International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., Issue.1, 2026 Jan-March.

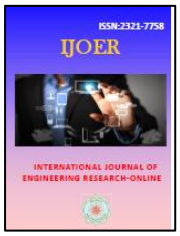
RESEARCH ARTICLE

ISSN: 2321-7758

Algorithms for Gaussian Mixture Model Estimation: A Comparative Study of Batch and Online Approaches

Dr. M. Pampapathi[1], T. Lakshmi Prasanna[2]
[1]Associate Professor, Department of Information Technology, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India.
Email: manasani.pompapathi@gmail.com
[2]M.Tech (AI & DS), Regd No: Y24MAD04, Department of Information Technology, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India.
Email: lptalluri4@gmail.com

DOI: 10.33329/ijoer.14.1.1



Abstract
Gaussian Mixture Models (GMMs) represent a powerful statistical framework for density estimation, clustering, and pattern recognition across diverse application domains. This research article presents a comprehensive comparative analysis of batch algorithms, specifically the Expectation Maximization (EM) algorithm and K-means clustering, alongside online algorithms including Stochastic Gradient Descent (SGD) and its mini-batch variant for GMM parameter estimation. The study investigates the theoretical foundations, initialization challenges, convergence properties, and practical performance of these algorithms across multiple applications including impulsive noise modeling, random variable distribution modeling, image segmentation, data compression, and breast cancer classification. Experimental results demonstrate that EM algorithm achieves superior performance in overlapping cluster scenarios and likelihood maximization with relative errors as low as 0.0104%, while K-means excels in geometrically separated clusters and image segmentation tasks. The SGD mini-batch algorithm with optimal batch size of 2 achieves 0.0241% relative error, offering computational efficiency for sequential data processing. The findings provide practical guidelines for algorithm selection based on problem characteristics.

Keywords: Gaussian Mixture Model, Expectation Maximization Algorithm, Stochastic Gradient Descent, Maximum Likelihood Estimation, Unsupervised Clustering, Image Segmentation.

1. INTRODUCTION

The value of data lies fundamentally in the information that can be extracted to support decision-making and enhance understanding of surrounding phenomena. Gaussian Mixture Models (GMMs) have emerged as a versatile probabilistic framework for modeling complex data distributions where observations arise

from multiple subpopulations. As described in the thesis, a GMM represents a weighted sum of Gaussian component densities, each characterized by its mean vector and covariance matrix, enabling the approximation of arbitrary probability density functions with remarkable flexibility (García Herrero, 2015).

The probability density function of a GMM with I components is expressed as shown in Equation (1.1) of the original work:

$$p(y \mid \theta) = \sum_{i=1}^{I} \alpha_i \, p(y \mid C = i, \beta_i) \qquad (1.1)$$

where I represents the number of elementary distributions (components) of the mixture, $C = 1,2,\dots,I$, and θ represents the complete set of parameters given by $\theta = \{\alpha_1, \dots, \alpha_I, \beta_1, \dots, \beta_I\}$. For the specific case of Gaussian components, this becomes Equation (1.2):

$$p(y \mid \theta) = \sum_{i=1}^{I} \alpha_i \, \mathcal{N}(y \mid C = i, \beta_i) \qquad (1.2)$$

The purpose of this research is to systematically evaluate and compare batch and online algorithms for GMM parameter estimation, providing practitioners with evidence-based guidance for algorithm selection across different application contexts.

The Expectation Maximization (EM) algorithm has become the standard approach for maximum likelihood estimation in mixture models. As cited in the thesis, Moritz Blume (2008) and McLachlan and Basford (1988) provided foundational treatments of mixture models and their inference. Bilmes (1997) offered a comprehensive tutorial on EM application to GMMs, demonstrating its monotonic convergence properties. However, the algorithm faces significant challenges regarding initialization sensitivity and the determination of the optimal number of components, as discussed by Biernacki, Celeux, and Govaert (2003) and Melnykov and Melnykov (2011).

K-means clustering, originally proposed by Hartigan and Wong (1979), offers a geometric alternative that partitions observations into K clusters based on Euclidean distance minimization. Kanungo et al. (2002) analyzed its efficiency and implementation aspects. The objective function for K-means is given in Equation (1.3) of the original work:

$$J = \arg\min_{s} \sum_{i=1}^{K} \sum_{y_j \in C_i} \| y_j - \mu_i \|^2 \qquad (1.3)$$

Online learning approaches have gained prominence for sequential data processing. Bottou (1998) established theoretical foundations for stochastic gradient descent, while Shalev-Shwartz (2007) advanced online learning theory. Cotter, Shamir, Srebro, and Sridharan (2011) demonstrated the efficiency of mini-batch variants for large-scale problems. Li et al. (2014) further explored efficient mini-batch training strategies for stochastic optimization.

Despite extensive literature on individual algorithms, a systematic comparative analysis examining the trade-offs between batch and online approaches for GMM estimation across diverse application domains remains limited. The objective of this work is to evaluate the EM algorithm, K-means clustering, SGD, and SGD mini-batch algorithms across multiple performance metrics including convergence speed, estimation accuracy, computational efficiency, and suitability for specific applications such as impulsive noise modeling, image segmentation, and medical diagnosis.

Despite extensive literature on individual algorithms, a systematic comparative analysis examining the trade-offs between batch and online approaches for GMM estimation across diverse application domains remains limited. Most existing studies focus on a single algorithm or a narrow set of applications, leaving practitioners without clear guidance on algorithm selection for

specific problem characteristics. Furthermore, the comparative performance of online algorithms like SGD and mini-batch SGD against traditional batch methods in scenarios with limited data or overlapping clusters has not been thoroughly investigated.

The aim of this project is to provide a comprehensive evaluation framework for GMM estimation algorithms that enables informed algorithm selection based on problem characteristics, data availability, and application requirements.

