Review Form 3

Journal Name:	**Asian Journal of Probability and Statistics**
Manuscript Number:	**Ms_AJPAS_128648**
Title of the Manuscript:	**Expectation Maximization Algorithm for Weighted k-components Gaussian Mixture: An Application to High Frequency Financia Data.**
Type of the Article	**Financial data modelling, inferential statistics, and model selection.**

General guidelines for the Peer Review process:

This journal's peer review policy states that **NO** manuscript should be rejected only on the basis of '**lack of Novelty'**, provided the manuscript is scientifically robust and technically sound.
To know the complete guidelines for the Peer Review process, reviewers are requested to visit this link:

https://r1.reviewerhub.org/general-editorial-policy/

Important Policies Regarding Peer Review

Peer review Comments Approval Policy: https://r1.reviewerhub.org/peer-review-comments-approval-policy/
Benefits for Reviewers: https://r1.reviewerhub.org/benefits-for-reviewers

PART 1: Comments

	Reviewer's comment	**Author's Feedback** *(Please correct the manuscript and highlight that part in the manuscript. It is mandatory that authors should write his/her feedback here)*
Please write a few sentences regarding the importance of this manuscript for the scientific community. A minimum of 3-4 sentences may be required for this part.	The paper investigates the application of finite Gaussian mixtures to model financial variations, a modelling framework within this domain. The proposed approach aims to capture the cyclic and inherently multi-modal nature of financial time series data, which cannot be adequately represented by standard unimodal Gaussian distributions. By leveraging univariate Gaussian mixture models, the study addresses a critical gap in existing methods that often fail to reflect the complex structure of financial variations. A notable advantage of this approach lies in its interpretability, which contrasts with the more complex and less transparent "black-box" methodologies commonly employed in the field, such as Long-Short Term Memory (LSTM) networks and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). This interpretability makes the model particularly suitable for practical applications where clarity of results is essential. The authors utilize the Expectation-Maximization (EM) algorithm, a well-established and widely adopted technique for estimating the parameters of Gaussian mixture models. The choice of EM is justified by its theoretical convergence properties and its flexibility in handling incomplete datasets. Furthermore, the paper introduces an innovative variation of the Bayesian Information Criterion (BIC) for determining the optimal number of components in the Gaussian mixture model. Specifically, the method involves analysing sequential variations of the BIC to define an appropriate cut-off point for selecting the final number of components. This refinement enhances the reliability of model selection, addressing a common challenge in Gaussian mixture modelling.	
Is the title of the article suitable? (If not please suggest an alternative title)	Apart from correcting spacing issues in the title, I would reframe it to highlight the contributions of the paper, specifically the first-ever application of Gaussian mixture models (GMMs) to model high-frequency financial datasets. Additionally, since finite mixture models are inherently "weighted," this term can be omitted, as can the reference to "k components." Finally, using the EM algorithm for parameter estimation in GMMs is the most popular approach, accordingly, I would not emphasize it as a aspect. Accordingly, I would suggest the following title that better captures the innovations brought by the paper: (or Unprecedent) Application of Finite (Univariate) Gaussian Mixture Models for High-Frequency Financial Data Modelling"	
Is the abstract of the article comprehensive? Do you suggest the addition (or deletion) of some points in this section? Please write your suggestions here.	The abstract of the paper is comprehensive and effectively outlines the structure, key simulations, and findings. However, a few improvements could enhance its clarity and conciseness: - Clarify "high-frequency" financial data: Define the minimal time point or ranges to qualify varying time points as high-frequency. - Provide a shorter description of the dataset used for evaluating the relevance of GMMs approaches. In particular, there's no need in the abstract to provide the exact time stamps used for extracting soy bean variations. - Remove redundant phrases: Notably, omit "under the assumption of missing information," as Gaussian mixture models (GMMs) are inherently suited to multi-modal distributions, and this assumption is implicit in their application. - Define non-standard terms: Provide a brief explanation of the "BIC gradient" approach for determining the number of components in GMMs (even though I would rather qualify the method as Elbow's method as it's very similar, report to section Optional comments) - Reduce abstract length: Condense the abstract to avoid unnecessary details while maintaining focus on the key contributions of the paper. Suggested Revision: The abstract effectively presents the study but could benefit from concise wording and clarity on terminology.	

