+1) = \mathbf{u}(t)\mathbf{\Gamma}.$$

If the Markov chain has stationary distribution $\boldsymbol{\delta} = (\delta_1 \ldots \delta_m)$, then $\boldsymbol{\delta}\mathbf{\Gamma}^{t-1} = \boldsymbol{\delta}$ for all $t \in \mathbb{N}$ and hence

$$\mathbf{P}(Z_t = z) = \boldsymbol{\delta}\mathbf{P}(z)\mathbf{1}'.$$

If $z_1, z_2, \ldots, z_n$ is the observation sequence generated by an HMM and $\boldsymbol{\delta}$ be the initial distribution which is assumed to be the same as the stationary distribution implied by the transition probability matrix $\mathbf{\Gamma}$, then the likelihood function is the following:

$$L = \boldsymbol{\delta}\mathbf{\Gamma}\mathbf{P}(z_1)\mathbf{\Gamma}\mathbf{P}(z_2)\ldots\mathbf{\Gamma}\mathbf{P}(z_n)\mathbf{1}'. \qquad (1)$$

A. Normal-HMM

Consider an m-state HMM with a Markov chain having transition probability matrix $\mathbf{\Gamma}$, stationary distribution $\boldsymbol{\delta}$ and univariate normal state dependent distribution

$$p_i(z) = (2\pi\sigma_i^{\,2})^{-1/2} exp(-\frac{1}{2\sigma_i^{\,2}}(z-\mu_i)^2); \qquad -\infty < z < \infty,\ -\infty < \mu_i < \infty,$$

and $\sigma_i > 0$. The parameters are estimated using EM algorithm. A detailed description of the iterative procedure involved in EM algorithm is available in [10]. The appropriate number of states m is decided on the basis of Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. AIC and BIC are defined as follows:

$$\begin{aligned} AIC &= -2 \log L + 2\,p \\ BIC &= -2 \log L + p \, log\, n \end{aligned}$$

where $logL$ is the log-likelihood of the model, p is the number of parameters and n is the total number of observations. The mean and variance of the distribution p_i is given below.

$$
\begin{aligned}
E(Z_t) &= \sum_{i=1}^{m}\delta_i\mu_i = \boldsymbol{\delta\mu}'. \\
\mathrm{Var}(Z_t) &= \sum_{i=1}^{m}\delta_i(\sigma_i^{\,2}+\mu_i^{\,2})-(\boldsymbol{\delta\mu}')^2.
\end{aligned} \tag{2}
$$

Using Viterbi algorithm, a decoding algorithm for finding the most likely sequence of hidden states conceived by [9], one can find the best state sequence with respect to the sequence of observations which maximize the likelihood function (1).

3. MODELING OF BANANA PRICE

A. Global Price

The data given in Table I is the monthly average global price of banana per metric tonne in USDollar during the period January 2011 to December 2016. The data is available at *https://fred.stlouisfed.org/series/PBANSOPUSDM* .

TABLE I. GLOBAL BANANA PRICES FROM JANUARY 2011 TO DECEMBER 2016 IN USD/METRIC TONNE.

Year	Jan	Feb	Mar	Apr	May	Jun
2011	949.42	1013.12	994.17	1013.47	1022.10	986.31
2012	942.99	1045.12	1151.43	1029.33	955.44	956.33
2013	933.82	925.41	938.57	912.23	908.00	911.60
2014	928.42	946.13	966.85	945.50	916.00	926.07
2015	911.60	966.85	1045.96	1059.14	946.79	929.18
2016	1011.25	1052.35	1020.90	998.09	987.57	993.22
	Jul	Aug	Sep	Oct	Nov	Dec
2011	964.22	960.25	945.50	956.33	959.32	946.76
2012	966.85	954.84	966.85	959.04	937.34	945.80
2013	925.41	937.60	939.23	936.82	922.13	925.41
2014	930.82	961.59	925.41	922.41	904.70	908.60
2015	938.33	956.66	951.78	933.58	932.32	932.32
2016	1004.67	1051.53	1013.31	976.39	959.94	959.94

The autocorrelation function (ACF) of the data given in Table 1 is given in Fig. 1.