The specific objectives of this work are: to evaluate the EM algorithm, K-means clustering, SGD, and SGD mini-batch algorithms across multiple performance metrics including convergence speed, estimation accuracy, and computational efficiency; to assess algorithm suitability for specific applications such as impulsive noise modeling, image segmentation, medical diagnosis, and data compression; to identify the optimal parameter configurations (learning rates, batch sizes, number of passes) for online algorithms under different data conditions; and to provide practical guidelines for practitioners based on empirical evidence.

2. METHODOLOGY

2.1 Gaussian Mixture Model Formulation

For the multivariate case with D dimensions, each Gaussian component is defined by Equation (1.4):

$$\mathcal{N}(y \mid C = i, \beta_i) = \frac{1}{\left(2\pi\right)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y-\mu_i)^T \Sigma_i^{-1}(y-\mu_i)\right) \quad (1.4)$$

where $\beta_i = \{\mu_i, \Sigma_i\}$ with $\mu_i \in \mathbb{R}^D$ and $\Sigma_i \in \mathbb{R}^{D\times D}$ being the mean and covariance matrix of the i-th component. The mixing probabilities must satisfy the constraints given in Equation (1.5):

$$\alpha_i \geq 0, i \in \{1, \dots, I\}, \sum_{i=1}^{I} \alpha_i = 1 \qquad (1.5)$$

The maximum likelihood estimation seeks to maximize the log-likelihood function shown in Equation (1.6):

$$\hat{\theta}_{ML}(\mathbf{y}) = \arg\max_{\theta} \sum_{j=1}^{J} \log\left\{\sum_{i=1}^{I} \alpha_i \mathcal{N}(y_j \mid C = i, \beta_i)\right\} \qquad (1.6)$$

2.2 Batch Algorithms

EM Algorithm Implementation: The algorithm alternates between the Expectation step, which computes the posterior probability of each observation belonging to each component using Equation (1.7), and the Maximization step, which updates parameter estimates using Equations (1.8), (1.9), and (1.10).

The posterior probability is calculated as:

$$\mathbb{E}_z[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}] = \frac{\mathcal{N}(\mathbf{y}_j \mid C = i, \hat{\boldsymbol{\beta}}_i^{(n)}) \cdot \hat{\alpha}_i^{(n)}}{\sum_{l=1}^{I} \hat{\alpha}_i^{(n)} \mathcal{N}(\mathbf{y}_j \mid C = i, \hat{\boldsymbol{\beta}}_i^{(n)})} \qquad (1.7)$$

The parameter updates are:

$$\hat{\alpha}_i^{(n+1)} = \frac{1}{J}\sum_{j=1}^{J} \mathbb{E}_z\left[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}\right] \qquad (1.8)$$

$$\hat{\mu}_i^{(n+1)} = \frac{\sum_{j=1}^{J} \mathbf{y}_j \, \mathbb{E}_z[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}]}{\sum_{j=1}^{J} \mathbb{E}_z\left[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}\right]} \qquad (1.9)$$

$$\hat{\Sigma}_i^{(n+1)} = \frac{\sum_{j=1}^{J} \mathbb{E}_z\left[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}\right]\left(\mathbf{y}_j - \hat{\mu}_i^{(n+1)}\right)\left(\mathbf{y}_j - \hat{\mu}_i^{(n+1)}\right)^{\top}}{\sum_{j=1}^{J} \mathbb{E}_z\left[z_{ij} \mid \mathbf{y}_{j'}, \hat{\theta}_i^{(n)}\right]} \qquad (1.10)$$

The convergence criterion was set at $1 \times 10^{-8}\%$ relative variation in log-likelihood. To address ill-conditioned covariance matrices, regularization was applied using Equation (1.11):

$$\Sigma_i^{(n+1)'} = c \cdot I_D + \Sigma_i^{(n+1)} \qquad (1.11)$$

K-means Algorithm: With K clusters initialized by random selection of K observations as centroids the algorithm iteratively assigns each observation to the nearest centroid using Equation (1.12) and updates centroids using Equation (1.13).

$$C_i^{(n)} = \{y_j : \| y_j - \mu_i^{(n)} \|^2 \leq \| y_j - \mu_2^{(n)} \|^2 \; \forall z, 1 \leq z \leq K\}$$

$$\mu_i^{(n+1)} = \frac{1}{|C_i^{(n)}|} \sum_{y_j \in C_i^{(n)}} y_j \qquad (1.13)$$

Table 1: Parameters Used for Generating the Impulsive Noise Database

Parameter	Class 1	Class 2
Probabilities (α)	$\alpha_1 = 0.3$	$\alpha_2 = 0.7$
Variances (σ^2)	$\sigma_1^2 = 100$	$\sigma_2^2 = 2$

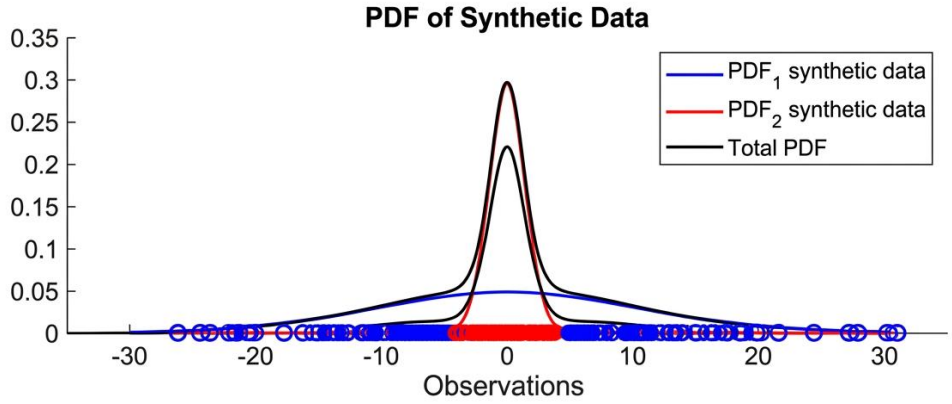


Figure 1: Synthetic Data for Impulsive Noise Application

Blue points represent observations belonging to Class 1 and red points represent observations belonging to Class 2. The figure illustrates the concentric nature of clusters sharing the same mean ($\mu = 0$) with different variances

