International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 1 ǁ January 2024 ǁ PP. 286-293

Exploring finite mixtures of t-distribution to model the white noise in AR(p) process

Anuj Nain
Department of Statistics, Central University of Rajasthan
Bandar Seendri, Ajmer, Rajasthan

Abstract

The idea of using mixture distribution for Innovations in time series models has come into the picture in recent years as this approach allows flexible modeling in cases where the observations are multimodal or clustered. In this regard, the present work explores the parameter estimation of AR(p) models in time series where the innovations are driven by a finite mixture of t-distribution. To support the idea, a simulation study is conducted for different cases of the model specifying a number of components in mixture distribution. Further, to show the application, an empirical analysis is done on a real-time data set that justifies the proposed theory.

Keywords: *Mixture distribution, innovations, AR(p) model, White Noise, t-distribution*

Date of Submission: 15-01-2024 Date of acceptance: 28-01-2024

I. Introduction

Many times in real-life situations the random variables are not generated from a single but from a mixture of several distributions. The mixture distribution is a weighted sum of K distributions where the weights sum up to one. Each weight corresponds to the proportion or contribution of the corresponding component density. Each density in the mixture is characterized by an unknown parameter (or a parameter vector). Mixture distributions are well used for capturing humps or multimodality in the data that provides relevant descriptions about the underlying clusters.

In time series modeling, we often encounter scenarios where the error (or innovation) terms arise from multiple subpopulations or clusters. Each cluster has its own weight, representing its contribution as a proportion to the overall error distribution. Analyzing such time series data visually reveals certain typical patterns.If the time series contains observations from the set of real numbers (R), meaning it includes both negative and positive values, the plot may exhibit humps or multimodal characteristics. On the other hand, if the series consists of non-negative values, the time series plot can still show multimodal behavior, or it may contain observations that are right-skewed and have heavy tails. In both situations, it is possible to model the innovations using a mixture distribution. By using a mixture distribution, we can capture the complexity and heterogeneity of the error terms in the time series. This allows the model to flexibly represent different subpopulations or clusters, each contributing its own weight to the overall distribution of errors. The mixture distribution enables us to better fit the data and accommodate various patterns that might not be adequately captured by a single simple error distribution. In this regard Maleki and Nematollahi (2017) employed a Gaussian mixture as the White Noise process in the autoregressive model. Their work was on Gaussian time series, meaning that the data followed a Gaussian (normal) distribution and the observations were real numbers (R), indicating that the data contained both negative and positive values.

Autoregressive models driven by a mixture of Gaussian Innovations might encounter cases where the tail covered by Gaussian innovations isn't heavy enough to support the actual data; this motivated us to introduce a finite mixture of t- Distribution as a White Noise in AR(p) model. Tiku et al. (2000) obtained parameter estimators for autoregressive models with non-normally distributed innovations modeled by Student's t-distribution with known degrees of freedom, utilizing the Modified Maximum Likelihood (MML) method. Tarami and Pourahmadi (2003) explored autoregressive models featuring dependent yet uncorrelated innovations distributed according to the t-distribution with known degrees of freedom, deriving the exact maximum likelihood equation for the model parameters.

Christmas and Everson (2011), employing the variational Bayesian method, incorporated the t-distribution into autoregressive models and proposed certain approximations to the posterior distribution of the model parameters. Regarding the t-distribution, Dempster et al. (1977) demonstrated the applicability of the EM algorithm for obtaining maximum likelihood estimates with completeunivariate data and fixed degrees of freedom.

The Expectation Maximization (EM) method has garnered significant interest following the influential work of Dempster et.al (1977). Particularly, when dealing with models characterized by challenging structures in maximum likelihood estimation of parameters, a missing data approach has been widely adopted. In the realm of parameter estimation procedures, Chen et.al (2014) introduced a method that maximizes a class of penalized likelihood functions. This approach explicitly incorporates modeling of missing data probabilities through the utilization of the EM algorithm. The EM algorithm stands out as a versatile method for iteratively computing maximum likelihood estimates, particularly when observations can be perceived as incomplete data. Its remarkable effectiveness is attributed to the simplicity of the associated theory and its extensive applicability across various domains. For a comprehensive understanding of the theory behind the EM algorithm and its extensions, refer to McLachlan and Krishnan (2007). The EM algorithm is commonly employed in maximum likelihood estimation due to its superior stability in convergence compared to the direct application of the Newton-Raphson algorithm, quasi-Newton, or conjugate gradient approaches, as outlined by Press et al. (2007). The hierarchy of the paper is as follows: In section 1 we present the introduction, here we give a brief literature review. In section 2, a brief introduction of mixture distribution has been provided (section 2.1), the model under studt has been presented in section 2.2 and the parameter estimation has been discussed in section 2.3. Next section 2.3 presents the simulation study followed by empirical analysis insection 4. Section 5 is about conclusion.

