EM Algorithm: A Versatile Tool for Parameter Estimation in Non-Ideal Data Scenarios where MLE Falls Short

Wilson Geronimo
Department of Mathematics and Statistics
Thompson Rivers University
British Columbia, Canada
geronimorodriguezw22@mytru.ca

***Abstract*—The Expectation-Maximization (EM) algorithm is a powerful tool for parameter estimation, especially in scenarios where Maximum Likelihood Estimation (MLE) encounters challenges. This study investigates EM's efficacy in handling non-ideal data scenarios where MLE falls short. Our aim is to showcase EM's superiority and versatility in providing accurate parameter estimates. Through empirical analysis using real-world datasets, including financial time series data, we illustrate instances where EM outperforms MLE, particularly in heavy-tailed distributions and complex data structures. Our results highlight EM's robustness, even in the presence of outliers and non-normal distributions. We discuss the implications for future research and practical applications, emphasizing EM's potential utility in finance, healthcare, and machine learning. This study underscores the importance of considering alternative estimation techniques like EM in situations where traditional methods may yield inferior results. By leveraging EM's flexibility and robustness, researchers and practitioners can analyze complex data more effectively, leading to more informed decisions and potentially groundbreaking advancements across various fields.**

***Index Terms*—Expectation-Maximization (EM), Maximum Likelihood Estimation (MLE), heavy-tailed distributions, financial time series data.**

I. Introduction

In the era of data-driven decision-making, robust and accurate parameter estimation techniques are fundamental for extracting meaningful insights from complex datasets. Maximum Likelihood Estimation (MLE) and the Expectation-Maximization (EM) algorithm stand as pillars in the realm of statistical estimation, playing pivotal roles in various fields of data science, including machine learning, econometrics, and computational biology.

A. *Maximum Likelihood Estimation (MLE)*

MLE is a cornerstone method for estimating the parameters of statistical models based on observed data. By maximizing the likelihood function, MLE aims to find the parameter values that make the observed data most probable under the assumed model. Due to its simplicity and asymptotic properties, MLE has become a standard approach for parameter estimation in many statistical models.

B. *Expectation-Maximization (EM) Algorithm*

While MLE offers a powerful framework for parameter estimation, it may encounter challenges in scenarios with incomplete or non-identifiable data. The EM algorithm, introduced by Dempster, Laird, and Rubin (1977) [2], addresses these limitations by providing a flexible framework for maximum likelihood estimation in the presence of latent variables or missing data. Through an iterative two-step procedure involving the E-step (Expectation) and M-step (Maximization), EM iteratively refines parameter estimates until convergence, even in complex data scenarios.

C. *Significance in Data Science*

In the era of big data, where datasets often exhibit intricate structures and may contain missing or incomplete information, the importance of robust parameter estimation techniques cannot be overstated. MLE and EM play critical roles in a wide range of data science applications, including clustering, classification, and density estimation. From financial modeling to healthcare analytics, these methods empower researchers and practitioners to derive actionable insights from diverse datasets, contributing to informed decision-making and innovation across industries.

II. Literature Review

Moon, T. K. (1996) [3] in their paper studied about EM algorithm. The estimation of probability distribution function parameters is a common task in signal processing, particularly in scenarios involving noisy signals. One prevalent problem is estimating the mean of a signal in noise. However, many estimation problems become more complex due to inaccessible data or missing data points, such as in binning operations or histograms. The Expectation-Maximization (EM) algorithm has emerged as a powerful tool for addressing such challenges. It is particularly well-suited for scenarios where there is a many-to-one mapping from an underlying distribution to the observed distribution. The EM algorithm comprises two main steps: an expectation step, where the unknown underlying variables are estimated, and a maximization step, which updates the parameter estimates based on the observed data.

These steps are iterated until convergence. Initially discovered and employed independently by various researchers, Dempster consolidated their ideas, proved convergence, and termed it the "EM algorithm." Since then, numerous studies across various domains have utilized the EM algorithm to tackle a wide range of estimation problems.