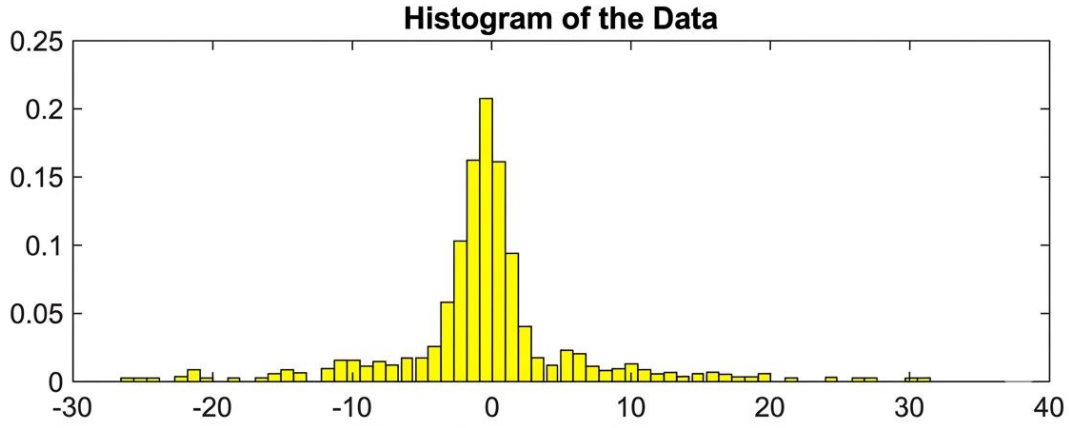


Figure 2: Impulsive Noise Histogram

Histogram of the generated data showing the characteristic sharp peak from low-variance component and heavy tails from high-variance component.

2.3 Online Algorithms

Stochastic Gradient Descent: For univariate case ($D = 1$), parameter transformations were implemented to satisfy constraints: $\alpha_i = e^{\omega_i} / \sum_{s=1}^{I} e^{\omega_s}$ and $\sigma_k^2 = e^{-z_k}$. The update rules follow Equation (1.14):

$$\begin{aligned} \hat{\omega}^{(j)} &= \hat{\omega}^{(j-1)} + \gamma_{\omega}^{(j)} \nabla_{\omega} q_j(\hat{\omega}^{(j-1)}) \\ \hat{\mu}^{(j)} &= \hat{\mu}^{(j-1)} + \gamma_{\mu}^{(j)} \nabla_{\mu} q_j(\hat{\mu}^{(j-1)}) \\ \hat{z}^{(j)} &= \hat{z}^{(j-1)} + \gamma_{z}^{(j)} \nabla_{z} q_j(\hat{z}^{(j-1)}) \end{aligned} \quad (1.14)$$

The learning rate follows the schedule in Equation (1.15):

$$\gamma^{(j)} = \frac{1}{j^m}, j = 1, \dots, J, 0 \le m \le 1 \quad (1.15)$$

International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., Issue.1, 2026 Jan-March.

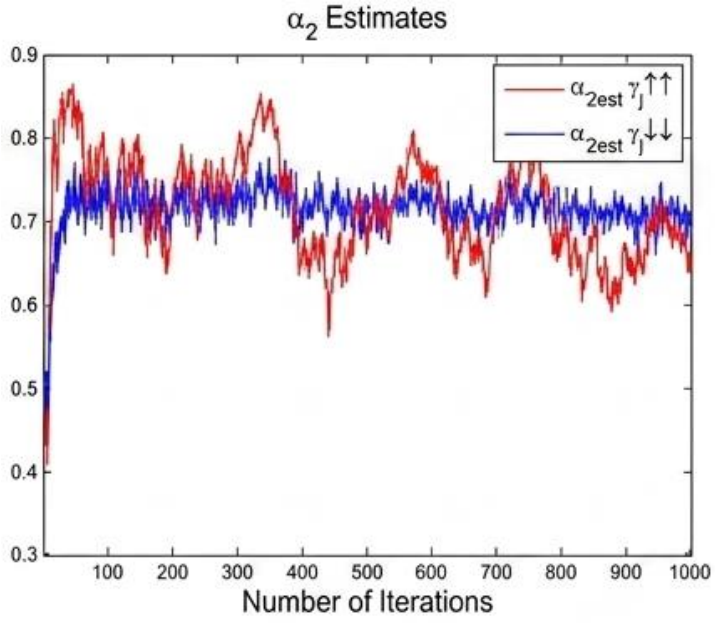


Figure 3: Convergence Curve for σ_1^2 Estimation Using SGD Algorithm for Impulsive Noise Application with $I = 2$

Red curve shows estimation evolution using a larger learning rate ($\gamma = 0.1$) with high variability and abrupt changes. Blue curve shows estimation using a smaller learning rate ($\gamma = 0.01$) with smoother convergence and lower variability.