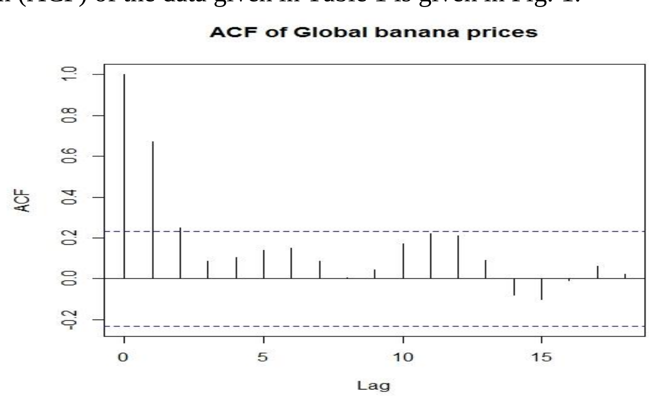

Figure 1. ACF of Global Price

The banana price series has a kurtosis of 2.7021 which is less than that of a normal distribution. In addition, the price series shows serial correlation for the first lag. In this situation, we model the price data using normal mixture models, assuming both independent observations and Markov dependent mixture models, known as normal-HMM. Let us fit normal-HMM with two, three and four states and Box-Jenkin's ARMA model to the data. Fitting of a normal-

HMM involves estimation of $\boldsymbol{\delta}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\Gamma}$ by maximising the likelihood function (1). AIC and BIC values of each fitted model are given in Table II.

TABLE II. COMPARISON OF FITTED MODELS BY AIC AND BIC.

Model	- log L	AIC	BIC
2-state HMM	345.8483	701.6966	713.0799
3-state HMM	332.8115	687.6230	712.6664
4-state HMM	330.7309	699.4618	742.7185
ARMA(1,1)	347.8700	703.7400	712.8400

On comparing AIC and BIC values one can see that 3-state normal-HMM is the best fitted model. For the fitted 3-state normal-HMM, the estimate of the transition probability matrix $\boldsymbol{\Gamma}$ obtained is the following:

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8319 & 0.1113 & 0.0568 \\ 0.2350 & 0.6491 & 0.1159 \\ 0.0000 & 0.1961 & 0.8039 \end{pmatrix}.$$

The corresponding estimates of $\boldsymbol{\delta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are shown in Table III.

TABLE III. STATIONARY DISTRIBUTION AND PARAMETERS OF 3-STATE NORMAL-HMM.

Parameter	State 1	State 2	State 3
$\boldsymbol{\delta}$	0.4119	0.2946	0.2935
$\boldsymbol{\mu}$	927.6933	956.1636	1020.9425
$\boldsymbol{\sigma}$	12.1558	7.8888	39.2086

The mean and variance of 3-state normal-HMM computed using equations (2) are 963.4493 and 2042.8030 respectively. Note that these values are very close to the sample mean 963.2601 and sample variance 2040.2260 .

Prediction of the most likely sequence of Markov states given the observed data set (decoding) of 3-state normal-HMM is done using Viterbi algorithm and is given in Table IV.

TABLE IV. THE MOST LIKELY SEQUENCE OF HIDDEN STATES OF 3-STATE NORMAL-HMM.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	2	3	3	3	3	3	2	2	2	2	2	2
2012	2	3	3	3	2	2	2	2	2	2	1	1
2013	1	1	1	1	1	1	1	1	1	1	1	1
2014	1	2	2	2	1	1	1	2	1	1	1	1
2015	1	2	2	3	2	1	1	2	2	1	1	1
2016	3	3	3	3	3	3	3	3	3	3	2	2

The state predictions of banana prices in the months of the year 2017 based on 3-state normal-HMM is given in Table V.

TABLE V. STATE PREDICTION USING 3-STATE HMM: THE PROBABILITY THAT THEMARKOV CHAIN WILL BE IN A GIVEN STATE IN THE SPECIFIED MONTH OF 2017.