Borman, Sean (2004) [1] in their paper explains how the EM algorithm is a powerful iterative technique designed to compute Maximum Likelihood (ML) estimates, especially in situations involving missing or hidden data. ML estimation aims to find the model parameter(s) that make the observed data most probable. Each iteration of the EM algorithm comprises two key steps: the E-step and the M-step. During the E-step, missing data are estimated based on the observed data and the current model parameter estimates. This estimation relies on the concept of conditional expectation, hence the name "expectation" step. In the subsequent M-step, the likelihood function is maximized, assuming that the missing data are known. The estimates of the missing data obtained from the E-step are used in place of the actual missing data.

One notable feature of the EM algorithm is its guarantee of convergence, as the likelihood is guaranteed to increase with each iteration. This property ensures that the algorithm will eventually reach an optimal solution.

In this study, we present a derivation of closed-form estimates for autoregressive (AR) models with Cauchy innovations using the Expectation-Maximization (EM) algorithm. We conduct a comparative analysis of the proposed EM algorithm against the Maximum Likelihood (ML) method using simulated data. ML estimation is performed using a built-in function in R. An advantage of the EM algorithm is its simultaneous calculation of both the model and innovation parameters. Our results, visualized through boxplots, demonstrate the superior performance of the EM algorithm over the ML method. Additionally, we explore an alternative approach based on characteristic functions (CF) for estimating AR model parameters with stable distributions. In future research, we aim to compare our proposed algorithm and the CF-based estimation method with existing techniques outlined in references [5, 6, 7] for AR models with infinite variance. Moreover, we anticipate applying the proposed model and methodologies to analyze real-life phenomena.

III. Objectives

The primary objective of this study is to investigate the effectiveness of the Expectation-Maximization (EM) algorithm for parameter estimation in scenarios where Maximum Likelihood Estimation (MLE) struggles due to non-ideal data distributions. Specifically, the study aims to achieve the following objectives:

- **Assessing Parameter Estimation Methods**: Compare the performance of MLE and EM algorithms in estimating model parameters using both real-world financial data (Bitcoin price data) and simulated data with a Cauchy distribution. Evaluate the accuracy and robustness of parameter estimates obtained from each method under varying levels of data non-normality and heavy-tailedness.
- **Identifying Non-Gaussian Patterns**: Identify and characterize non-Gaussian patterns and behaviors present in the Bitcoin price data and simulated data with a Cauchy distribution. Explore the impact of extreme price movements, volatility clustering, and other non-Gaussian phenomena on parameter estimation outcomes.
- **Understanding Algorithmic Advantages**: Investigate the theoretical underpinnings and computational advantages of the EM algorithm compared to MLE in handling non-Gaussian data distributions. Analyze how the iterative nature of EM facilitates parameter estimation in the presence of data outliers and heavy tails.
- **Implications for Financial Modeling**: Explore the implications of using EM-based parameter estimation techniques in financial modeling and risk management applications. Assess the potential benefits of incorporating non-Gaussian modeling approaches for capturing tail risk, volatility dynamics, and other features commonly observed in financial time series data.
- **Contributions to Data Science**: Contribute to the broader field of data science by advancing understanding and methodologies for parameter estimation in non-ideal data scenarios. Provide insights into the practical applications of EM algorithms for addressing challenges associated with non-Gaussian data distributions in various domains beyond finance.

By addressing these objectives, this study aims to shed light on the suitability and effectiveness of EM-based parameter estimation techniques for analyzing complex data with non-Gaussian characteristics, thereby enhancing our ability to model and interpret real-world phenomena in data science and beyond.