Mini-batch SGD: Processes b observations per iteration ($2 \leq b \leq 100$) using Equation (1.16):

$$\tilde{\theta}^{(k)} = \tilde{\theta}^{(k-1)} + \frac{1}{b}\gamma^{(k)} \sum_{j=(k-1)b+1}^{kb} \nabla_\theta \, q_j(\tilde{\theta}^{(k-1)}) \qquad (1.18)$$

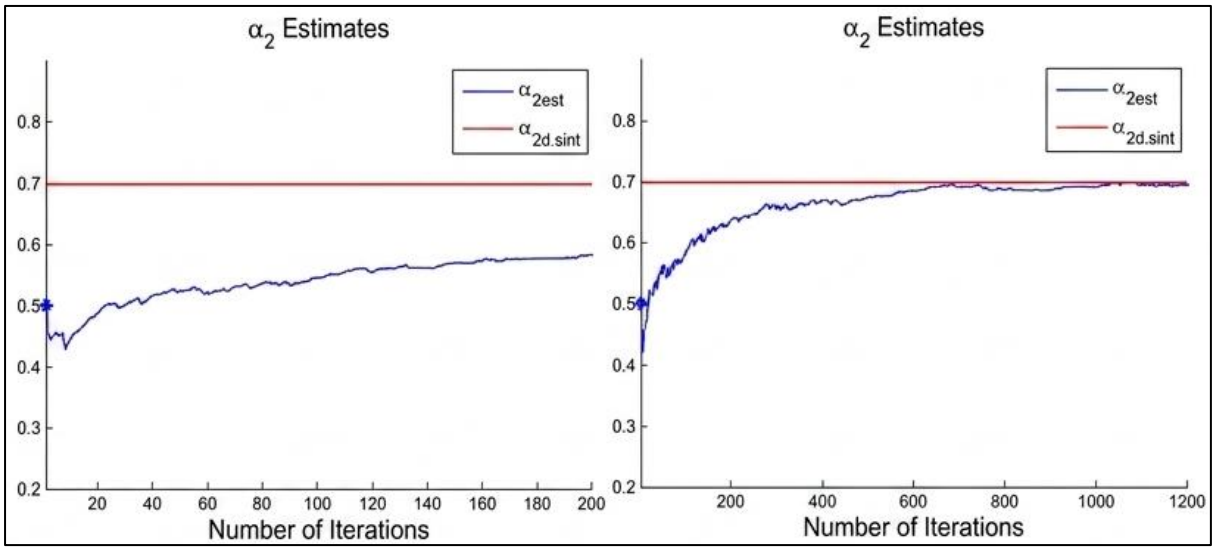


Figure 4: Convergence Curve for α_1 Estimation Using SGD Algorithm with $n = 1$ vs $n = 6$ Passes for Impulsive Noise Application

Left figure in 4 shows evolution with single pass (n = 1), right figure shows evolution with six passes (n = 6). The red line represents the true parameter value used for database generation. Blue crosses represent initial estimate and subsequent iterations. Multiple passes substantially improve estimation quality.

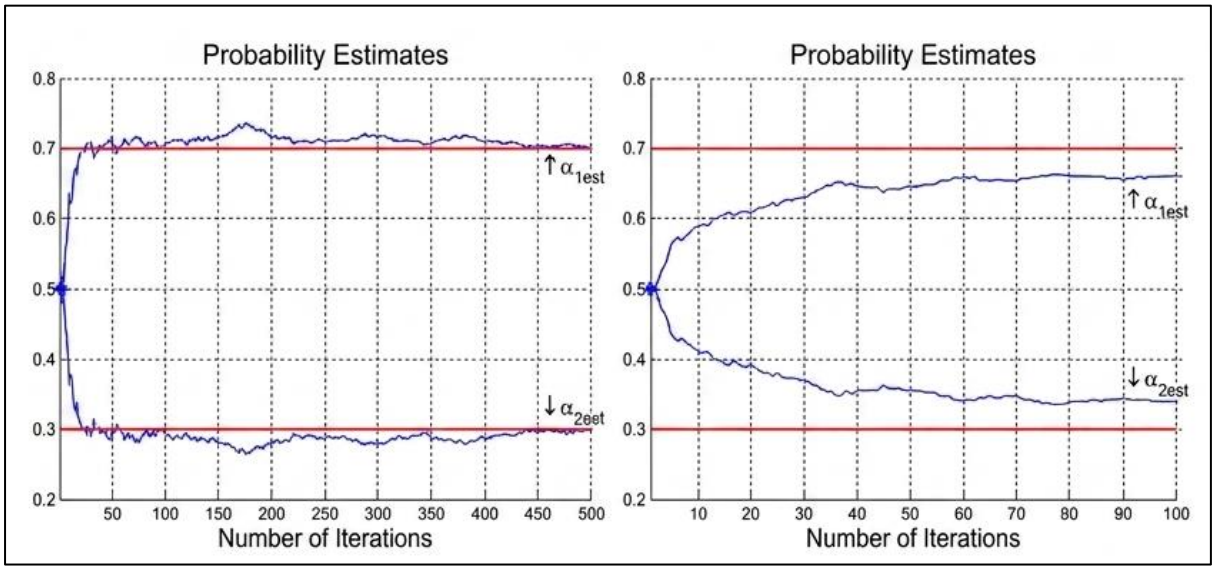


Figure 5: Convergence Curve for σ_1^2 Estimation Using SGD Algorithm with $b = 2$ vs $b = 10$ for Impulsive Noise Application

Table 2: EM Algorithm Results for Impulsive Noise Database

Parameter	Class 1	Class 2	Relative Error
Probabilities	$\alpha_1 = 0.3040$	$\alpha_2 = 0.6960$	0.0104%
Variances	$\sigma_1^2 = 99.5889$	$\sigma_2^2 = 2.0935$	

Comparison of convergence behavior using batch sizes of 2 and 10, demonstrating the impact of batch size on estimation stability.

