Concept Note: MaDoctoral Study

|  |  |  |
| --- | --- | --- |
| **Full names and surname** | NJAGI HENRY MURIMI | |
| **Student number** | PA301/S/21076/23 | |
| **Telephone number** | +254 714819645 | |
| **Email address** | HNJAGIM@GMAIL.COM | |
| **Proposed Supervisor** | DR. VICTOR MUSAU | |
| **Proposed Co-Supervisor if applicable** |  | |
|  | □ | Msc. Actuarial Science |
|  | □ | Msc. Applied Mathematics |
|  | □ | Msc. Data Science and Analytics |
| **Degree**  (fill to mark your choice on the tick box) | □  □ | Msc. Pure Mathematics  Msc. Statistics |
|  | □ | Ph.D. Applied Mathematics |
|  | □ | Ph.D. Pure Mathematics |
|  |  | Ph.D. Statistics |

**PROPOSED AREA OF STUDY**

The proposed area of study is Statistical Modelling and Data Analysis with a Focus on Time Series

Clustering Using Gaussian Finite Mixture Models.

**PROPOSED TITLE**

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

**Brief description and statement of the research question:**

**Brief description**

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. The adjustment

aims at overcoming flaws arising from a single Gaussian density and providing some flexibility

enables the model to adapt quickly enough toward complicated dynamics coming across time series patterns.

**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.

**Stipulate three to five keywords describing your research.**

Versatile Modification of Gaussian Finite Mixture Models; Time Series Clustering; Multiple Quantile

Clustering; Statistical Modeling.

**Introduction and rationale**

**Introduction**

Statistics is key for the development of models used to comprehend and describe real-life phenomena. The

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.

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 times.

series data.

Model-based clustering is a prevalent approach in the field of statistical modeling, 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, 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.

**Problem Statement and Gap/Niche**

This research is motivated by the perceived limitations of current statistical modeling approaches, especially

in respect to 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 which will enhance our knowledge of dynamics clusters.

**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.

**Summary of Preliminary 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 Scrucca (2015) emphasizes that the most common assumption is a one-to-one correspondence between

clusters and components of mixtures; he argues 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. 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.

**Proposed Research Methodology and Method**

**Introduction**

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.

**Research Design/Approach**

They will have an exploratory research design that combines quantitative and qualitative methods for

unmasking patterns within time series data sets. This combination of quantitative and qualitative techniques

facilitates a comprehensive analysis that looks at the numerical efficacy of modifying GMM while capturing

important factors identified clusters.

**Population, Sample, and Sampling Method**

This study population includes time series data sets with different features and structures. A purposive

sample will be collected from time series datasets available in the public domain and considered relevant to

the research aim. 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. At least 50 time series datasets will be selected to ensure

adequate data for model building and verification.

**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.

**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.

**Ethical Considerations**

Participant Anonymity and Confidentiality: All measures should 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 a team of researchers engaged with the data. A

data encryption technique will 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 research team will perform a risk assessment to address possible ethical risks arising

from the study. Risks identified will be communicated clearly to participants during the informed consent

process.

**REFERENCES**

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.