IV. Methodology

A. *Parameter Estimation Methods*

- **Maximum Likelihood Estimation (MLE)**: MLE was employed to estimate the parameters of the underlying probability distribution function (PDF) based on the observed data. The negative log-likelihood function was minimized using optimization techniques to obtain the MLE estimates.
- **Expectation-Maximization (EM) Algorithm**: The EM algorithm was utilized as an alternative parameter estimation method, particularly suitable for scenarios where the data distribution is unknown or non-Gaussian. EM iteratively estimates the parameters by maximizing the expected log-likelihood, leveraging latent variables to model complex data structures.

B. *Model Fitting and Evaluation*

- **MLE Parameter Estimation**: The MLE algorithm was applied to the Bitcoin price data and simulated Cauchy data to estimate the location and scale parameters of the underlying distributions. The optimization process

aimed to minimize the negative log-likelihood function, capturing the discrepancy between the observed data and the model predictions.

- **EM Parameter Estimation**: The EM algorithm, implemented through the normalmixEM function in the mixtools package, was employed to fit mixture models to the observed data. EM iteratively updated the component means, variances, and mixture weights to maximize the likelihood of the observed data under the mixture model framework.

C. Performance Comparison

- **Accuracy Assessment**: The parameter estimates obtained from MLE and EM methods were compared in terms of their accuracy, robustness, and computational efficiency. Statistical metrics such as mean, scale, and covariance were used to evaluate the quality of parameter estimates and assess the fit of the models to the data.

D. Software and Tools

- **R Statistical Software**: Data processing, parameter estimation, and statistical analysis were conducted using the R programming language. R packages including mixtools, mvtnorm, and MASS were utilized for fitting mixture models, generating synthetic data, and performing numerical optimization.

By following this methodology, the study aims to comprehensively compare the performance of MLE and EM algorithms in estimating model parameters for non-Gaussian data, providing insights into their strengths, limitations, and practical implications for data analysis and modeling.

V. Data and variables

A. Data Collection

- **Real-World Financial Data**: Historical Bitcoin price data was obtained from Yahoo Finance, covering the period from April 1, 2021, to July 31, 2021. This dataset provides a representative sample of volatile and non-Gaussian financial time series data.
- **Simulated Data**: Synthetic data following a Cauchy distribution was generated to simulate heavy-tailed and non-Gaussian data characteristics. The simulated dataset was constructed to mimic the extreme fluctuations and fat-tailed distributions observed in certain financial markets.

B. Data Preprocessing

Before conducting parameter estimation experiments, both datasets undergo preprocessing steps to ensure their suitability for analysis. This includes handling missing values, filtering data to the desired time period (for Bitcoin data), and selecting relevant features for estimation. Additionally, data exploration techniques may be employed to gain insights into Bitcoin's price behavior and identify patterns that could influence parameter estimation results.

The dataset used in this study comprises historical price data of Bitcoin (BTC-USD) obtained from Yahoo Finance, as well as simulated data with a Cauchy distribution. Bitcoin, a decentralized digital currency, has garnered significant attention as a volatile and widely traded asset in financial markets. The availability of high-frequency price data makes Bitcoin an ideal candidate for studying parameter estimation techniques in the context of non-ideal data scenarios.

The Bitcoin price data spans April 1, 2021, to July 31, 2021, covering a period of 4 months. It consists of daily opening, high, low, and closing prices, along with trading volumes and adjusted closing prices. The historical price data provide insights into Bitcoin's price dynamics, including trends, volatility, and trading activity, which are essential for assessing the performance of parameter estimation algorithms.

The simulated data with a Cauchy distribution is generated to mimic non-Gaussian, heavy-tailed data characteristics often observed in financial markets. The Cauchy distribution is known for its fat tails and lack of finite moments, making it a suitable choice for modeling extreme price movements and other non-Gaussian phenomena.

C. Statistical Properties

The Bitcoin price data exhibit characteristics typical of financial time series, including volatility clustering, non-normality, and occasional extreme fluctuations. These properties pose challenges for traditional parameter estimation methods like Maximum Likelihood Estimation (MLE), which assume Gaussian distributions and may struggle to capture the nuances of heavy-tailed or non-Gaussian data distributions.