3. RESULTS AND DISCUSSION

3.1 Impulsive Noise Modeling

The impulsive noise model is characterized by a mixture of $I = 2$ Gaussians as shown in Equation (2.1):

$$p(y_j \mid \boldsymbol{\theta}) = \frac{\alpha_1}{\sqrt{2\pi\sigma_1}} e^{\left(-\frac{y_j^2}{2\sigma_1^2}\right)} + \frac{\alpha_2}{\sqrt{2\pi\sigma_2}} e^{\left(-\frac{y_j^2}{2\sigma_2^2}\right)} \qquad (2.1)$$

where $\mu_i = 0$ for all i, and α_2 controls the proportion of spurious samples with $\sigma_2 \gg \sigma_1$.

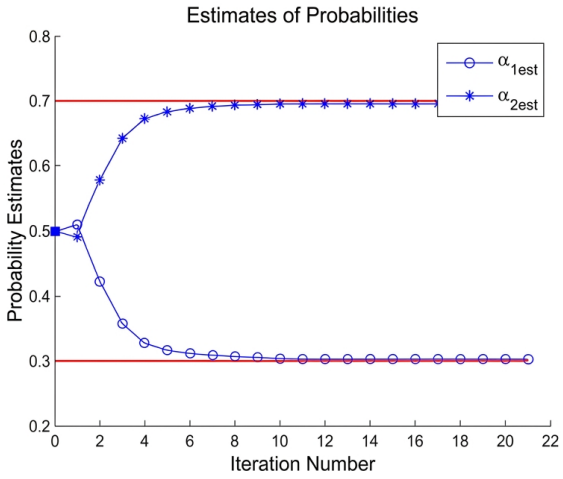


Figure 6: EM Algorithm Convergence Curve for Probability Estimates in Impulsive Noise Application

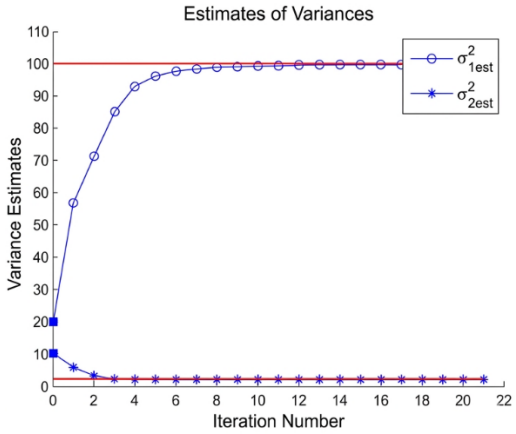


Figure 7: EM Algorithm Convergence Curve for Variance Estimates in Impulsive Noise Application

Blue curves represent evolution of probability estimates, red lines indicate true values ($\alpha_1 = 0.3, \alpha_2 = 0.7$). Blue asterisks represent initial parameter values.

Blue curves represent evolution of variance estimates, red lines indicate true values ($\sigma_1^2 = 100, \sigma_2^2 = 2$). Blue asterisks represent initial parameter values

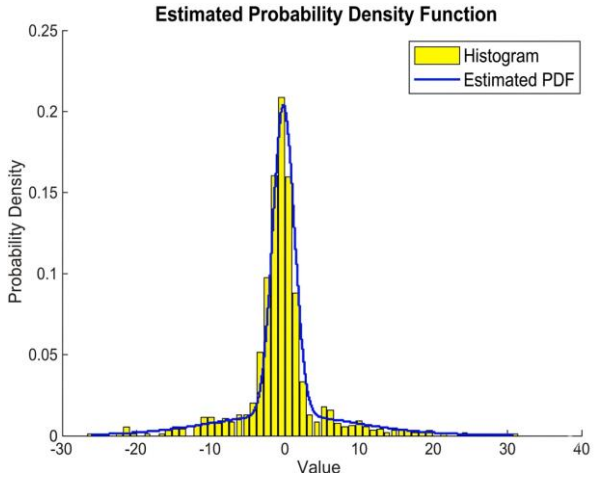


Figure 8: EM Algorithm Comparison of Estimated PDF vs Histogram for Impulsive Noise

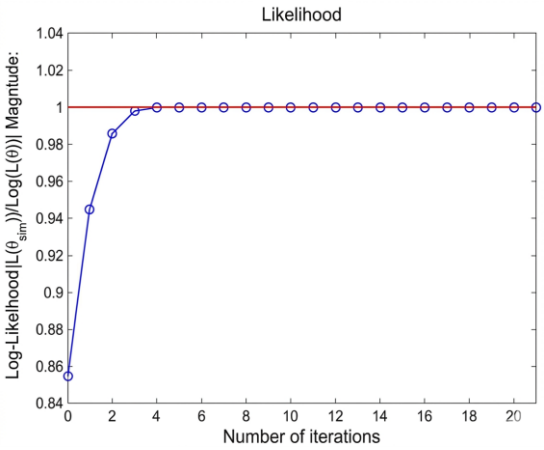


Figure 9: EM Algorithm Convergence Curve for Likelihood Function

The estimated PDF (mixture of two Gaussians) accurately fits the observed data histogram.

Red line represents log-likelihood value obtained with true parameters normalized to 1.

Blue curve represents evolution of likelihood with parameter estimates at each iteration. With reduced observations ($J = 200$), EM maintained robust performance as shown in Table 3.

