Predicting raga notes through an ARIMA model: A Case Study in raga Bhupali

Moujhuri Patra[1], Soubhik Chakraborty[2] and Dipak Ghosh[3]
[1]Department of MCA, Netaji Subhash Engineering College, Kolkata-700152, India
[2]Department of Applied Mathematics, B. I. T. Mesra, Ranchi- 835215, India
[3]C V Raman Centre for Physics and Music, Jadavpur University, Kolkata-700 032,India

ABSTRACT: From the aesthetic side, emotion and devotion are the essential characteristics of Indian music. Here we take only the technical side of the music, talking about its pitch which is nothing but a function of time. In the present work, we build a structure of raga notes through ARIMA modeling and predict its nature through this model.
KEYWORDS: Time series, modeling, prediction, musical notes, raga

Introduction

A great deal of attention has been paid in the recent past in modeling musical structure and performance [BM99]. Musical data will always be the function of time and the stochastic processes such as the time series ARIMA models can be used as predictive models in music. Here we analyze and identify the nature of the phenomenon represented by the sequence of pitch and forecasting (predicting future values of the time series variable).Finally we build a predictive model of Bhupali, a north Indian raga using Box and Jenkins ARIMA model. Raga has been defined in section 6. See also appendix.

The paper is organized as follows. Section 1 gives a brief outline on the identification of ARIMA model. Sections 1.1 and 1.2 describe

autoregressive components and moving average components respectively. Section 1.3 describes a mixed model. Section 2 discusses about ACFs and PACFs (autocorrelation functions and partial autocorrelation functions respectively). Section 3 describes the general idea of estimating the model parameters. In Section 4, we describe the diagnosis of a model and Ljung-Box (Q) Statistic for diagnostic checking. Section 5 describes the modeling structure of raga Bhupali. Section 6 is for results and discussion and the last one is reserved for conclusion. For comprehensive literature on time series analysis and ARIMA, we refer to [BJ76] and the following website:
http://wps.ablongman.com/wps/media/objects/2829/2897573/ch18.pdf

1. Identification of ARIMA (p, d, q) Models

The ARIMA (Auto-regressive, integrated, moving average) model of a time series is define by three terms (p, d, q). The meaning of these terms will be explained later (sec 2.1 and 2.2). Identification of a time series is the process of finding integer, usually very small (e.g., 0, 1 or 2), values of p, d and q that model the patterns in the data. When the value is 0, the element is not needed in the model. The middle element, d is investigated before p and q. The goal is to determine if the process is stationary and if not, to make it stationary before determining the values of p and q. Recall that a stationary process has a constant mean and variance over the time period of the study.

In the simplest time series, an observation at a time period simply reflects the random shock at that time period , a_t, that is:

$$Y_t = a_t$$

The random shocks are independent with constant mean and variance and so are the observations. If there is trend in data , however, the score also reflects that trend as represented by slope of the process. In this slightly more complex model, the observation at the current time, Y_t, depends on the value of the previous observation, the slope and the random shock at the current time period:

$$Y_t = \theta_0(Y_{t-1}) + a_t$$

1.1 Auto-regressive components

The auto-regressive components represent the memory of the process for preceding observations. The value of p is the number of auto-regressive components in an ARIMA (p, d, q) model. The value of p is 0 if there is no relation between adjacent observations. When the value p is 1, there is a relationship between the observations at lag 1 and correlation coefficient ϕ_1 is the magnitude of the relationship. When the value of p is 2, there is a relationship between the observations at lag 2 and the correlation coefficient ϕ_2 is the magnitude of the relationship. Thus p indicates the lag in the dependence of observations which is needed to build the relationship.