II. Modeling and Estimation

2.1 The Mixture distribution:

Let X_1 X_2 . . . $X_n$be an independent and identically distributed random sample from a K-component finite mixture of probability distributions. This mixture distribution is represented as [see Chandra (1977), Dempster et. al (1977)]

$$f(x;\,\Theta) = \;;\sum_{k=1}^{K} \pi_k f_k(x;\ \theta_K)\ , \qquad (1)$$

$$\text{subject to } \sum_{k=1}^{K} \pi_k = 1$$

where $\boldsymbol{\Theta} = (\boldsymbol{\pi}', \boldsymbol{\theta}')$ = (π_1 , π_2 ,. . . , π_{k-1} , θ_1 , θ_2 , . . . θ_k) is the vector of unknown parameters and $0 < \pi_i \leq 1$. These K distributions may or may not be from the same family. In this paper, we assume that for the mixture density given in (1) the component densities $f_k(.)$ are from the same family (which is t- distributed) . The associated pdf is given by

$$f_k(x;\ \theta_k) = \frac{1}{\sqrt{\beta\left(\frac{1}{2},\frac{a_k}{2}\right)}}\left(1+\frac{\left(\frac{x-\mu_k}{\sigma_k}\right)^2}{a_k}\right)^{-\frac{a_{k+1}}{2}} \qquad , -\infty < x < \infty \qquad (2)$$

where $\beta(a,b) = Gamma(a)Gamma(b)/Gamma(b+b)$.

$\theta_k = a_k, \mu_k$ *and* σ_k are the parameters of the distribution. Next we discuss the model under study

2.2 The Model under study

The pth order Autoregressive model is given by

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-1} + \mathcal{E}_t \qquad \text{t= 1, 2, 3 . . . n} \qquad (3)$$

Where $|\phi_i|$ < 1 , $\mathcal{E}_t$ are iid from the density $f\,(\,.\,)$ such that

$$f(\mathcal{E}_t) = \sum_{k=1}^{K} \pi_k f_k(\mathcal{E}_t;\ \theta_K\ ) \ ,$$

$$\text{subject to } \sum_{k=1}^{K} \pi_k = 1 \qquad (4)$$

Where $f_k\left(\mathcal{E}_t;\ \theta_K\ \right)$follows the density defined in (2)

We use $\mathcal{E}_t = \sum_{i=1}^{p} \phi_i y_{t-1} - y_t$, as this representation alows flexible parameter estimation.

2.3 Parameter estimation

The white noise process described for the model in section 2.2 assumes that clusters are the source of the Innovations, enabling us to represent them using a mixture density. Here we assume that the entire set of innovations contains observed terms and Latent variables. By employing the EM Algorithm, we can attain maximum likelihood estimates when handling this mixture density.We call the density given in (1) as the incomplete data density. with the following likelihood function and log likelihood functions respectively.

$$\boldsymbol{l}\,(\boldsymbol{\Theta};\,\boldsymbol{X}) = \prod_{i=1}^{n} f\,(x_i;\ \boldsymbol{\Theta}) \ = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k\left(\mathcal{E}_\mathrm{i};\ \theta_k\ \right) \qquad (5)$$

$$\boldsymbol{L}\,(\boldsymbol{\Theta};\,\boldsymbol{X}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} log f\ (\mathcal{E}_i;\ \boldsymbol{\Theta}) \qquad (6)$$

We call the likelihood given in (5) as incomplete data log-likelihood whereas for implementing EM Algorithm, complete data log likelihood is required. So, at first we assume that the complete data-set consists of **D = (E, Z**

) but that only **E** is observed , whereas **Z** = (z_{ik} Є { 0 , 1}, k = 1 , 2 . . . K , i = 1,2, . . .n) is the set of latent variables. The complete-data likelihood is then denoted by

$$\boldsymbol{l\,(\Theta;\,E,\,Z)} \text{ as } \boldsymbol{l(\Theta;\,E\,,\,Z)} = \prod_{i=1}^{n}\sum_{k=1}^{K}(\pi_k f_k(\mathcal{E}_i;\ \theta_k\,))^{z_{ik}} \qquad (7)$$

where z_{ik} = 1 if the observation $\mathcal{E}_i$ comes from kth component density, otherwise z_{ik} = 0. Taking logarithm of (7) gives us complete data log-likelihood.

$$\boldsymbol{L\,(\Theta;\,X,\,Z)} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}[\log(\pi_k) + \log(f_k(\mathcal{E}_i;\ \theta_k\,))] \qquad (8)$$

where $\boldsymbol{\Theta}$ is the unknown parameter vector for which we wish to find the MLE. The term EM stands for Expectation – Maximization, which are also the two steps of the algorithm.

