Concept Paper:

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

NJAGI HENRY MURIMI

PA301/S/21076/23

February 15, 2024

Presentation Outline

1. Description of proposed study
   - Statement of Research Question
2. Introduction and Rationale
   - Statement of the Problem
   - Objectives of the study
3. Materials and Methods
   - Model development
   - Data Analysis
   - Data Source
4. References

Description of proposed study

- The research aims to develop new statistical modeling techniques for time series data clustering.
- Offer a modified GMM for time series clustering with multiple quantiles.
- Model will overcome flaws in single Gaussian density, provide flexibility for adapting to complex time series patterns.

Statement of the Research Question

How does a generalized version of the Gaussian Finite Mixture Model improve accuracy and flexibility in time series clustering, particularly clustering at multiple quantiles?

Introduction

- Most of the proposed approaches concern univariate time series (UTS) while clustering of multivariate time series (MTS) has received much less attention (López-Oriona et al., 2022)
- Within time series analysis, a versatile modification of Gaussian Finite Mixture Models enhances accuracy and flexibility in clustering. The method will utilize a slight modification of the parameter López-Oriona and Vilar (2021)
- Musau et al. (2022) proposed the need to perform clustering at multiple quantiles instead of fixing the levels of quantiles while taking caution to avoid the issue of crossing quantiles.

Cont'

When performing clustering at multiple quantiles, there could be a risk of inconsistent or conflicting cluster assignments across different quantiles.

- Alfo et al. (2017) proposed a versatile approach that incorporates a broader application of finite mixture models in statistical analysis, particularly extending to multivariate cases.

Rationale

- To enhance accurate and adaptive clustering in time series, the research will develop an innovative approach integrating a slight modification of parameters.
- The goal is to improve the model's adaptability and learning capabilities, providing a more accurate representation of complex time series data patterns.

Statement of the Problem

- The current state of Clustering relies largely on traditional models that often overlook the dynamic interplay of multiple quantiles.
- To address the challenges, the study proposes a more customizable modification of Gaussian Finite Mixture Models addressing the drawbacks of a single Gaussian density and exploring clustering at various quantiles.
- Integrating time series data and advanced reinforcement learning technique, the proposed model will provide a more accurate clustering.

Objectives of the Study

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

Specific Objectives

1. To develop and Implement a Generalized Gaussian Finite Mixture Model (GMM) for Time Series Clustering.
2. To estimate model parameters of the model using Expectation - Maximization algorithm.
3. To evaluate and Quantify the Improvement in Clustering Accuracy.
4. Investigate and Measure the Effect of Clustering for Several Quantiles.

Model development

- Design the probability density function through a finite mixture model of G components.

$$f(x_i;\Psi) = \sum_{k=1}^{G} \pi_k f_k\left(x_i;\theta_k\right) \tag{1}$$

In this case $\Psi = \{\pi_1, ..., \pi_{G-1}, \theta_1, ..., \theta_G\}$ are the parameters of the mixture model. $f_k\left(x_i;\theta_k\right)$ is the kth component density for observation x_i with parameter vector $\theta_k = \left\{\underline{\mu}_k, \sum_k\right\}$.

Cont'd

- Mixing weights or probabilities $(\pi_1, ..., \pi_{G-1})$ (such that $\pi_k > 0, \sum_{k=1}^{G} \pi_k = 1$) and G is the number of mixture components.
- Compute the membership weights given parameters Ψ:

$$w_{ik} = p\left(z_{ik} = 1 | \underline{x}_i, \Psi\right) = \frac{p_k\left(\underline{x}_i | z_k, \theta_k\right).\pi_k}{\sum_{m=1}^{K} p_m\left(\underline{x}_i | z_m, \theta_m\right).\pi_m} \tag{2}$$

Cont'd

- Define the Expectation-Maximization (EM) algorithm. Start from some initial estimate of Ψ (random), update Ψ iteratively until convergence is detected.
- E-Step: Denote the current parameter values as Ψ, compute w_{ik} for all data points $\underline{x}_i, 1 \leq i \leq N$ and all mixture components $1 \leq k \leq K$ to yield an $N \times K$ matrix of membership weights. Calculate the expected value of the log-likelihood function given the current parameter estimates,

Cont'd

- M-Step: Use the membership weights and the data to calculate new parameter values. Update the parameter estimates to maximize the expected log-likelihood calculated in the E-Step.

$$N_k = \sum_{i=1}^{N} w_{ik} \tag{3}$$

The sum of the membership weights for the kth component - this is the effective number of data points assigned to component k.

- Compute the new mixture weights

$$\pi_k^{new} = \frac{N_k}{N}, 1 \leq k \leq K \tag{4}$$

Cont'd

- Update the mean

$$\underline{\mu}_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik}.\underline{x}_i 1 \leq k \leq K \tag{5}$$

- Compute the Covariance matrix

$$\sum_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik}.\left(\underline{x}_i - \underline{\mu}_k^{new}\right)(\underline{x}_i - \underline{\mu}_k^{new})^t 1 \leq k \leq K \tag{6}$$

Equation 6 requires first compute the K new π's, then the K new $\underline{\mu}_k$'s and finally the K new $\sum_k$'s.

Data Analysis

- The developed model will be implemented using Python programming language.
- The Expectation-Maximization algorithm will be used for parameter estimation.
- Initialize the GMM with the desired number of components using techniques such as, Bayesian Information Criteria (BIC) Fraley and Raftery (1998), Integrated complete-data likelihood criteria (ICL) Biernacki et al. (2000)
- Quantitative results will inform the development and evaluation of the model.

Data Source

- Economic data will be sourced from different sources such as the Kenya National Bureau of Statistics, World bank, and international trade databases for GDP Growth Rate, unemployment rate, and trade balance data; Central bank of Kenya and foreign investment reports for Interest rate, Exchange Rate, Foreign Direct Investment (FDI)

Selected References

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588.

López-Oriona, Á. and Vilar, J. A. (2021). Quantile cross-spectral density: A novel and effective tool for clustering multivariate time series. *Expert Systems with Applications*, 185:115677.

López-Oriona, Á., Vilar, J. A., and D'Urso, P. (2022). Quantile-based fuzzy clustering of multivariate time series in the frequency domain. *Fuzzy Sets and Systems*, 443:115–154.

THANK YOU