For example, a model with p=2, ARIMA (2, 0, 0), is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t$$

1.2 Moving Average Components

The moving average components represent the memory for preceding random shocks. The value q indicates the number of moving average components in an ARIMA (p, d, q). When q is zero, there are no moving average components. When q is 1, there is a relationship between the current score and the random shock at lag 1 and correlation coefficient θ_1 represents the magnitude of the relationship. When q is 2, there is a relationship between the current score and the random shock at lag 2 and the correlation coefficient θ_2 represents the magnitude of the relationship.

Thus, an ARIMA (0, 0, 2) model is

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

1.3 Mixed Models

A series has both auto-regressive and moving average components so that both types of correlations are required to model the patterns. If both elements are present only at lag 1, the equation is:

$$Y_t = \phi_1 Y_{t-1} - \theta_1 a_{t-1} + a_t$$

In the present work, this is the model that we have found to be a suitable one.

2. ACFs and PACFs

Models are identified through patterns in their ACFs (autocorrelation functions) and PACFs (partial autocorrelation functions). Both autocorrelations and partial autocorrelations are computed for sequential lags in the series. The first lag has an autocorrelation between Y_{t-1} and Y_t, the second lag has both an autocorrelation and partial autocorrelation between Y_{t-2} and Y_t, and so on. ACFs and PACFs are the functions across all the lags.

The equation for autocorrelation is that the mean of the response Y-series is subtracted from each Y_t and from each Y_{t-k}, and the denominator is the variance of the whole series (more correctly, it is the mean sum of squares of y-series).

$$r_k = \frac{\dfrac{1}{N-K}\sum_{t=1}^{N-k}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\dfrac{1}{N-1}\sum_{t=1}^{N}(Y_t - \bar{Y})^2}$$

where N is the number of observations in the whole series, k is the lag. $\bar{Y}$ is the mean of the whole series.

The standard error of an autocorrelation is based on the squared autocorrelations from all previous lags. At lag 1, there are no previous autocorrelations, so r_0^2 is set to 0.

$$SE_{r_k} = \sqrt{\frac{1 + 2k - \sum_{l=0}^{k-1} r_l^2}{N}}$$

If an autocorrelation at some lag is significantly different from zero, the correlation is included in the ARIMA model. Similarly, if a partial autocorrelation at some lag is significantly different from zero, it too, is included in the ARIMA model. The significance of full and partial autocorrelations is assessed using their standard errors.

3. Estimating Model Parameters

Estimating the values of parameters in models consists of estimating the parameter(s) from an auto-regressive model or the parameters from a moving average model .as indicated by [D+80], the following rules apply:

1. Parameters must differ significantly from zero and all significant parameters must be included in the model.

2. Because they are correlations, all auto-regressive parameters must be between -1 and 1.If there are two such parameters (p=2) they must also meet the following requirements:

$$\phi_1 + \phi_2 < 1$$
$$\phi_1 - \phi_2 < 1$$

These are called the *bounds of stationarity* for the auto-regressive parameter(s).

3. Because they are also correlations, all moving average parameters, must also be between -1 and 1.If there are two such parameters (q=2) they must also meet the following requirements:

$$\theta_1 + \theta_2 < 1$$
$$\theta_1 - \theta_2 < 1$$

These are called the *bounds of invertibility* for the moving average parameter (s).

Complex and iterative likelihood procedures are used to estimate these parameters. The equation for θ_1 is:

$$\theta_1 = \frac{-\mathrm{cov}(a_t a_{t-1})}{\sigma_N^2}$$

4. Model Diagnostics

How well does the model fit the data? Are the values of observations predicted from the model close to actual ones?

If the model is good, the residuals (differences between actual and predicted values) of the model are a series of random errors. These residuals form a set of observations that are examined the same way as any time series.

Ljung –Box (Q) statistic for diagnostic Checking

The Ljung-Box Q statistic or Q(r) statistic can be employed to check independence instead of visual inspection of the sample autocorrelations. A test of the hypothesis can be done for the model adequacy by choosing a level of significance and then comparing the value of calculated x^2 with the x^2-table at (k-m) degree of freedom, k be the maximum considered lag and m is the no. of parameters in the model. The Q(r) Statistic is calculated by the following equation [LB78]:

$$Q(r) = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}r_k^2$$

where n is the no. of observation in the data series.