Year 2017	Jan	Feb	Mar	Apr	May	Jun
State=1	0.2011	0.2997	0.3513	0.3795	0.3956	0.4051
State=2	0.5888	0.4154	0.3362	0.2994	0.2818	0.2732
State=3	0.2101	0.2849	0.3125	0.3211	0.3226	0.3217

	Jul	Aug	Sept	Oct	Nov	Dec
State=1	0.4109	0.4150	0.4168	0.4183	0.4192	0.4198
State=2	0.2688	0.2664	0.2651	0.2644	0.2639	0.2637
State=3	0.3203	0.3191	0.3181	0.3173	0.3169	0.3165

B. Local Price

Let us consider the maximum of monthly banana prices per 100 kg of our own local market, namely Pala, during the period January 2013 to December 2017. The data is available at *http://agmarknet.gov.in* and given in Table VI.

TABLE VI. LOCAL BANANA PRICES FROM JANUARY 2013 TO DECEMBER 2017 IN RUPEES/100 KG.

Year	Jan	Feb	Mar	Apr	May	Jun
2013	4200	4200	4300	4200	3600	4600
2014	4200	3400	3000	3400	3400	3400
2015	4400	3800	3800	3200	3800	3800
2016	3400	3600	3600	5400	5400	6400
2017	5800	5800	4800	5200	4700	4700
	Jul	Aug	Sep	Oct	Nov	Dec
2013	5400	5800	6000	5600	5400	4400
2014	3800	6400	6800	6800	5000	4600
2015	3800	5200	5400	4400	3800	3800
2016	6800	7400	7400	4000	4400	4400
2017	4400	6000	6800	5000	5000	4500

The autocorrelation function (ACF) of the data given in Table 3.2.1 is given in Fig. 2.

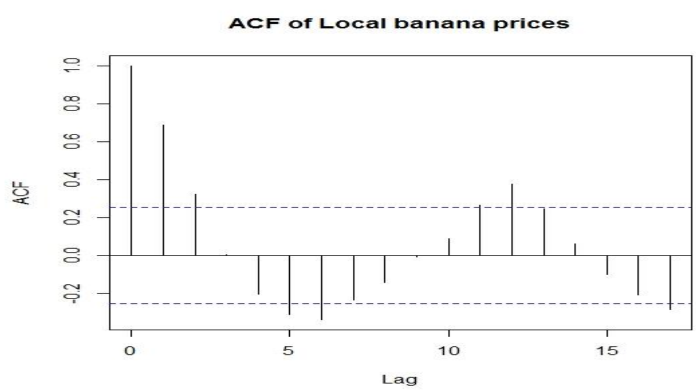

Figure 2. ACF of Local price

Here also we fit normal-HMM with two, three and four states and Box-Jenkin's ARMA model. AIC and BIC values of the fitted models are given in Table VII.

TABLE VII. COMPARISON OF FITTED MODELS BY AIC AND BIC.

Model	- log L	AIC	BIC
2-state HMM	488.8214	989.6428	1002.1145
3-state HMM	474.1935	972.3870	997.5191
4-state HMM	468.1055	976.2111	1018.0980
ARMA(2,1)	483.5400	977.0900	987.5600

AIC value selects 3-state HMM whereas BIC selects ARMA model for the data. The transition probability matrix Γ for the fitted 3-state normal-HMM is the following:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.7899 & 0.0000 & 0.2101 \\ 0.3691 & 0.5148 & 0.1161 \\ 0.0000 & 0.1624 & 0.8376 \end{pmatrix}.$$

The corresponding estimates of $\boldsymbol{\delta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are shown in Table VIII.

TABLE VIII. STATIONARY DISTRIBUTION AND PARAMETERS OF 3-STATE NORMAL-HMM.

Parameter	State 1	State 2	State 3
$\boldsymbol{\delta}$	0.3333	0.1667	0.5000
$\boldsymbol{\mu}$	3569.791	4318.944	5630.15
$\boldsymbol{\sigma}$	239.6449	132.1290	885.4618

The mean and standard deviation of 3-state normal-HMM computed using equation (2) are 4724.854 and 1138.419 respectively which are close to the sample mean 4733.333 and sample standard deviation 1135.956.

Prediction of the most likely sequence of Markov states given the observed data set (decoding) of 3-state normal-HMM is done using Viterbi algorithm and is given in Table IX.

