Variational Fuzzy K-Means: A Free Energy-Based Approach

Abhishek Sharma[1]

[1]Independent Researcher

June 13, 2025

Abstract

We propose a variational reformulation of fuzzy clustering through a method we term Variational Fuzzy K-Means (VFKM). Rather than treating clustering as a purely geometric or heuristic procedure, VFKM minimizes a variational free energy objective derived from approximate Bayesian inference. It treats the cluster memberships as variational distributions over latent discrete variables and optimizes an energy functional composed of expected reconstruction error, entropy-based exploration, and a KL-divergence alignment with prior (or earlier) assignments. This formulation generalizes classical soft clustering models and offers a flexible pathway toward modeling uncertainty, performing stable optimization, and enabling future semi-supervised or prior-guided extensions. We empirically validate VFKM on benchmark datasets and show its competitive performance while enabling interpretable and principled fuzzy partitions.

Keywords: Clustering, Variational Inference, Free Energy, Fuzzy K-Means, Soft Assignments, Bayesian Learning

1 Introduction

Clustering is a cornerstone of unsupervised learning with broad applications in vision, health, and natural language domains. While classical methods like K-Means (MacQueen, 1967) provide efficient and interpretable partitioning, they rely on hard assignments and assume isotropic cluster structure. Probabilistic models like Gaussian Mixture Models (GMMs) offer soft assignments and density estimation, but introduce higher model complexity and sensitivity to dimensionality.

Fuzzy clustering approaches, including Fuzzy C-Means (Bezdek, 1981), relax hard assignments by introducing a membership matrix, allowing each point to belong to multiple clusters. However, such methods often lack a probabilistic grounding and can become unstable without principled regularization.

We propose a method termed **Variational Fuzzy K-Means (VFKM)**, which offers a variational perspective on fuzzy clustering. In VFKM, the soft cluster assignments are treated as variational distributions over latent discrete variables. This variational approach enables scalable uncertainty-aware clustering by providing interpretable posterior distributions and flexible regularization mechanisms absent in purely geometric methods. The model minimizes a free energy objective composed of: (1) an expected reconstruction term, corresponding to the expected negative log-likelihood; (2) an entropy regularizer that encourages uncertainty and exploration in the assignments; and (3) a KL divergence term that anchors the current posterior to prior estimates or previous assignments. This combination enables VFKM to act as a free-energy-minimizing fuzzy clustering framework—balancing data fit, stability, and probabilistic interpretability.

2 Background and Motivation

2.1 Fuzzy Clustering and Soft Assignments

Fuzzy clustering relaxes the hard assignment constraint of K-Means, allowing a point x_i to have membership weights u_{ik} for cluster k such that $\sum_k u_{ik} = 1$. This enables each point to softly belong to multiple clusters, offering a more nuanced representation of cluster boundaries.

In Soft K-Means, these assignments are computed as:

$$u_{ik} = \frac{\exp\left(-d_{ik}/T\right)}{\sum_j \exp\left(-d_{ij}/T\right)} \tag{1}$$

where $d_{ik} = \|x_i - \mu_k\|^2$ is the squared distance to cluster k, and T is a temperature parameter that controls the softness of the assignments.

The temperature T plays a critical role in shaping the distribution of memberships. At high temperatures, the exponentiated distances become less sensitive to differences, leading to more uniform (high-entropy) assignments across clusters. This encourages exploration and uncertainty. As T decreases, the softmax becomes sharper, and assignments become more concentrated around the nearest cluster, approximating hard clustering. Annealing T during training allows the algorithm to transition from exploratory to confident assignments, which is particularly useful in optimizing non-convex objectives and improving convergence.