The E-step of the EM algorithm computes the expected value of $\boldsymbol{L\,(\Theta;\,E,\,Z)}$given the observed data, $\boldsymbol{E}$, and the current parameter estimate, $\boldsymbol{\Theta}_{old}$ say. In particular, we define w_{ik} as

$$w_{ik} = \mathrm{E}\,[z_{ik} \mid \mathcal{E}_i,\, \boldsymbol{\Theta}_{old}] = \frac{\pi_k f_k(\mathcal{E}_i;\,\theta_k\,)}{\sum_{k=1}^{K}\pi_k f_k(\mathcal{E}_i;\,\theta_k\,)} \qquad (9)$$

We replace z_{ik} by w_{ik}. in (9). The result is called as **Q** function , given by

$$\mathbf{Q}\,(\boldsymbol{\Theta} \mid \boldsymbol{\Theta_{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}[\log(\pi_k) + \log(f_k(\mathcal{E}_i;\ \theta_k\,))] \qquad (10)$$

In the M-step the **Q** function is maximized to obtain new estimates of $\boldsymbol{\Theta}$(θ and π) . The two steps are repeated as necessary until the sequence of new estimates of θ and π converges. This optimization problem can be solved separately for π and for θ. The estimates of π are updated in the sth iteration by

$$\pi_k^{(s)} = \frac{\sum_{i=1}^{n} w_{ik}^{(s)}}{n}$$

III. Simulation Study

Simulation in statistical analysis involves generating artificial data using the model's assumptions and specifications. This artificial data undergoes the same statistical analysis as the actual data, allowing us to grasp the model's behavior under diverse dimensions, parameters, or conditions. To assess our proposed methodology, an extensive simulation study has been conducted , showcasing how effectively the model captures the clusters in the Innovation.

According to the model described in section 2.2, "K" denotes the the number of t-distributed components in the mixture density. The total number of parameters to be estimated in the Modal is (3K + 1). By selecting different values of K, various models can be crafted. Two cases have been simulated in this paper:

Model 1: AR (1) model with 2-K t-distribution as White Noise(K = 2)

Model 2: AR (1) model with 3-K t-distribution as White Noise (K = 3)

For each case, the samples have been simulated 20000 times with respective sample sizes 50 , 100 , 150 , 200 , 250 , 300, 350 , 400 , 450 , 500. Plots of Absolute Bias and Mean Squared Error have ben constructed to verify the estimates. The simulation results for the two models have been discussed in the subsections below.

3.1 Computations for Model 1

The White Noise process in model 1 is governed by two t-distributed components, so that on specifying K=2 , the model in (3) would be represented as

$$y_t = \phi y_{t-1} + \mathcal{E}_t \qquad t = 0\,,1,\,2\ \ldots\ldots n$$

Where $|\rho| < 1$ and $\mathcal{E}_t$ are iid from the density $f\,(\,.\,)$ such that

$f\,(\mathcal{E}_t) = \sum_{k=1}^{2}\pi_k f_k(\mathcal{E}_t;\ \theta_K\,)$,

subject to $\sum_{k=1}^{2}\pi_k = 1$

$$where\ \ f_k(x;\ \theta_k) = \frac{1}{\sqrt{\beta\left(\frac{1}{2},\frac{a_k}{2}\right)}}\left(1 + \frac{\left(\frac{x-\mu_k}{\sigma_k}\right)^2}{a_k}\right)^{-\frac{a_{k+1}}{2}} \qquad ,-\infty < x < \infty$$

With $\theta_k = \{a_k, \mu_k\,, \sigma_k\}$, $a_k, \sigma_k > 0$, k = 1, 2

There are nine parameters in the model. Following values have been used for simulation purpose. $\phi =\ 0.489$, $\pi_1 =\ 0.21\,, a_1 =\ 1.412\,,\ \mu_1 =\ 2.956 \quad \pi_2 =\ 0.79\,,\ a_2 =\ 3.042,\ \mu_2 = 0.089\ \sigma_1 = 2\ \sigma_2 = 3.$ For each sample size the Absolute bias and Mean Squarred Error (MSE) of the parameter estimates have been obtained and their plots against the sample size have been constructed (see figue 1 and figure 2)

Figure 1: Plot of Absolute bias for parameter estimates of model 1

It can be seen for each parameter estimate that with the increase in sample size the Bias of the parameter estimate decreases to zero. Similarly with the increase in the sample size MSE of the parameter estimates also decreases.