TABLE IX. THE MOST LIKELY SEQUENCE OF HIDDEN STATES OF 3-STATE NORMAL-HMM.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013	2	2	2	2	1	3	3	3	3	3	3	2
2014	2	1	1	1	1	1	1	3	3	3	3	3
2015	2	1	1	1	1	1	1	3	3	2	1	1
2016	1	1	1	1	3	3	3	3	3	2	2	2
2017	3	3	3	3	3	3	3	3	3	3	3	3

The state predictions of banana prices in the months of the year 2018 based on 3-state normal-HMM is given in Table X.

TABLE X. STATE PREDICTION USING 3-STATE HMM: THE PROBABILITY THAT THE MARKOV CHAIN WILL BE IN A GIVEN STATE IN THE SPECIFIED MONTH OF 2018.

Year 2018	Jan	Feb	Mar	Apr	May	Jun
State=1	0.1847	0.2646	0.3003	0.3168	0.3247	0.3287
State=2	0.2690	0.2076	0.1830	0.1732	0.1693	0.1677
State=3	0.5463	0.5278	0.5167	0.5100	0.5060	0.5036
	Jul	Aug	Sept	Oct	Nov	Dec
State=1	0.3308	0.3319	0.3325	0.3328	0.3330	0.3332
State=2	0.1671	0.1668	0.1667	0.1667	0.1667	0.1667
State=3	0.5021	0.5013	0.5008	0.5005	0.5003	0.5001

CONCLUSION

For the global banana price data, the best fitted model is found to be the 3-state normal-HMM having stationary distribution $\boldsymbol{\delta}$ = (0.4119 0.2946 0.2935), state dependent mean vector $\boldsymbol{\mu}$ = (927.6933 956.1636 1020.9425) and $\boldsymbol{\sigma}$ = (12.1558 7.8888 39.2086). On studying the Viterbi path of states of 3-state HMM in relation with the prices, it is found that state 1 corresponds to less than 940 USD, state 2 corresponds to price between 940 USD and 970 USD and state 3 corresponds to more than 970 USD. From Table V it can be seen that the probability that the Markov chain will be in state 2 is high for the months January and February and the probability is high for state 1 for the rest of the months.

The best fitted model for the local banana price is found to be the 3-state normal-HMM having stationary distribution $\boldsymbol{\delta}$ = (0.3333 0.1667 0.5000), state dependent mean vector $\boldsymbol{\mu}$ = (3569.7910 4318.9440 5630.15) and $\boldsymbol{\sigma}$ =

(239.6449 132.1290 885.4618). On studying the Viterbi path of states of 3-state HMM in relation with the local market prices, it is found that state 1 corresponds to less than 4000 rupees, state 2 corresponds to price between 4000 rupees and 4400 rupees and state 3 corresponds to more than 4400 rupees.

The present study reveals that HMMs can be effectively used to model and study the hidden factors (states) which affect the prices of agricultural commodities.

REFERENCES

[1] L. E. Baum and T. Petrie, ``Statistical inference on probabilistic functions of finite state Markov chains", Annals of Mathematical Statistics, 37, 1966, pp. 1554-1563.

[2] L. E. Baum, T. Petrie, G. Soules and N. Weiss, ``A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains", Annals of Mathematical Statistics, 41(1), 1970, pp. 164-171.

[3] G. A. Churchill, `` Stochastic Models for Heterogeneous DNA sequences", Bull Math Biol, 51(1), 1989, pp. 79-94.

[4] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

[5] M. Gales and S. Young, ``The application of hidden Markov models in speech recognition", Foundations and Trends in Signal Processing 1 (3), 2007, pp.195-304.

[6] K. Munch and A. Krogh, Automatic generation of gene finders for eukaryotic species. BMC Bioinformatics, 7, 2006, pp.263.

[7] A. Petropoulos, S. P. Chatzis and S. Xanthopoulos, "A Novel Corporate Credit Rating System Based on Student's-t Hidden Markov Models," Expert Systems with Applications, 53, 2016, pp. 87-105.

[8] L. R. Rabiner, ``A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 77, 1989, pp.257-286.

[9] A. J. Viterbi , ``Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Transactions on Information Theory, 13, 1967, pp.260-269.

[10] W. Zucchini and I. L. MacDonald, Hidden Markov Models for time series: An Introduction Using R, Monographs on statistics and applied probability ; 110, 2009.

[11] https://fred.stlouisfed.org/series/PBANSOPUSDM

[12] http://agmarknet.gov.in