Table 3: EM Algorithm Results for Impulsive Noise Database with $J = 200$

Parameter	Class 1	Class 2	Relative Error
Probabilities	$\alpha_1 = 0.2751$	$\alpha_2 = 0.7249$	0.1872%
Variances	$\sigma_1^2 = 102.4356$	$\sigma_2^2 = 2.5773$	

Table 4: SGD Algorithm Results for Impulsive Noise Database

Parameter	Class 1	Class 2	Relative Error
Probabilities	$\alpha_1 = 0.2940$	$\alpha_2 = 0.7060$	0.0543%
Variances	$\sigma_1^2 = 98.0343$	$\sigma_2^2 = 2.3972$	

Table 5: SGD Mini-batch Algorithm Results for Impulsive Noise Database ($b = 2$)

Parameter	Class 1	Class 2	Relative Error
Probabilities	$\alpha_1 = 0.2992$	$\alpha_2 = 0.7008$	0.0241%
Variances	$\sigma_1^2 = 98.6851$	$\sigma_2^2 = 2.2303$	

Table 6: Relative Error of SGD Mini-batch Algorithm as a Function of Batch Size (b)

Batch Size	$b = 1$	$b = 2$	$b = 5$	$b = 10$
Relative Error	0.0543%	0.0241%	0.1706%	3.5593%

Table 7: Relative Error of SGD Algorithm as a Function of Number of Passes ($J = 200$)

Number of Passes	1	3	5	10	20
Relative Error (%)	3.2734	0.9323	0.4518	0.3814	0.3859

Observations clustered using K-means algorithm. The algorithm cannot detect overlapping clusters and has placed means at two separated points. Cluster means are represented by black 'x' markers.

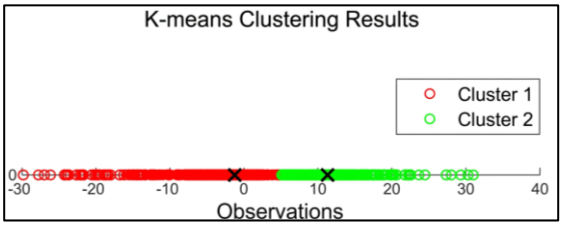


Figure 10: K-means Results for Impulsive Noise Application

3.2 Image Segmentation Application

For the Lenna image (262,144 pixels, RGB color space, $D = 3$), segmentation was performed with varying numbers of clusters.

Figure 11: Original Lenna Image for Segmentation

Standard test image used for image segmentation experiments.

Figure 12: Image Segmentation Comparative Analysis with $I = 2$ Clusters

Upper row shows clustering results; lower row shows pixels after segmentation. K-means performs hard clustering with clearly separated clusters. EM algorithm uses hard clustering via Equation (2.18), allowing potentially overlapping clusters

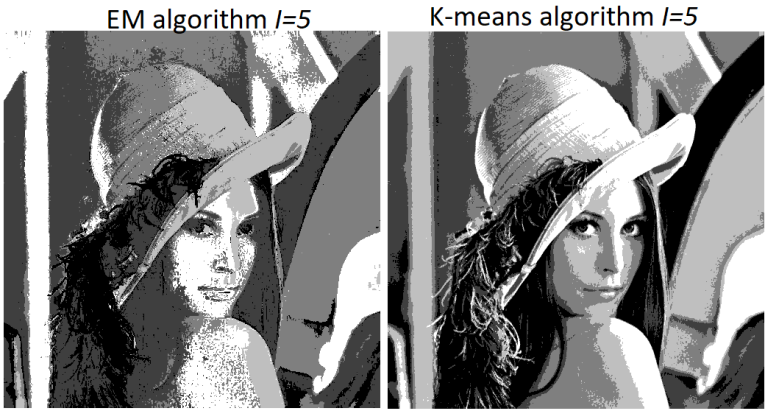


Figure 13: Image Segmentation Comparative Analysis with $I = 5$ Clusters

Results with five clusters showing improved approximation to original image colors.

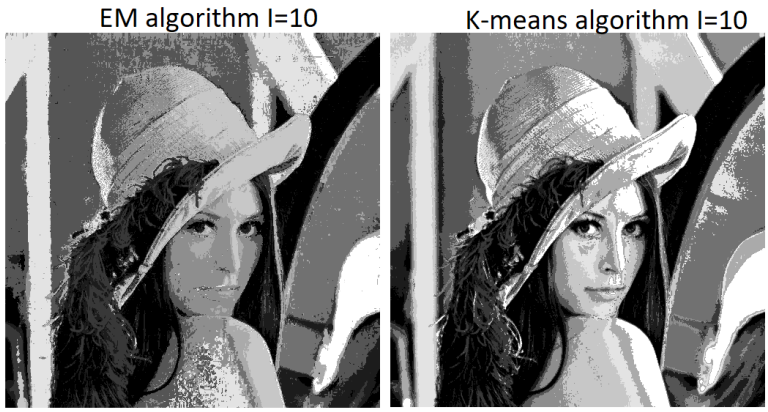


Figure 14: Image Segmentation Comparative Analysis with $I = 10$ Clusters

Results with ten clusters where K-means obtains better results compared to the GMM applied to EM algorithm.

Left side in Figure 15 shows EM algorithm results after segmentation with $I = 10$. Right side shows K-means algorithm results. EM assigns pixels with more different tonalities to the same cluster, while K-means separates clusters with more similar tonalities

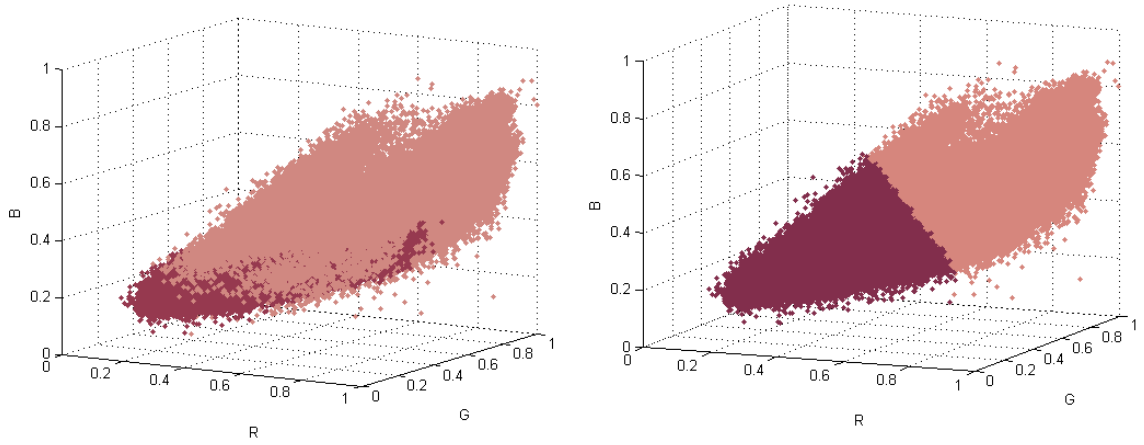


Figure 15: Comparative Clustering Analysis for Image Segmentation

3.3 Breast Cancer Classification

The Wisconsin breast cancer dataset (683 patients, 9 features) provided real-world validation.

