CONCEPT NOTE

A Versatile Modification of Gaussian Finite Mixture Models for Clustering in Time Series Analysis

NJAGI HENRY MURIMI

PA301/S/21076/23

Keywords: Versatile Modification of Gaussian Finite Mixture Models; Time Series Clustering; Multiple Quantile Clustering; Statistical Modeling.

Proposed Title: A Versatile Modification of Gaussian Finite Mixture Models for Clustering in Time Series Analysis.

The proposed study falls under the following areas: Probability distributions and Data Analysis.

Description of the Proposed Study

The objective of this study is to develop new statistical modeling techniques, especially in time series data clustering. The main goal is to offer a modified GMM specifically for time series clustering with multiple quantiles (Zhang et al., 2019). The adjustment aims at overcoming flaws arising from a single Gaussian density (Luca & Scrucca, 2015) and providing some flexibility that enables the model to adapt quickly enough toward complicated dynamics coming across time series patterns.

Statement of the Research Question

How does a generalized version of the Gaussian Finite Mixture Model, tailor-made for clustering in time series data improve accuracy and flexibility while addressing one-to-one correspondence assumptions, particularly in terms that we can cluster at multiple quantiles?

1. Introduction and rationale

Statistics is key for the development of models used to comprehend and describe real-life phenomena. The primary goal of statistical modeling is to ascertain the uncertainties that exist in different fields, especially time series analysis. These models are based on probability distributions which enable the researchers to determine and investigate uncertain events quantitatively.

A tremendous amount of energy has been directed toward enhancing statistical modeling, and more recently, the emphasis moved to expanding distribution families by adding parameters (Jose K., 2011). This method also has potential since it leads to more dynamic models that can reflect the complexity of actual data. Consistent with this fresh take, our study seeks to enrich the developing face of statistics modeling by offering a generalized variation of Gaussian Finite Mixture Models for

clustering in time. series data.

Model-based clustering is a prevalent approach in the field of statistical modeling (Baudry et al., 2010), especially regarding multivariate continuous data. Nevertheless, the implicit assumption of a direct relationship between mixture components and clusters does not always apply. This spurs our research in which we seek to develop a flexible modification of Gaussian Finite Mixture Models designed precisely for the difficult conditions that arise from clustering time series data. Since relying solely on a single Gaussian density is inherently limited (A. Azzalini *et al., 2007*), we investigated new approaches.

A promising direction is to generalize our model's clustering at multiple quantiles, rather than strictly limiting the levels. We aim to improve our ability to time series intrinsic patterns, revealing a detailed picture of cluster dynamics. However, care must be exercised in avoiding the traps that follow from traversing quantiles as our suggested clustering method is both robust and interpretable. Our attempt does not stop at improving on the accuracy and adaptability of time series clustering but extends to exploring an uncharted territory- multi-quantile-based clustering. By doing so, we aim to provide a vital contribution to the general area of statistics that focuses on statistical modeling and data analysis.

2. Statement of the Problem

This research is motivated by the perceived limitations of current statistical modeling approaches, especially concerning time series clustering. The widespread one-to-one correspondence assumption between mixture components and clusters as well as a strong Gaussian density use render it problematic to portray the volatility in time series data. Specifically, the gap or niche we wish to address concerns a more customizable alteration of Gaussian Finite Mixture Models that Perform Time Series Data Clustering. This alteration should not only address the drawbacks of a single Gaussian density but also unravel new possibilities such as clustering at various quantiles (Muthama Musau et al., 2022) which will enhance our knowledge of dynamics clusters.

2.1 Objectives

General objective

To develop a versatile Modification of Gaussian Finite Mixture Models for Clustering in Time Series data.

Specific objectives

1. Develop and Implement a Generalized Gaussian Finite Mixture Model (GMM) for Time Series Clustering: An altered GMM algorithm is expected to yield better results in clustering time series data.
2. Evaluate and Quantify the Improvement in Clustering Accuracy: Evaluate and compare the clustering accuracy of the proposed model with standard GMM methods according to such measures as silhouette score, and adjusted Rand index.
3. Introduce and Evaluate Modification as Flexibility: Compare the flexibility of modified GMM by determining how well it adapts to different patterns of time series compared with traditional GMM.
4. Investigate and Measure the Effect of Clustering for Several Quantiles: Analyze how the use of multiple quantiles impacts model performance, assessing measurable benefits and threats.
5. Advancing Statistical Modeling Techniques: Explain how the proposed modification and clustering approach bring elements of a novel contribution by showing empirical improvements in statistical modeling, whereby accuracy increases, and flexibility rises alongside adaptability.