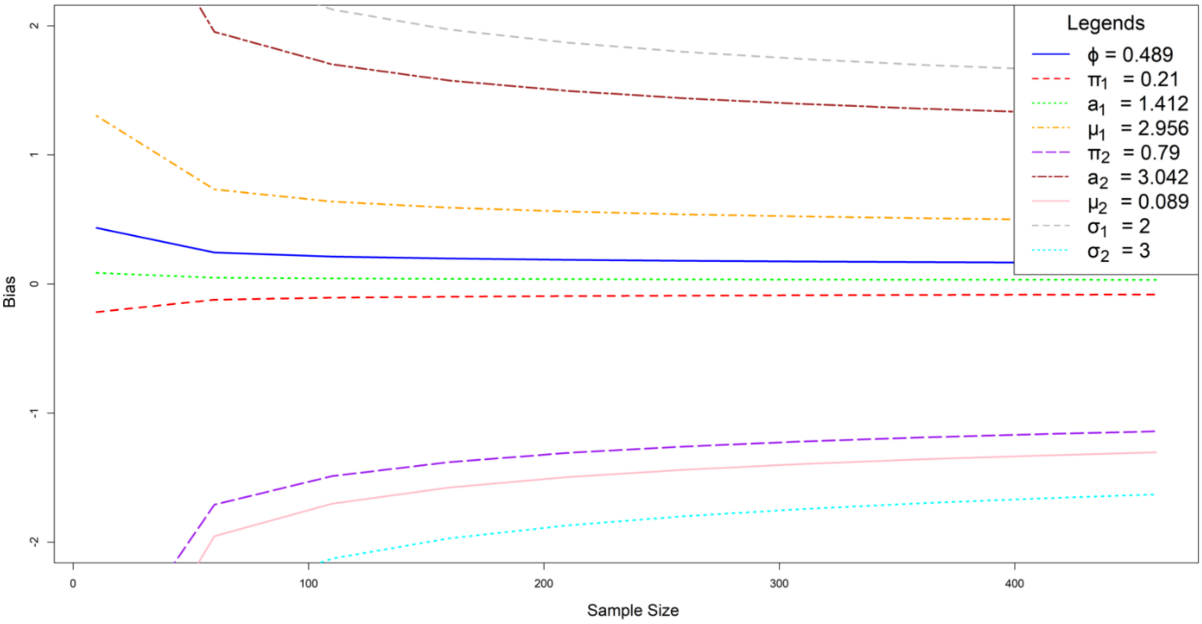


Figure 1: Plot of BIAS for parameter estimates of model 1

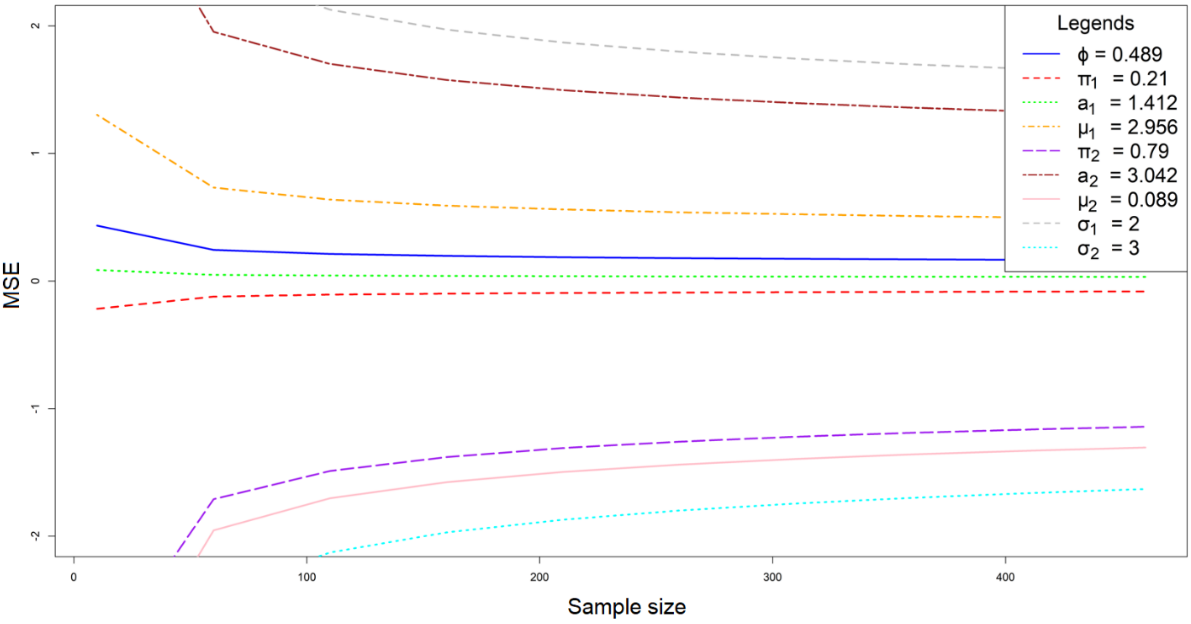


Figure 2: Plot of Mean Square Error (MSE) for parameter estimates of model 1

3.2 Computations for model 2

The White Noise process in model 2 is governed by three t-distributed components, so that on specifying K=3 , the model in (3) would be represented as

$$y_t = \phi y_{t-1} + \mathcal{E}_t \qquad t = 0, 1, 2 \dots\dots n$$

Where $|\rho| < 1$ and $\mathcal{E}_t$ are iid from the density $f(\,.\,)$ such that