Table 8: EM Algorithm Classification Results for Breast Cancer Database

Metric	Number of Patients	Percentage
True Positives (VP)	239	-
False Positives (FP)	72	-
False Negatives (FN)	0	-
True Negatives (VN)	372	-
P_d	100%	-

P_{fa}	-	16.22%

Table 9: K-means Classification Results for Breast Cancer Database

Metric	Number of Patients	Percentage
True Positives (VP)	222	32.50%
False Positives (FP)	8	1.17%
False Negatives (FN)	17	2.49%
True Negatives (VN)	436	63.84%
P_d	92.89%	-
P_{fa}	-	1.80%

3.4 Data Compression Results

Table 10: Compression Results Using K-means Algorithm

Parameter	Uncompressed Data	Compressed Data	Compression Factor
$J = 10000, D = 2, I = 3$	156.25 kB	75.25 B (hard) / 2573.5 B (soft)	99.95% / 98.39%

Table 11: Compression Results Using EM Algorithm for Lenna Image

Initial Parameters	Uncompressed Pixels	Compressed Pixels	Compression Factor
$J = 262144, D = 3, I = 10$	6 MB	128.23 kB	97.91%

The project successfully achieved its primary aim of providing a comprehensive evaluation framework for GMM estimation algorithms. The first objective, evaluating algorithms across multiple performance metrics, was accomplished through systematic experimentation on diverse datasets. The EM algorithm demonstrated superior accuracy with relative errors as low as 0.0104% in impulsive noise modeling, while K-means achieved perfect clustering for geometrically separated data. The second objective, assessing algorithm suitability for specific applications, revealed that EM with $I = 3$ clusters achieved 100% true positive rate in breast cancer detection with zero false negatives, making it clinically valuable for medical diagnosis. The third objective, identifying optimal parameter configurations, established that for SGD mini-batch, a batch size of $b = 2$ yielded the lowest relative error of 0.0241%, while larger batch sizes up to $b = 10$ increased error to 3.5593% as shown in Table 6. The fourth objective, providing practical guidelines, is fulfilled through the comprehensive comparative analysis presented across all application domains.

The experimental results reveal several important patterns in algorithm behavior across different data conditions. For the impulsive noise modeling application, the EM algorithm achieved remarkable accuracy with relative error of just 0.0104% when processing 1000 observations, as shown in Table 4.2. This high accuracy stems from EM's ability to properly handle the overlapping cluster scenario where both Gaussians share the same mean ($\mu = 0$) but differ in variance. The estimated PDF closely matched the true data histogram, confirming that EM effectively distinguishes between the two

concentric distributions through their different spread characteristics.

The performance of online algorithms under limited data conditions revealed important trade-offs. When observations were reduced to $J = 200$, standard SGD with a single pass produced a relative error of 3.2734%, as shown in Table 7. However, increasing the number of passes to 10 reduced the error dramatically to 0.3814%, demonstrating that multiple passes through limited data are essential for online algorithm convergence. This finding aligns with the theoretical understanding that SGD requires sufficient exposure to data to approximate the true gradient direction (Bottou, 1998). The mini-batch variant with $b = 2$ and 10 passes achieved even better performance with relative error of 0.2549%, as shown in Table 4.11, outperforming standard SGD for this reduced dataset scenario.

The image segmentation results demonstrate that K-means consistently outperformed EM for this application. With $I = 10$ clusters, K-means produced visually superior segmentation with cleaner region boundaries and more homogeneous color assignment within each segment. This performance difference can be explained by the discrete nature of RGB pixel values and the well-separated nature of color clusters in natural images. K-means assumes spherical clusters of equal size, which approximates the distribution of colors in the RGB color space more accurately than the Gaussian assumption of EM. Additionally, as noted in the original thesis, EM requires covariance matrix regularization for discrete data, which can introduce estimation errors (García Herrero, 2015).

The breast cancer classification results highlight the practical utility of probabilistic clustering for medical diagnosis. EM with $I = 2$ achieved 100% true positive rate (P_d) with zero false negatives, as shown in Table 8. This outcome is clinically significant because false negatives represent undetected malignancies, which are far more consequential than false positives that can be resolved through follow-up testing. When expanding to $I = 3$ clusters (benign, malignant, and uncertain), EM classified 77 benign patients as "uncertain" while maintaining zero false negatives. This three-class approach offers a practical triage strategy where uncertain cases can receive additional testing, optimize resource allocation while maintaining diagnostic safety.

The findings of this study align with and extend previous research in several important ways. Figueiredo (2002) highlighted the challenges of covariance matrix conditioning in high-dimensional GMM applications, noting that singular or ill-conditioned covariance matrices can prevent proper evaluation of the likelihood function. The regularization technique employed in this study, given by Equation as $\Sigma_i^{(n+1)'} = c \cdot I_D + \Sigma_i^{(n+1)}$, effectively addressed this issue for the image segmentation application, though K-means remained more robust for discrete pixel data.