2.2 Variational Inference and Free Energy

Variational inference approximates a posterior $p(Z|X)$ using a variational distribution $q(Z)$ by minimizing the Kullback-Leibler divergence $\mathrm{KL}(q(Z)||p(Z|X))$. This is equivalent to minimizing the **variational free energy**:

$$\mathcal{F}(q) = \mathbb{E}_q[-\log p(X|Z)] + \mathrm{KL}(q(Z)||p(Z)) \tag{2}$$

In our model, the membership weights u_{ik} serve as the variational distribution over discrete latent cluster indicators Z_i. The first term captures expected reconstruction cost under q, while the second encourages regularization via entropy (encouraging uncertainty) or KL divergence (encouraging closeness to priors or previous estimates). The free energy objective hence combines reconstruction fidelity with regularization: entropy maximization encourages exploration, while KL divergence ensures closeness to a prior assignment distribution $p(Z)$, promoting stability.

3 Model Formulation

3.1 Objective Function

The Variational Fuzzy K-Means (VFKM) objective is derived from a free energy minimization framework, balancing data reconstruction, uncertainty regularization, and prior alignment. Given N data points $\{x_i\}_{i=1}^N$ and K clusters, the objective minimizes the following functional:

$$\mathcal{L} = \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}\|x_i - \mu_k\|^2}_{\text{Expected Reconstruction}} - \underbrace{\lambda_{\text{entropy}} \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}\log u_{ik}}_{\text{Entropy Regularization}} + \underbrace{\lambda_{\text{kl}} \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}\log \frac{u_{ik}}{u_{ik}^{\text{prev}}}}_{\text{KL Prior Matching}} \tag{3}$$

The first term represents the expected reconstruction cost, quantifying the discrepancy between data points x_i and cluster centroids μ_k, weighted by the membership probabilities u_{ik}. This encourages accurate representation of the data by minimizing distortion.

The second term introduces entropy regularization. Here, $\sum u_{ik}\log u_{ik}$ is the negative entropy of the membership distribution. In a minimization objective, this term inherently discourages confident (low-entropy) assignments, encouraging soft and exploratory memberships. The hyperparameter λ_{entropy} controls its influence: higher values enforce smoother cluster assignments.

The third term represents KL prior matching, which penalizes deviations from a reference distribution u_{ik}^{prev}. This stabilizes optimization by discouraging abrupt shifts in cluster assignments. The coefficient λ_{kl} governs the strength of this temporal regularization.

Unlike heuristic fuzzy clustering, this formulation explicitly connects each component to the ELBO principle: minimizing reconstruction error, promoting assignment entropy, and enforcing stability through KL divergence. Together, these terms define a variational free energy functional that guides robust and interpretable fuzzy clustering.

3.2 Optimization Strategy

The optimization procedure of VFKM alternates between updating cluster centers and membership probabilities. The cluster centroids μ_k are updated by computing a weighted expectation over data points, where the membership probabilities u_{ik} act as soft responsibilities. This step aligns with a variational M-step and is formally expressed as:

$$\mu_k = \frac{\sum_{i=1}^{N} u_{ik} x_i}{\sum_{i=1}^{N} u_{ik}} \tag{4}$$

Following the centroid update, the membership probabilities u_{ik} are refined by balancing data fidelity, entropy smoothing, and KL anchoring derived from the variational free energy objective. The update takes the form:

$$u_{ik} \propto (u_{ik}^{\text{prev}})^{-\frac{\lambda_{\text{kl}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}}} \cdot \exp\left(-\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}}\right) \tag{5}$$

The first term adjusts the current membership probabilities by penalizing deviations from the previous assignments u_{ik}^{prev}. A higher λ_{kl} amplifies this prior anchoring effect, enforcing temporal consistency and stability in the clustering assignments. The second term encourages proximity to cluster centroids, where the effective scaling factor $\lambda_{\text{entropy}} - \lambda_{\text{kl}}$ acts analogously to a temperature parameter, modulating the sharpness of the assignment distribution. A larger denominator softens the assignments, promoting exploration, while a smaller value sharpens cluster memberships.

It is important to note that while the entropy regularization appears with a negative sign in the variational objective (promoting entropy maximization), its effect is inherently embedded through the scaling of the distance term. The interplay between λ_{entropy} and λ_{kl} governs the trade-off between exploratory soft assignments and stabilizing prior consistency, thereby guiding VFKM towards robust and interpretable fuzzy partitions.