3. Literature Review

The literature review includes various works from distinguished authors in statistical modeling, who focus on specific elements of clustering and distribution models. In his work on Gaussian finite mixture models, Luca and Scrucca (2015) emphasize that the most common assumption is a one-to-one correspondence between clusters and components of mixtures; they argue for a clustering algorithm based on locating areas with high densities. This draws attention to the need for changes that must be introduced into conventional clustering methodologies, coinciding with our goal of improving a modification of Gaussian Finite Mixture Models in terms of time series data clustering.

Musau and Gaetan (2021) present a two-dimensional clustering approach for bivariate time series based on a quantile regression model. However, their approach of clustering at various quantile levels reflects the inherent challenges associated with methods based on average values over entire periods on the issue of overlapping components (Baudry et al., 2010). This is in line with our goal of examining clustering behavior at multiple quantiles over time series data to gain a deeper insight into the underlying distribution patterns.

Katherine Morris et al., (2018) introduce mixtures of contaminated shifted asymmetric Laplace distributions highlighting the necessity to model unbalanced cluster structure with outliers. This corresponds to the stated goal of resolving these limitations within a single Gaussian density and exploring clustering across various quantiles, creating an all-encompassing strategy for effectively handling different forms of data.

The article by Nuttanan Wichitaksorn, S. T. Boris Choy, and Richard Gerlach (2019) suggests a generalized class of skew distributions based on a mixture of normal random variables scaled to have the same variance/covariance structure. This observation is closely aligned with our desire to create a generalized variation of Gaussian Finite Mixture Models that would be suitable for clustering time series data based on the concept of multiple quantiles (Carreira-Perpiñán and Williams 2003).

4. Materials and Methods

4.1 Model Development

Quantitative model development

The proposed research will use both quantitative and qualitative approaches to construct a model, which is the flexible modification of Gaussian Finite Mixture Models (GMMs) for time series clustering. This broader method is selected to conduct a detailed investigation of both quantitative measures regarding performance metrics and qualitative insights into the structures grasped by the changed model .

1. In the model development, we will start with the traditional representation of Gaussian Finite Mixture Models (GMMs).

$$GMM_{original} = \sum_{k=1}^{K} \pi_k \times \mathrm{N}(\mu_k, \Sigma_k) \quad (1)$$

2. The main assumption under this model (1) is that the data can be represented as a combination of several Gaussian distributions. This model will only work as a baseline for comparison with the modified model.

$$GMM_{modified} = \sum_{k=1}^{K} \sum_{q=1}^{Q} \pi_{k,q} \times \mathrm{N}\left(\mu_{k,q}, \Sigma_{k,q}\right) \quad (2)$$

3. The parameter in Equation 2 will be estimated through the Expectation-Maximization (EM) algorithm. Given the model, the log-likelihood of the observed data will be maximized following equations 3 for computing the posterior probabilities, $P\left(Z|X, \theta^{(t)}\right)$ and 4 for maximizing the expected log-likelihood $Q\left(\theta, \theta^{(t)}\right)$.

Expectation (E-Step)

$$Q\left(\theta, \theta^{(t)}\right) = \sum_{Z} P\left(Z|X, \theta^{(t)}\right) \log P(X, Z|\theta) \quad (3)$$

Maximization (M-step)

$$\theta^{(t+1)} = arg\ max_{\theta}\ Q\left(\theta, \theta^{(t)}\right) \quad (4)$$

4. The cluster assignment will be done after parameter estimation, in this, each observation will be assigned to the cluster that maximizes the posterior probability while considering both the quantile and cluster.