5. A prediction model: Structure of raga Bhupali

We define a raga, the nucleus of Indian Classical music, as a melodic structure with fixed notes and a set of rules characterizing a certain mood conveyed by performance [C+09a]. Here we take a sequence of raga Bhupali notes taken from a standard text [Dut06]. Here Y_t values are pitch of the raga notes musical notes which are a function of time (the instance t of realization of musical notes, here the serial numbers 1, 2, 3….157). See Appendix for more details. In Statistics, stochastic processes are the families of random variables which are functions of time measure where time is taken as a realization of the response (here musical notes) and musical data from an analytical perspective can be realized as the outcome of a stochastic process in which the true model is unknown even to the performer although his/her decision is deterministic. Here we are only analyzing the structure of the raga.

6. Results and Discussions

The original series was found non-stationary so a different transformation is required. We try to fit a model which gives minimum error (by error we mean the difference between the predicted value from the model and the actual ones). We checked the residuals from the model. To do that, we did another Ljung-Box (L-B test) test which indicates that residuals. A model is good if the residuals were independent over the various lags for all the models. In this test, the statistics is 36.934 which is larger than the value for the statistical test for raga Malkauns which is 19.800.There is also a large difference between the significance between Malkauns and Bhupali, which is 0.285 and 0.003 respectively. We have 261 instance values for raga Malkauns where as Bhupali has only 157.The results were obtained using SPSS statistical package version 17.0.

Table 1. Fitted ARIMA Model

Note	Model	Fit Statistics			Model Parameters			Model Chosen
		R-squared	RMSE	MAPE	Constant	Difference	Lag	
Bhupali	ARIMA (1,1,0)	.646	2.981	50.569	.019	1	AR(1) -.232	ARIMA (0,1,1)
	ARIMA (0,1,1)	.644	2.992	50.548	.018	1	MA(1) .206	
	ARIMA (1,1,1)	.648	2.983	50.585	.019	1	AR(1) -.490 MA(1) -.271	

Table-2 Results of L-B Test

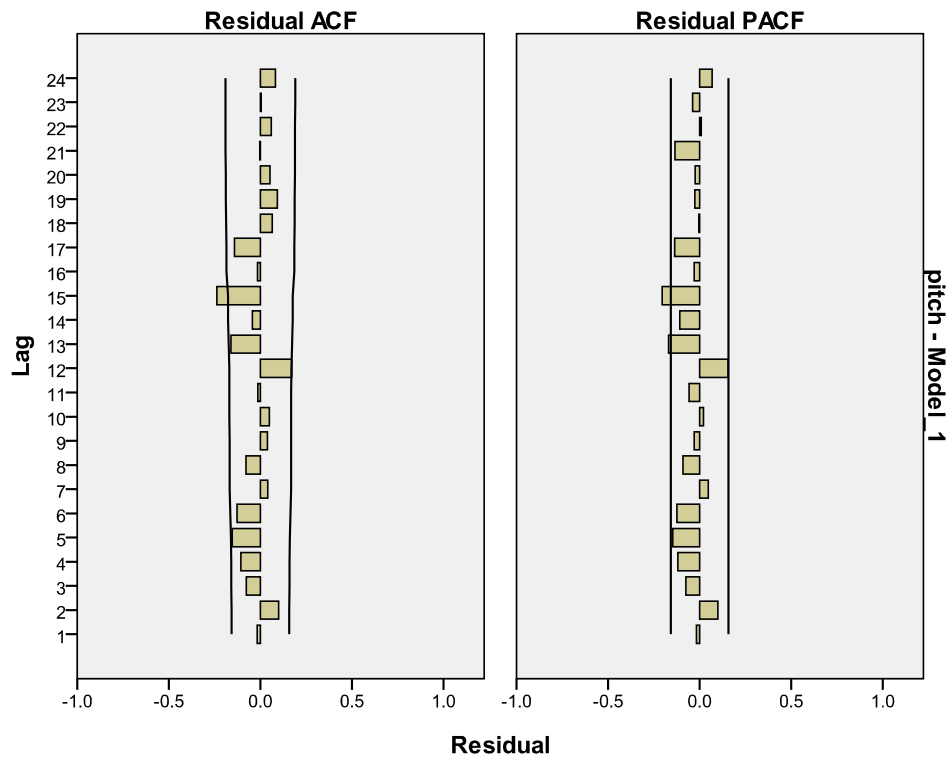| Note	Model	L-B Test		
		Statistics	DF	Sig
Bhupali	ARIMA (1,1,0)	34.080	17	.008
	ARIMA (0,1,1)	36.934	17	.003
	ARIMA (1,1,1)	30.233	16	.017