3.3 Derivation of Membership Update from ELBO

The variational objective for VFKM minimizes the free energy functional:

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}\|x_i - \mu_k\|^2 - \lambda_{\text{entropy}} \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik} \log u_{ik} + \lambda_{\text{kl}} \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik} \log \frac{u_{ik}}{u_{ik}^{\text{prev}}} \tag{6}$$

We derive the optimal memberships u_{ik} by setting the derivative of $\mathcal{L}$ with respect to u_{ik} to zero, while enforcing $\sum_{k=1}^{K} u_{ik} = 1$ via Lagrange multipliers ζ_i:

$$\frac{\partial \mathcal{L}}{\partial u_{ik}} - \zeta_i = 0 \tag{7}$$

The Lagrangian for a single data point x_i is:

$$\mathcal{L}_i = \sum_{k=1}^{K} u_{ik}\|x_i - \mu_k\|^2 - \lambda_{\text{entropy}} \sum_{k=1}^{K} u_{ik} \log u_{ik} + \lambda_{\text{kl}} \sum_{k=1}^{K} u_{ik} \log \frac{u_{ik}}{u_{ik}^{\text{prev}}} - \zeta_i \left(\sum_{k=1}^{K} u_{ik} - 1\right) \tag{8}$$

To find the optimal u_{ik}, we take the partial derivative of $\mathcal{L}_i$ with respect to u_{ik} and set it to zero:

$$\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial u_{ik}} &= \|x_i - \mu_k\|^2 - \lambda_{\text{entropy}}(\log u_{ik} + 1) + \lambda_{\text{kl}} \left(\log \frac{u_{ik}}{u_{ik}^{\text{prev}}} + 1\right) - \zeta_i = 0 \\
&= \|x_i - \mu_k\|^2 - \lambda_{\text{entropy}} \log u_{ik} - \lambda_{\text{entropy}} + \lambda_{\text{kl}} \log u_{ik} - \lambda_{\text{kl}} \log u_{ik}^{\text{prev}} + \lambda_{\text{kl}} - \zeta_i = 0
\end{aligned}$$

Rearranging terms to solve for $\log u_{ik}$:

$$
\begin{aligned}
(\lambda_{\text{kl}} - \lambda_{\text{entropy}}) \log u_{ik} &= -\|x_i - \mu_k\|^2 + \lambda_{\text{entropy}} - \lambda_{\text{kl}} + \lambda_{\text{kl}} \log u_{ik}^{\text{prev}} + \zeta_i \\
\log u_{ik} &= \frac{-\|x_i - \mu_k\|^2 + \lambda_{\text{entropy}} - \lambda_{\text{kl}} + \lambda_{\text{kl}} \log u_{ik}^{\text{prev}} + \zeta_i}{\lambda_{\text{kl}} - \lambda_{\text{entropy}}} \\
\log u_{ik} &= -\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} - 1 - \frac{\lambda_{\text{kl}} \log u_{ik}^{\text{prev}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} - \frac{\zeta_i}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}}
\end{aligned}
$$

Now, exponentiate both sides:

$$
u_{ik} = \exp\left(-\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} - 1 - \frac{\lambda_{\text{kl}} \log u_{ik}^{\text{prev}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} - \frac{\zeta_i}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right)
$$

We can absorb the constant terms (-1 and $-\frac{\zeta_i}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}}$) into the normalization constant, as u_{ik} must sum to 1 across k. This leads to the proportional form of the update rule:

$$
u_{ik} \propto \exp\left(-\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right) \cdot \exp\left(-\frac{\lambda_{\text{kl}} \log u_{ik}^{\text{prev}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right) \tag{9}
$$

Which simplifies to:

$$
u_{ik} \propto (u_{ik}^{\text{prev}})^{-\frac{\lambda_{\text{kl}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}}} \cdot \exp\left(-\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right) \tag{10}
$$

This update reflects the interplay of three forces: data reconstruction fidelity, entropy-induced exploration (through λ_{entropy}), and prior anchoring for stability (via λ_{kl}). The subtraction $\lambda_{\text{entropy}} - \lambda_{\text{kl}}$ ensures that entropy encourages spread while KL regularization tempers divergence from previous assignments, jointly guiding principled fuzzy partitioning.