$f(\mathcal{E}_t) = \sum_{k=1}^{3} \pi_k f_k(\mathcal{E}_t;\ \theta_K)$,

subject to $\sum_{k=1}^{3} \pi_k = 1$

$$where\ \ f_k(x;\ \theta_k) = \frac{1}{\sqrt{\beta\left(\frac{1}{2},\frac{a_k}{2}\right)}}\left(1 + \frac{\left(\frac{x-\mu_k}{\sigma_k}\right)^2}{a_k}\right)^{-\frac{a_{k+1}}{2}} \qquad , -\infty < x < \infty$$

With $\theta_k = \{a_k, \mu_k, \sigma_k\}$, $a_k, \sigma_k > 0$, k = 1, 2,3

There are thirteen parameters in the model. Following values have been used for simulation purpose. $\phi = 0.350$, $\pi_1 = 0.3$, $a_1 = 1.5$, $\mu_1 = 1.2$, $\pi_2 = 0.4$, $a_2 = 3.5$, $\mu_2 = 3.2$, $\pi_3 = 0.2$, $a_3 = 5$, $\mu_3 = 0.33$ $\sigma_1 = 1$ $\sigma_2 = 2$ $\sigma_3 = 3$. For simulation corresponding to each sample size the Absolute bias and Mean Squared Error (MSE) of the parameter estimates have been obtained and their plots against the sample size have been constructed (see Figure 3 and Figure 4). It can be seen for each parameter estimate that with the increase in sample size the Absolute bias of the parameter estimate decreases to zero. Similarly It can be seen for each parameter estimate that with the increase in sample size the MSE of the parameter estimate decreases.