Fig.-1 Residual ACF and PACF plot

From the above mentioned results it is clear that ARIMA (0, 1, 1) will be the most suitable model. It is clearly revealed from table-1 that R2, MAPE are lower in ARIMA (0, 1, 1) compare to the rest of the models.Fig.1 shows the residual plot for ACF and PACF. The plot exhibits the differences between the actual and predicted values of the model. Using this ARIMA (0, 1, 1) model, we obtain fig.2 which shows the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted series is very close to the observed series.
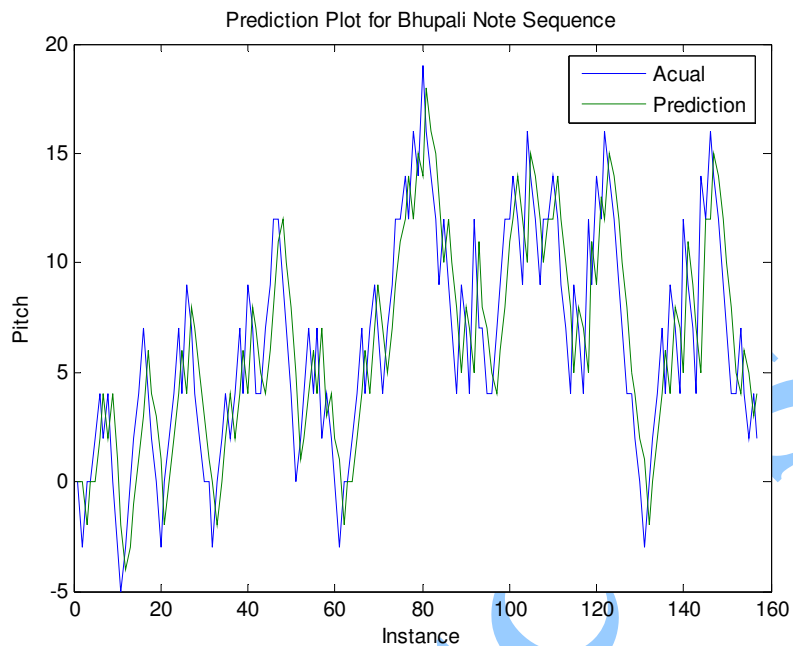
Fig.-2 Actual versus predicted through ARIMA (0,1,1)

Conclusion

Our result shows that ARIMA (0, 1, 1) model is best capturing the Bhupali raga sequence. The lag of unity in the ARIMA model confirms that a first order Markov chain may be used to simulate a sequence of notes of this raga as in our previous work on raga Malkauns [C+09b, CS09].

Appendix

Table 3 gives the numbers representing pitches in different octaves which will be useful in understanding the note progression of the raga in question.