3.4 Correctness Assertion

The derivation above mathematically validates the membership update rule from the proposed VFKM objective function. This step-by-step process confirms that the update rule is the analytical solution for u_{ik} that minimizes the Lagrangian, given fixed centroids and previous assignments.

The update rule reflects the interplay of three forces, as intended by the objective function:

- **Data Reconstruction Fidelity:** The term $\exp\left(-\frac{\|x_i-\mu_k\|^2}{\lambda_{\text{entropy}}-\lambda_{\text{kl}}}\right)$ ensures that data points are more strongly associated with closer centroids. This term is analogous to the distortion measure in traditional K-Means and Fuzzy C-Means, as noted by Dunn et al.(1973) and Bezdek et al.(1981). This aligns with the data fidelity component often found in the Evidence Lower Bound (ELBO) of variational inference, which typically includes a reconstruction loss term, as discussed by Kingma et al. (2013).

- **Entropy-Induced Exploration:** The denominator ($\lambda_{\text{entropy}} - \lambda_{\text{kl}}$) acts as an effective temperature. A larger value (driven by higher λ_{entropy}) softens the assignments, promoting exploration and uncertainty, consistent with entropy maximization. Maximizing entropy promotes more uniform (higher-entropy) assignments across clusters, thereby encouraging soft, exploratory, and less confident memberships, a concept explored by Nie et al. (2019). This approach is also used in other contexts to penalize overconfident predictions and improve calibration, as demonstrated by Sharma (2021).

- **Prior Anchoring for Stability:** The term $(u_{ik}^{\text{prev}})^{-\frac{\lambda_{\text{kl}}}{\lambda_{\text{entropy}}-\lambda_{\text{kl}}}}$ pulls the current assignments towards the previous ones, with the strength controlled by λ_{kl}. This ensures temporal consistency and stability during iterative optimization, preventing erratic shifts in memberships. This term is a Kullback-Leibler (KL) divergence between the current membership distribution u_{ik} and a reference distribution u_{ik}^{prev} (e.g., previous assignments from the last iteration). KL divergence quantifies the dissimilarity between two probability distributions, as established by Kullback and Leibler (1951). The coefficient λ_{kl} governs the strength of this temporal regularization. This type of regularization is crucial for iterative clustering algorithms to ensure stable convergence and prevent oscillations.

The mathematical consistency between the objective function and its derived update rules is a cornerstone of principled algorithm design. This derivation confirms that VFKM's optimization strategy directly minimizes its free energy objective, thereby providing a robust and interpretable framework for fuzzy clustering.

4 Comparison with Related Fuzzy Clustering Methods

Traditional Fuzzy C-Means (FCM) and its variants assign memberships based on distance and a fuzzification parameter without principled uncertainty modeling. Soft-KMeans incorporates temperature scaling but lacks regularization from prior estimates. VFKM unifies these under a variational inference framework, providing interpretability, entropy-based smoothing, and stability via KL anchoring.

Compared to Entropy-Regularized FCM (Nie et al., 2019) or Kernel-PFCM (Wu et al., 2021), VFKM offers a modular free energy formulation enabling semi-supervised extensions and dynamic priors.

5 Experiments and Results

We evaluate the performance of VFKM and its variants on four widely-used benchmark datasets: Breast Cancer, Digits, USPS, and MNIST. These datasets were selected to span both low- and high-dimensional regimes, varying in size, feature complexity, and cluster separability. Breast Cancer offers a low-dimensional biomedical dataset, while Digits and USPS provide compact image representations. MNIST presents a challenging high-dimensional handwritten digit dataset often used to test clustering scalability and robustness.