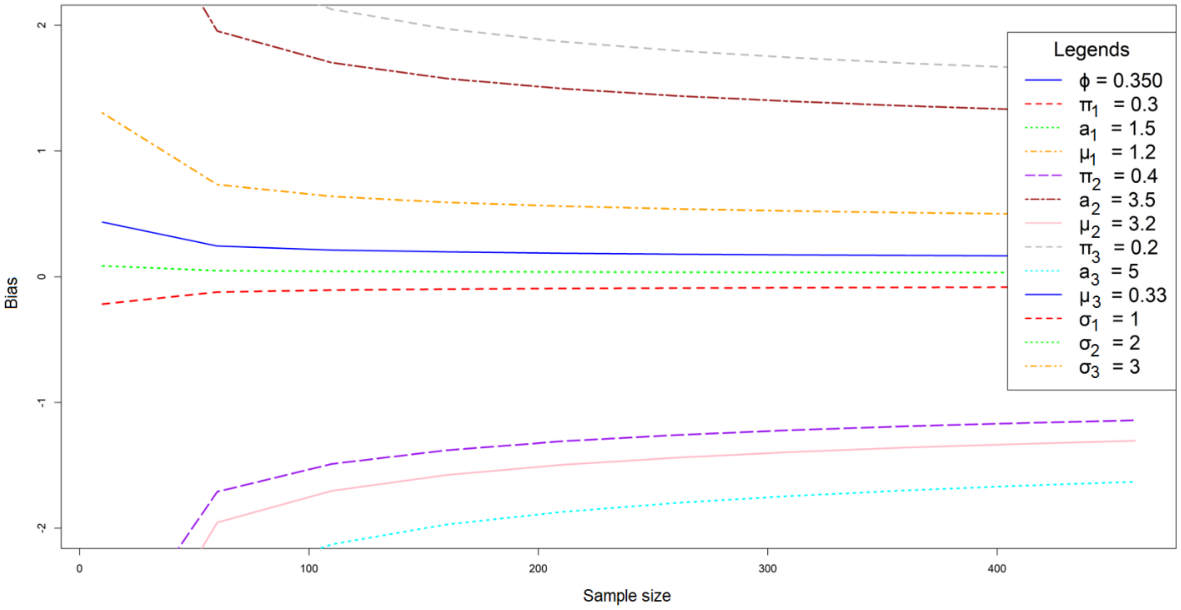


Figure 3: Plot of Absolute bias for parameter estimates of model 2

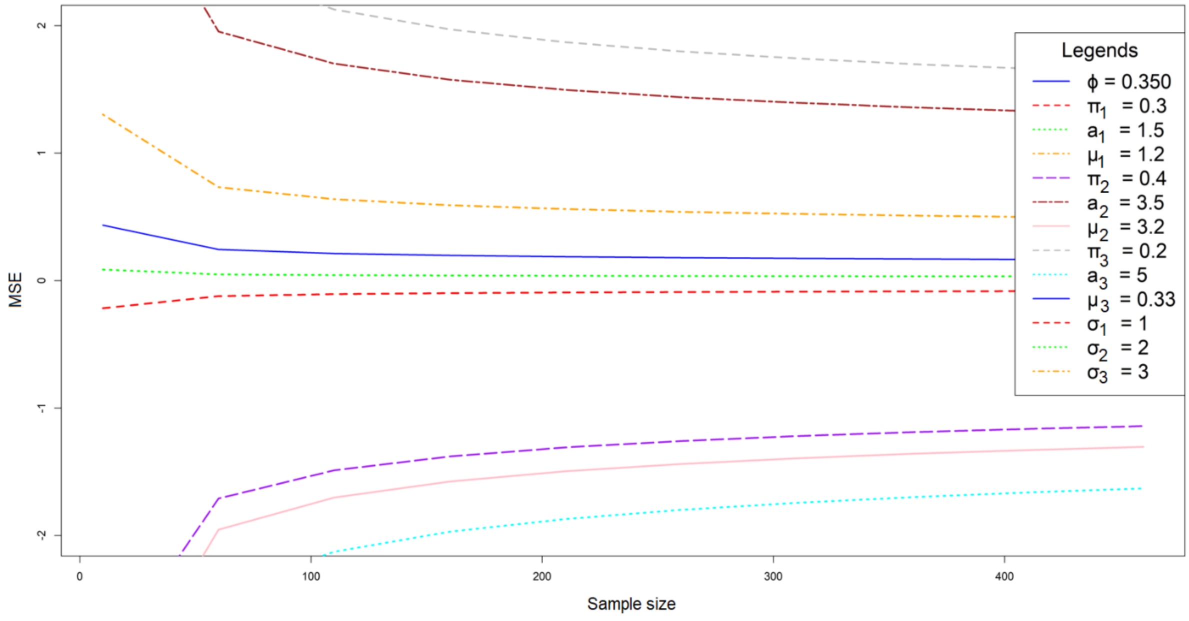


Figure 4: Plot of Mean Square Error (MSE) for parameter estimates of model 2

IV. Empirical analysis

In order to show the applicability of our proposed methodology an empirical analysis on a real-time data has been done in this section. The daily box-office collection of the movie “Jawan” has been considerd for this purpose. The data was available on the website BOX OFFICE INDIA from Sept 9 , 2022 to Oct 28, 2022. During this timeframe, the website continuously recorded the data. However, it's important to note that once the movie's viewership in theaters declined, the corresponding data was removed from the website. If needed, this removed data can be obtained through a special request either from BOX OFFICE INDIA or Dharma Productions. The time series data has been represented using the notation “ y_t“ . Figure 5 shows the time series plot of the data. Figure 6 shows the PACF plot along with ADF test. The ADF test shows that the series is stationary and the PACF plot shows that the series is AR(1).

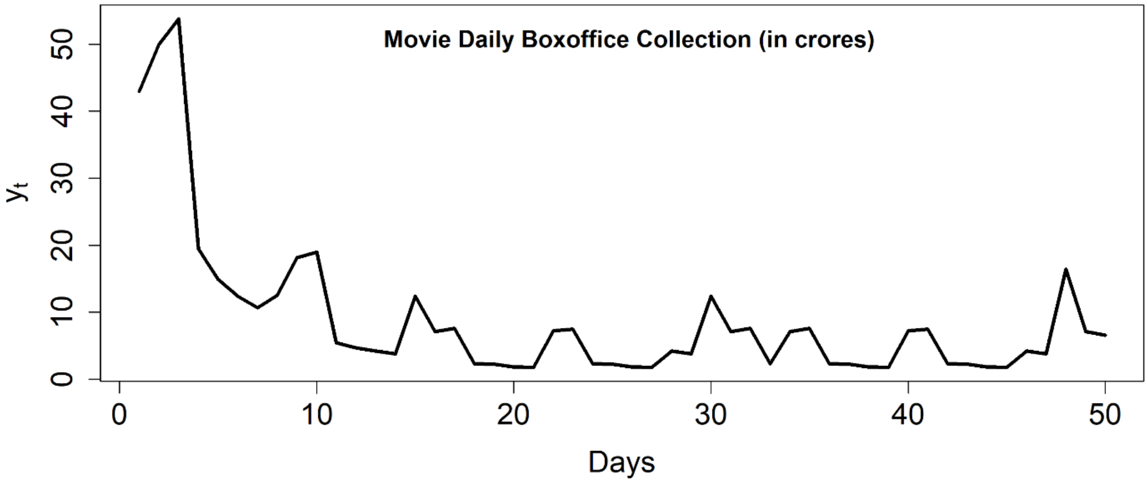
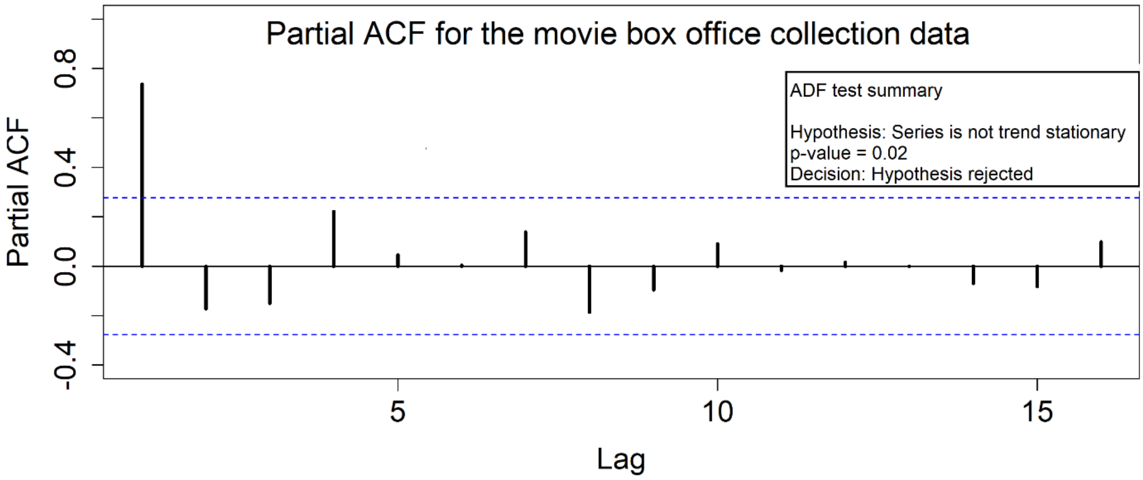