VI. Results

A. Parameter Estimates

Maximum Likelihood Estimation (MLE):

For the Bitcoin price data collected from April 1, 2021, to July 31, 2021, MLE yielded parameter estimates of the underlying distribution as follows:

- **Mean:** $36,980.51$
- **Scale:** $5,016.306$

Expectation-Maximization (EM) Algorithm:

In contrast, the EM algorithm produced the following parameter estimates for the same dataset:

- **Component 1:**
  - **Mean:** $35,719.7$
  - **Scale:** $3,037.96$
  - **Mixture Weight:** 62.21%
- **Component 2:**
  - **Mean:** $56,134.2$
  - **Scale:** $4,211.60$
  - **Mixture Weight:** 37.79%
- **Robustness to Outliers**: In the EM Parameter Estimates, the mean (μ) for component I is approximately \$35,719.70, while for component II, it is around \$56,134.20. These estimates reflect the central tendency of the data, even in the presence of extreme values or outliers. The EM algorithm assigns a relatively low weight (λ) to component 1 (around 62.20%), indicating

Estimate	Component I	Component II
λ	6.22059×10^{-1}	3.77941×10^{-1}
μ	3.57197×10^{4}	5.61342×10^{4}
σ	3.03796×10^{3}	4.21160×10^{3}
loglik at estimate	-1246.217	

TABLE I
SUMMARY OF NORMALMIXEM OBJECT

that it captures the majority of the data with moderate fluctuations. Component 2, with a higher mean and lower weight, likely represents the tail of the distribution, capturing extreme price movements or outliers in the Bitcoin prices.

- **Iterative Optimization Process**: The iterative nature of the EM algorithm enables it to refine parameter estimates over multiple iterations, gradually converging to a better solution. The log likelihood at estimate, approximately -1,246.217, indicates the goodness of fit of the model to the data. Through successive iterations, EM maximizes this likelihood, leading to more precise parameter estimates.

In conclusion, the MLE estimates for the mean and scale of the Cauchy distribution deviate notably from the EM estimates, suggesting that MLE may struggle to accurately model the distribution of Bitcoin prices during periods of high volatility. Meanwhile, the EM algorithm appears to provide more reasonable estimates for the distribution parameters.

This outcome supports the scenario where MLE fails to find accurate estimates while EM succeeds in capturing the underlying distribution of the data, demonstrating the robustness of the EM algorithm in handling heavy-tailed distributions such as the Cauchy distribution.

VII. SUMMARY AND DISCUSSIONS

In this study, we explored the effectiveness of Maximum Likelihood Estimation (MLE) and the Expectation-Maximization (EM) algorithm in estimating parameters for non-ideal data scenarios. We began by collecting Bitcoin price data from April 1, 2021, to July 31, 2021, and simulated data using a Cauchy distribution.

Our results indicate that while MLE provides parameter estimates, it struggles with non-ideal data distributions such as those with heavy tails. In contrast, the EM algorithm demonstrates robustness in such scenarios and provides accurate parameter estimates even in the presence of outliers or heavy-tailed distributions.

These findings underscore the versatility and importance of the EM algorithm in parameter estimation tasks, particularly when dealing with non-ideal data scenarios. The ability of EM to handle complex data distributions makes it a valuable tool in various fields, including finance, healthcare, and natural sciences.

Overall, our study highlights the significance of considering alternative estimation methods like the EM algorithm when MLE falls short, emphasizing the importance of selecting appropriate methods based on the characteristics of the data at hand.

REFERENCES

[1] Sean Borman. The expectation maximization algorithm-a short tutorial. *Submitted for publication*, 41, 2004.

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[3] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

Github Link:
https://github.com/wgeronimor/EMAlgorithmwhereMLEFalls