Review Form 3

Is the manuscript scientifically, correct? Please write here.	• Throughout the manuscript, I noticed the interchangeable use of the terms "mixture models" and "weighted models." I recommend removing any mention of "weighted models," as this terminology is equivocal: indeed, “weighted models” often refer to weighted inferential approach, where weights reflect the importance of specific observations, and whose scope is much larger than finite mixture models (for instance, can be used to trim outliers). Furthermore, variations of mixture models’ programmatic implementations that incorporate weighting for observations do exist, such as those implemented in the **mclust** R package (see https://mclust-org.github.io/mclust/reference/me.weighted.html vignette, with argument *weights*), or for uncertain prior cluster assignments in the **bgmm** R package, implementing “soft-label” and “belief-based” modelling: https://www.jstatsoft.org/article/view/v047i03 -> as a conclusion, to avoid any ambiguity, I suggest consistently referring to the methodology as "mixture models."
	• Regarding the motivation for using the EM algorithm, the justification of "missing data" alone is insufficient alone. First, I suggest elaborating that the "missing data" here specifically refers to **unknown cluster assignments**, which render direct maximization of the log-likelihood function infeasible. In greater details, a standard maximisation of the incomplete log-likelihood is unfeasible as the formula involves cancelling a log of sums, preventing the derivation of a closed-form solution for the roots of its gradient (aka the **mapping function**): explaining this computational challenge would provide more clarity and rigor to the argument.
	• There are typographical errors and inconsistencies in the formulas that should be addressed:
	  ○ The product of scalar terms is inconsistently denoted using the symbol * in some equations, and sometimes not.
	  ○ Use standard LaTeX commands for operators such as **$\log$**, **$\arg$**, and **$\max$** for better readability and conformity with conventional formatting. When using non-native mathematical operators, use the following set of Latex commands: *\DeclareMathOperator{\operatorname}{DisplayName}* used as *\operatorname{}*
	  ○ Between equations (12) and (13), the second component proportion, 1−ω, should be enclosed in brackets, .
	  ○ Several equations exceed the page margins and should be reformatted to fit within the column width.
	  ○ (While the use of ^{Z_i} and ^{1 - Z_i} is valid for only considering in the estimation observations associated with the cluster indexed as *i*, replacing these with an indicator variable or a Dirac function may improve consistency with statistical notations.)
	• In scientific writing, **Equations**, **Tables**, and **Figures** must be embedded within float environments and referenced in the main body text with proper cross-references using $\ref$ command (e.g., “as shown in Equation (12)” or “refer to Table 3”,). This ensures seamless integration of mathematical content with the text and improves the flow of the document.
	• Additionally, in the Adobe Acrobat file, I noticed the presence of unresolved comments or annotations. These should be trimmed or finalized before the submission of the final version to ensure the manuscript's professional presentation.Addressing these points will significantly enhance the technical clarity, mathematical precision, and overall presentation of the paper.	

Are the references sufficient and recent? If you have suggestions of additional references, please mention them in the review form.	Identified at least two papers that are worth citing, and certainly benchmarked against, that use parametric or non-parametric mixture models to capture the dynamics of “high-frequency” financial datasets: - *Time-Varying Gaussian-Cauchy Mixture Models for Financial Risk Management,* https://arxiv.org/abs/2002.06102, Uses Gaussian-Cauchy mixture models for financial risk management. - *Gaussian Mixture and Kernel Density-Based Hybrid Model for Volatility Behavior Extraction From Public Financial Data*, in which a hybrid model utilizing a Gaussian Mixture Model is used for analyzing foreign exchange market volatility, https://www.mdpi.com/2306-5729/4/1/19. - See also this post that lists comprehensively clustering approaches for dealing with longitudinal datasets: https://www.linkedin.com/embed/feed/update/urn:li:ugcPost:7258503366748651521, slide deck on which you could focus your interest on density-based estimation methods, notably on the Latent Class Growth Analysis clustering approach. I would additionally recommend reading and mentioning the following benchmarks of univariate (Nityasuddhi and Böhning 2003; Lourens et al. 2013; Leytham 1984; Xu and Knight 2010) and multivariate and high-dimensional (Chassagnol et al. 2024) Gaussian mixture models to address and select the best triple combination of an initialisation algorithm – stopping criteria – metric to optimise (RMSE and Variance).	
Is the language/English quality of the article suitable for scholarly communications?	While the paper is clear and well-structured, making it accessible even to non-experts in the field, I noticed several grammatical errors, spelling mistakes, and awkward and non-natural phrasing that detract from its readability and professionalism. Below are some specific examples: - “which does not change it’s meaning”: The word "it’s" is incorrectly used instead of "its," leading to a grammatical error. - “more than one group of data coming from more than one Gaussian distribution”: **Corrected**: "multiple groups of data, each originating from different Gaussian distributions", to avoid redundancy. - "We therefore could not make conclusions" **Corrected**: ". Therefore, we could not draw conclusions - …. (and many others) - The ChatGPT prompt I used to that end: For this text snippet <original_text>, identify all spelling or grammatical mistakes, using bullet points for each identified error with this format: "original text": grammatically correct and rephrased text. To address these issues, I strongly recommend either LLM-based tools such as *Quillbot* or *Grammarly,* or referring to a translator, to systematically identify and correct grammatical and spelling mistakes while improving the overall fluency and naturalness of the language: I’m not indeed a native English speaker.	