Figure 5 Time series plot for the series y_t



Figur 6 : PACF plot for the series y_t

We consider fitting models for different values of “ K “ , i.e the number of t-distributed components in Innovation. For the model selection criteria we use AIC and BIC. As a criteria used by many researchers, we stop increasing the number of components at the stage when there is no improvement in BIC, for details on this one may refer [Miljkovic and Grün (2016)].

Table 1: AR (1) Models for different values of K

K (Number of Components)	AIC	BIC
1	191.652	503.388
2	141.308	418.692
3	180.468	349.588

Table 1 displays the performance of various AR(1) models with different values of K. It can be observed that using K = 2 or K = 3 for the model, shows improved fit compared to K = 1. This indicates that employing a mixture density for innovations is a better choice than using a single density model. Particulary, the model with two t-distributed components (K=2) provides the best fit. Figure 7 provides a clear depiction of how effectively this model, i.e fitted AR(1) model with 2-K t-distributed white noise aligns with the original series. The parameter estimates of this model, along with the density of Innovations, are illustrated in Figure 8. The presence of two humps in Figure 8 justifies the choice of two t-distributed components, effectively capturing the clusters in the density of Innovations. The interpretation of the parameter estimates of the density is as follows: When an error or innovation occurs, it is governed by a mixture of two t-distributions. The first component contributes 25.3% of the weight and has a shape parameter of 3.023 and a scale parameter of 3.129. The second component contributes 74.7% of the weight and has a shape parameter of 2.234 and a scale parameter of 0.216. This mixture of two t-distributions effectively characterizes the density of innovations, with each component capturing a different aspect of the underlying data pattern.