Qualitative model development

5. Pattern identification through observing the cluster membership of observation i and the q-th quantile ($C_{i,q}$) and the posterior probability of observation i that belong to the cluster k at q-th quantile ($P(C_k|x_i, q)$)

$$C_{i,q} = arg\ max_k\ P(C_k|x_i, q) \quad (5)$$

6. Assessing the observations and their similarities within each cluster at distinguished quantiles.

Model Evaluation and Validation

7. Train the model using a time series data.
8. Evaluate the model's performance by comparing it with traditional finite GMMs to ensure robustness and generalization; this is crucial for real-world applicability.

4.2 Population, Sample, and Sampling Method

This study population includes time series data sets with different features and structures. Purposive Samples will be collected from time series datasets available in the public domain and considered relevant to the research aims. The purposive sampling technique will make it possible to select datasets with various time series patterns and correlating features that traditionally are difficult for GMMs. Based on the diversity of available time series datasets and to ensure that a variety of patterns are captured, the intended sample size will be calculated.

4.3 Methods of Data Collection

Data collection will involve a multi-step process:

Selection of Time Series Datasets: Purposive sampling will be used to select appropriate time series datasets that are available publicly for the study’s purpose.

Preprocessing of Time Series Data: After selecting the required datasets, they will be preprocessed to deal with missing values, outliers, and standardization for them to become compatible.

Quantitative Model Development: Programming languages like Python and R will be used for the versatile modification of GMMs which involve clustering at varying quantiles to identify varied distribution patterns. For assessment, quantitative measures such as clustering accuracy and log likelihood values will be used.

Qualitative Evaluation: Patterns formulated through clustering at different quantiles will be qualitatively. evaluated to determine interpretability and implication value.

4.4 Data Analysis

The analysis will encompass both quantitative and qualitative assessments:

Quantitative Analysis: The proposed GMM will be compared to standard GMMs through statistical. measures like clustering accuracy, silhouette scores, and log-likelihood values.

Qualitative Analysis: Clustering at several quantiles will reveal patterns that are qualitatively evaluated for meaningful interpretation and useful utility.

5.Ethical Considerations

Participant Anonymity and Confidentiality: All measures will be taken to preserve the anonymity and confidentiality of respondents. Identifiable information including name and contact details will be detached from the acquired data. Participants will be assigned identification numbers to ensure anonymity. The identifiable information will be accessible only to the research team and it will not be merged with anonymized data.

Secure Data Storage: There will be data security through the storage of information in a digital environment with limited access. The analysis shall only be accessible to the researchers engaged with the data. A data encryption technique is likely to be used to keep the information secure from unauthorized access, and backup measures shall also be taken so as not to lose any of its content.

Risk Mitigation: The researcher will perform a risk assessment to address possible ethical risks arising from the study.

REFERENCES

Azzalini, A., & Torelli, N. (2007). Clustering via nonparametric density estimation. Statistics and Computing, 17, 71-80.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. Journal of computational and graphical statistics, 19(2), 332-353.

Carreira-Perpinán, M. A., & Williams, C. K. (2003, June). On the number of modes of a Gaussian mixture. In International Conference on Scale-Space Theories in Computer Vision (pp. 625-640). Berlin, Heidelberg: Springer Berlin Heidelberg.

Jose, K. K. (2011). Marshall-Olkin family of distributions and their applications in reliability theory, time series modeling and stress-strength analysis. Proc. ISI 58th World Statist. Congr Int Stat Inst, 21st-26th August, 201, 3918-3923.

Muthama Musau, V., Gaetan, C., & Girardi, P. (2022). Clustering of bivariate satellite time series: a quantile approach. arXiv e-prints, arXiv-2207.

Morris, K., Punzo, A., McNicholas, P. D., & Browne, R. P. (2019). Asymmetric clusters and outliers: mixtures of multivariate contaminated shifted asymmetric Laplace distributions. Computational Statistics & Data Analysis, 132, 145-166.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data dependent. priors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(1), 159-180.

Wichitaksorn, N., Choy, S. B., & Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. Canadian Journal of Statistics, 42(4), 579-596.

Zhang, Y., Wang, H. J., & Zhu, Z. (2019). Quantile-regression-based clustering for panel data. Journal of Econometrics, 213, 54–67.