Review Form 3

Optional/General comments	To be considered for acceptance, I would require the following **minimum revisions** to address critical issues:
	
	• **Code and Dataset Availability**: Provide the code and dataset (or at least a URL link) necessary to reproduce the results presented in the paper, particularly the parameter estimations and simulations used to generate Figure 2. Add the programming language used along with the software and hardware configuration of your modelling environment (for instance in R, use the reprex https://reprex.tidyverse.org/ recommendations + *set.seed()* if using random approach for initialising the parameters of your mixture model). Reproducibility is a fundamental requirement for the validation of scientific contributions.
	• **Extended Benchmarking**: Broaden the scope of benchmarking by including additional datasets, whether real-world or simulated, of longitudinal nature. This will enhance the robustness of the conclusions and ensure the approach generalizes beyond the specific dataset analyzed in the paper.
	• **Address Rolling/Moving Window Selection**:
	  ○ In Figure 2, the selected rolling window appears to involve both left- and right-truncated data, as the time points are partially removed. This raises questions about the validity of the parameter estimates derived within such truncated intervals.
	  ○ To address this issue, I recommend either:
	  ▪ Proposing an automated approach to define a **rolling window** that avoids truncation, ensuring that only complete cycles are captured (although I do not specifically recommend this restrictive approach).
	  ▪ Alternatively, implement a mechanism to **detect truncation**—both left- and right-truncation—so its effect can be captured during the inferential process (for instance, by assuming the left and the right-side proceed from left and right-truncated Gaussian distributions, while the potential remaining modes and cycles could be captured using standard Gaussian mixtures).
	• **Mandatory Addition of One of the Following Benchmarking Approaches**:
	To address the lack of comparative analysis and methodological depth, I recommend incorporating **at least one** of the following extensions to the study:
	1. **Comparison with Alternative Time-Series Models**: Benchmark the univariate Gaussian Mixture Model (GMM) approach against other well-established methods for financial time series modelling, such as Long Short-Term Memory (LSTM) networks or ARIMA. These approaches were mentioned in the introduction as alternatives but were not empirically compared to the proposed GMM.
	2. **Evaluation of Alternative Distributions**: Assess the performance of the GMM framework against other parametric or non-parametric distributions that may be better suited to capturing financial time series dynamics, especially given the data's potential heavy tails, skewness, or multi-modal nature.
	3. **Alternative Metrics and EM Variants**: Investigate whether alternative metrics (e.g., Integrated Completed Likelihood, or ICL) or modifications of the Expectation-Maximization (EM) algorithm (e.g., Variational EM or Stochastic EM) could improve parameter estimation or convergence speed. Specifically, compare their performance in terms of both convergence behaviour and parameter precision.
	4. **Model Selection Criteria**: Benchmark global model selection metrics (such as BIC or AIC) against more specialized techniques, such as the MML (Minimum Message Length) framework implemented in tools like the *GMKMcharlie* package. This would allow for a more rigorous evaluation of the chosen mixture components and parameterization.
	
	These revisions are critical to address the methodological gaps in the paper, without any of them, the study lacks sufficient statistical breadth and rigor to be accepted.	

PART 2:

	Reviewer's comment	**Author's comment** ***(if agreed with reviewer, correct the manuscript and highlight that part in the manuscript. It is mandatory that authors should write his/her feedback here)***
Are there ethical issues in this manuscript?	***(If yes, Kindly please write down the ethical issues here in details)***	

Reviewer Details:

Name:	**Bastien Chassagnol**
Department, University & Country	**Sorbonne University, France**