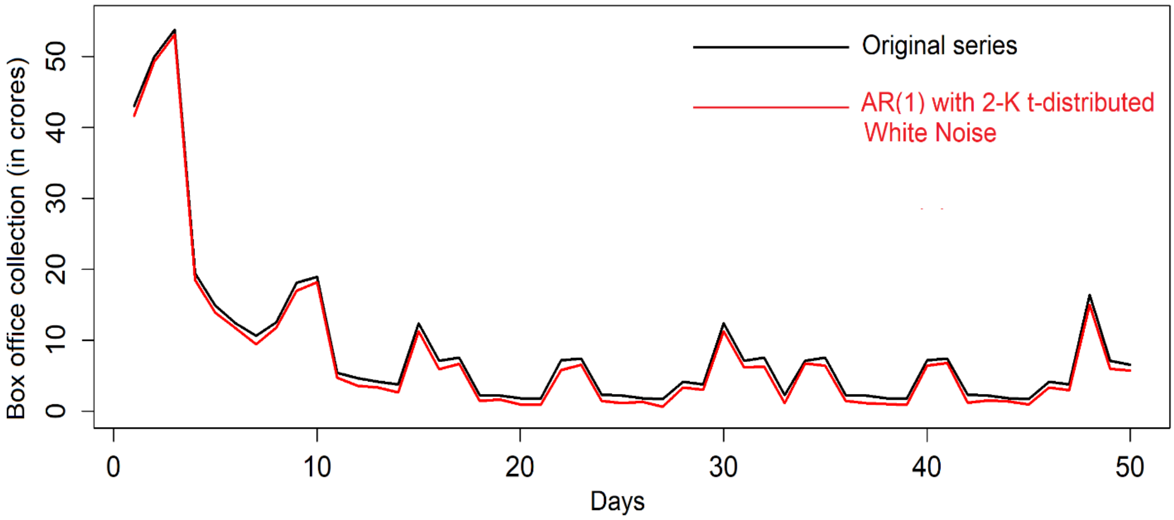


Figure 7 : Plot of original series and the fitted series

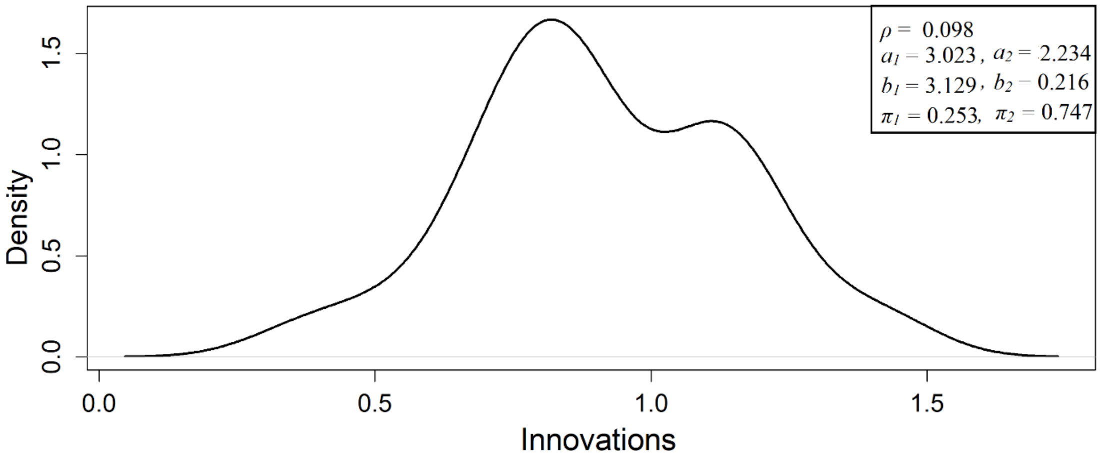


Figure Conclusion

From our perspective, we are the pioneers in conducting this particular study, as no similar research has been previously undertaken. Our work offers a distinctive and unparalleled contribution to the field by providing a comprehensive understanding of models in describing non-Gaussian time series, particularly in capturing the multimodal or clustering nature of the data.

Based on the findings from the simulation study, empirical analysis, and realization study, our conclusion suggests that for non-Gaussian time series data governed by an AR(1) process, using a mixture of t-distributed distributions as the White noise distribution is preferable over choosing a single density or specifically the exponential density. The proposed mixture of t-distributions offers distinct advantages in capturing the clustering or multimodal nature of the time series data, which the traditional single density or exponential density models may struggle to represent accurately. This improved capability makes the proposed approach a more suitable and robust choice for modeling complex and diverse time series patterns.

In this study , the focus was solely on thet-distribution for modeling non-Gaussian AR(1) time series data. However, it is essential to recognize that the underlying theory can be extended to include other appropriate distributions as well. Different distributions might be better suited to capture specific characteristics of the data, and exploring alternative families could offer valuable insights into model performance and flexibility.

Furthermore, in our study, we made the assumption of using the same distribution family for each component in the mixture model. While this simplification was useful for our initial investigation, future research could delve into the potential advantages of considering different distribution families for distinct components.

References

[1]. Andel, J., 1988. On AR (1) processes with exponential white noise. Communications in Statistics-Theory and Methods, 17(5), pp.1481-1495.

[2]. Maleki, M. and Nematollahi, A.R., 2017. Autoregressive models with mixture of scale mixtures of Gaussian innovations. Iranian Journal of Science and Technology, Transactions A: Science, 41, pp.1099-1107.

[3]. GAVER D.P. & LEWIS, P.A.W. (1980). First-order autoregressive gamma sequences and point processes. Adv. in Appl. Probab. 12, 726–745

[4]. LAWRANCE, A.J. (1982). The innovation distribution of a gamma distributed autoregressive process. Scand. J. Statist. 9, 234–236.

[5]. DEWALD, L.S. & LEWIS, P.A.W. (1985). A new Laplace second-order autoregressive time series model — NLAR(2). IEEE Trans. Inf. Theory 31, 645–651

[6]. BELL, C.B. & SMITH, E.P. (1986). Inference for non-negative autoregressive schemes. Comm. Statist. Theory Methods 15, 2267–2293

[7]. RAO, P.S. & JOHNSON, D.H. (1988). A first-order AR model for non-Gaussian time series. In Proceedings of IEEE International Conference on ASSP. Vol. 3, pp.1534–1537

[8]. HUTTON, J.L. (1990). Non-negative time series models for dry riverflow. J. Appl. Probab. 27, 171–182

[9]. SIM, C.H. (1993). First-order autoregressive logistic processes. J. Appl. Probab. 30, 467–470

[10]. LYE, J.N. & MARTIN, V.L. (1994). Non-linear time series modelling and distributional flexibility. J. Time Ser. Anal. 15, 65–84.

[11]. BROCKWELL, P.J. & DAVIS, R.A. (1991). Time Series: Theory and Methods, 2nd edn. New York: SpringerVerlag

[12]. [Chandra (1977)] Chandra, S. (1977). On the mixtures of probability distributions. Scandinavian Journal of Statistics, 105-112.

[13]. [Dempster et al. (1977)] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

[14]. https://www.boxofficeindia.com/