Table 3: Numbers representing pitch of notes [ANO06]

```
 C  Db   D Eb   E  F F#   G Ab   A Bb  B
 S   r   R  g   G  M  m   P  d   D  n  N      (lower octave)
-12 -11 -10 -9 -8 -7 -6  -5 -4  -3 -2 -1

 S   r   R  g   G  M  m   P  d   D  n  N      (middle octave)
 0   1   2  3   4  5  6   7  8   9 10 11

 S   r   R  g   G  M  m   P  d   D  n  N      (higher octave)
12  13  14 15  16 17 18  19 20  21 22 23
```

Abbreviations: The letters S, R, G, M, P, D and N stand for Sa, *Sudh* Re, *Sudh* Ga, *Sudh* Ma, Pa, *Sudh* Dha and *Sudh* Ni respectively. The letters r, g, m, d, n represent *Komal* Re, *Komal* Ga, *Tibra* Ma, *Komal* Dha and *Komal* Ni respectively. Normal type indicates the note belongs to middle octave; italics implies that the note belongs to the octave just lower than the middle octave while a bold type indicates it belongs to the octave just higher than the middle octave. Sa, the tonic in Indian music, is taken at C. Corresponding Western notation is also provided. The terms "*Sudh*", "*Komal*" and "*Tibra*" imply, respectively, natural, flat and sharp.

We close this section giving some general feature of raga Bhupali:-

Raga: Bhupali

   Thaat (a specific way of grouping ragas according to scale): Kalyan
   Aroh (ascent): S R G P, D, **S**
   Awaroh (descent): **S**, D P, G, R, S
   Jati: *Aurabh-Aurabh* (5 distinct notes allowed in ascent and 5 in descent)
   Vadi Swar (most important note): G
   Samvadi Swar (second most important note): D

Anga: *Poorvanga pradhan* (first half more important)
Prakriti (nature): Restful
Pakad (catch): G, R, S, *D*, S R G, P G, D P G, R, S
Speciality: Meend (glide) from S to *D* (or **S** to D) and from P to G
Nyas Swar (Stay notes): G, P and D
Time of rendition: First phase of night (6 PM to 9 PM)

References

[ANO06]   K. Adiloglu, T. Noll, K. Obermayer – *A Paradigmatic Approach to Extract the melodic Structure of a Musical Piece*, Journal of New Music Research, 2006, Vol. 35, no. 3, 221-236.

[BM99]   J. Beran, G. Mazzola - *Analyzing Musical Structure and Performance –A Statistical Approach*, Statistical Science, 1999, vol. 14, no. 1, p. 47-79.

[BJ76]   G.E.P. Box, G.M. Jenkins - *Time Series Analysis, Forecasting and Control*, San Francisco, Holden Day, 1976

[C+09a]   S. Chakraborty, K. Krishnapriya, Loveleen, S. Chauhan, S.S. Solanki - *Analyzing the melodic structure of a North Indian Raga: A Statistical approach*, Electronic Musicological Review, Vol.XII, 2009.

[C+09b]   S. Chakraborty, M. Kumari, S.S. Solanki, S. Chatterjee - *On What Probability can and cannot do: A case study in Raga Malkauns*, Journal of Acoustical Society of India, Vol. 36, No. 4, 2009, p. 176-180

[CS09]   S. Chakraborty, R. Shukla - *Raga Malkauns Revisited with Special Emphasis on Modeling*, Ninad, Journal of ITC Sangeet Research Academy, Vol. 23, Dec 2009

[Dut06]   D. Dutta - *Sangeet Tattwa (Pratham Khanda)*, Brati prakashani, 5th ed, 2006 (Bengali)

[D+80]   D. McDowall, R. McCleary, E.E. Meidinger, R. Hay Jr. - *Interrupted time series analysis*. Thousand Oaks, CA: Sage Publications, 1980

[LB78]   G.M. Ljung, G.E.P. Box - *On a Measure of Lack of Fit in Time Series Models*, Biometrika, Vol.- 65, 297-303.