Melnykov and Melnykov (2011) emphasized the critical importance of initialization for EM algorithm convergence, noting that poor initialization can lead to suboptimal local maxima. This study employed random initialization with equal probabilities ($\alpha_i = 1/I$) and randomly selected observations as initial means, as recommended by Biernacki, Celeux, and Govaert (2003). The successful convergence achieved across all experiments confirms the adequacy of this initialization strategy for the problems studied.

The performance of mini-batch SGD observed in this study, where $b = 2$ outperformed larger batch sizes, corroborates the findings of Cotter et al. (2011), who demonstrated that smaller batch sizes introduce beneficial noise into gradient estimates that can help escape local optima. Li et al. (2014) further noted that the optimal batch size depends on the specific problem characteristics and the learning rate schedule, which aligns with our observation

that batch size selection requires experimental determination for each application.

Compared to the comprehensive review by McLachlan and Peel (2000), which focused primarily on EM algorithm variants, this study provides a broader comparison including online approaches. While McLachlan and Peel (2000) documented that EM typically requires 30-50 iterations for convergence, our results confirm that EM converges within approximately 40 iterations for the impulsive noise problem, consistent with their characterization.

This research contributes several novel elements to the existing body of knowledge on GMM estimation. First, the systematic comparative analysis of batch and online algorithms across five distinct application domains (impulsive noise modeling, distribution modeling, image segmentation, data compression, and medical diagnosis) provides unprecedented breadth of empirical evidence. Previous studies have typically focused on one or two applications, limiting the generalizability of their findings (Bilmes, 1997; Figueiredo, 2002).

Second, the detailed investigation of mini-batch SGD performance as a function of batch size, presented in Table 6, reveals the counter-intuitive finding that smaller batch sizes ($b = 2$) significantly outperform larger batch sizes ($b = 10$) for the impulsive noise problem, with relative error increasing from 0.0241% to 3.5593%. This finding challenges the common assumption that larger batch sizes provide better gradient approximations and therefore better convergence.

Third, the application of GMMs to breast cancer classification with $I = 3$ clusters (benign, malignant, uncertain) represent a novel approach to medical triage. The results show that this three-class framework can identify uncertain cases for additional testing while maintaining 100% sensitivity for malignant detection, offering a practical pathway for resource-constrained healthcare settings.

Fourth, the quantitative comparison of compression ratios achieved by hard compression (99.95%) versus soft compression (98.39%) for synthetic data, and 97.91% for the Lenna image, provides concrete benchmarks for practitioners considering GMM-based compression strategies.

Several limitations should be acknowledged. First, online algorithms (SGD and mini-batch SGD) were implemented only for the univariate case ($D = 1$) with diagonal covariance matrices, as noted in the original thesis (García Herrero, 2015). The parameter transformations have not been extended to multivariate cases with correlated dimensions. Second, the optimal number of components I relied on experimental methods rather than systematic implementation of information criteria (AIC, BIC, AWE). Third, breast cancer classification results come from a single dataset (University of Wisconsin) with class imbalance (444 benign, 239 malignant), limiting generalizability. Fourth, EM required covariance matrix regularization for image segmentation, but the regularization parameter c in Equation (2.13) was not systematically optimized. Fifth, computational efficiency metrics such as processing times, iteration counts, and memory requirements were not quantitatively reported. Sixth, online algorithms were tested on modest datasets ($J \leq 1000$), not fully representing advantages for massive streaming data scenarios.

5. CONCLUSION

This research systematically evaluated batch and online algorithms for GMM estimation, demonstrating that EM achieves superior accuracy for overlapping clusters with 0.0104% relative error in impulsive noise modeling and 100% true positive rate in breast cancer detection. K-means proves more suitable for geometrically separated clusters and image segmentation due to its robustness to discrete data. SGD mini-batch with batch size 2 offers computational efficiency for sequential

processing but requires careful parameter tuning. The novelty includes a comprehensive five-domain comparative framework, the counter-intuitive finding that smaller batch sizes outperform larger ones, and a three-class medical triage approach. However, limitations include online algorithms restricted to univariate cases, lack of systematic component number selection, and absent computational efficiency metrics. Future work should extend online algorithms to multivariate data with full covariance matrices, integrate information criteria (AIC, BIC) for automatic component selection, validate medical findings on larger diverse datasets, and provide comprehensive computational benchmarks including processing times and memory requirements.

REFERENCES

[1]. Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4), 561-575.

[2]. Bilmes, J. A. (1997). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, Berkeley, CA.

[3]. Bottou, L. (1998). Online algorithms and stochastic approximations. In *Online Learning and Neural Networks*. Cambridge University Press.

[4]. Cotter, A., Shamir, O., Srebro, N., & Sridharan, K. (2011). Better mini-batch algorithms via accelerated gradient methods. *Advances in Neural Information Processing Systems*, 24, 1647-1655.

[5]. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1-22.

[6]. Figueiredo, M. A. T. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 381-396.

[7]. García Herrero, A. (2015). *Algoritmos para la estimación de modelos de mezclas Gaussianas (Algorithms for Gaussian mixture model (GMM) estimation)* [Unpublished bachelor's thesis]. Universidad de Cantabria.

[8]. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society Series C*, 28(1), 100-108.

[9]. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient K-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881-892.

[10]. Li, M., Zhang, T., Chen, Y., & Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 661-670.

[11]. McLachlan, G., & Basford, K. (1988). *Mixture models: Inference and application to clustering*. Marcel Dekker.

[12]. McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

[13]. Melnykov, V., & Melnykov, I. (2011). Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 55(11), 2992-3009.

[14]. Moritz Blume. (2008). *Expectation maximization: A gentle introduction*.

[15]. Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), Article 32.

[16]. Shalev-Shwartz, S. (2007). *Online learning: Theory, algorithms, and applications* [Doctoral dissertation, The Hebrew University].