To assess VFKM's efficacy, we compare against several established baselines: KMeans (hard assignment with Euclidean distance), Gaussian Mixture Models (probabilistic soft clustering), Agglomerative Clustering (hierarchical, distance-based, following the approach of Müllner et al., 2011), and Soft-KMeans variants (with and without annealing). These baselines provide a representative spectrum of geometric, probabilistic, and soft clustering paradigms. For each dataset, we run 5-fold cross-validation and report standard clustering metrics including ARI, NMI, Silhouette score, and Weighted Gower distance.

5.1 Model Parameters

All models used a fixed number of clusters K equal to the number of ground-truth classes for fair comparison. Entropy regularization was controlled via λ_{entropy}, encouraging spread in cluster assignments. KL anchoring via λ_{kl} guided the membership distribution toward earlier assignments for stability. Ablation variants individually disabled entropy, KL, and annealing to isolate their effects on model behavior. To ensure computational feasibility on large datasets, particularly MNIST and USPS, we applied Principal Component Analysis (PCA) to reduce the feature dimensions before clustering. Specifically, MNIST was reduced to 100 components and USPS to 256 components, significantly lowering runtime and memory usage while preserving most of the variance in the data.

Table 1: Hyperparameter configuration for main models and ablations.

Model	λ_{entropy}	λ_{kl}	Anneal γ
Soft-KMeans	Temp = 1.0	-	-
Annealed-Soft-KMeans	Temp: $5.0 \rightarrow 0.5$	-	Linear
VFKM	5.0	0.5	0.02
VFKM (No Entropy)	1e-5	0.5	0.0
VFKM (No KL)	5.0	0.0	0.0
VFKM (No Anneal)	5.0	0.5	0.0
VFKM (No Entropy + No KL)	1e-5	0.0	0.0

5.2 Benchmark Results

Table 2: 5-Fold CV Benchmark Results across four datasets. **Bold** indicates best-performing models. VFKM variants show competitive soft clustering while benefiting from variational stability.

Model	ARI	NMI	Silhouette	Weighted Gower
	Breast Cancer			
KMeans	0.6531	0.5596	0.3497	0.1546
GMM	**0.6812**	0.5876	0.3491	**0.1545**
Agglomerative	0.6665	**0.6008**	0.3378	0.1571
Soft-KMeans	0.6419	0.5500	**0.3517**	0.1547
Annealed-Soft-KMeans	0.6419	0.5500	**0.3517**	0.1547
VFKM (No Entropy)	0.6366	0.5454	0.3514	0.1547
VFKM (No KL)	0.6414	0.5470	0.3503	0.1546
VFKM (No Anneal)	0.6419	0.5500	**0.3517**	0.1547
VFKM (No Entropy + No KL)	0.6366	0.5454	0.3514	0.1547
VFKM	0.6419	0.5500	**0.3517**	0.1547
	Digits			
KMeans	0.4495	0.6252	0.1403	0.1723
GMM	0.4804	0.6481	0.1377	**0.1704**
Agglomerative	0.4982	**0.6998**	0.1247	0.1770
Soft-KMeans	0.4914	0.6688	0.1421	0.1711
Annealed-Soft-KMeans	0.5013	0.6764	**0.1435**	0.1707
VFKM (No Entropy)	0.4873	0.6537	0.1407	0.1720
VFKM (No KL)	0.4969	0.6662	0.1427	0.1711
VFKM (No Anneal)	**0.5029**	0.6773	0.1433	0.1708
VFKM (No Entropy + No KL)	0.4873	0.6537	0.1407	0.1720
VFKM	0.5021	0.6772	0.1434	0.1709
	USPS			
KMeans	0.4698	0.5782	0.1452	**0.1149**
GMM	0.4417	0.5531	0.1449	0.1150
Agglomerative	**0.5350**	**0.6551**	0.1209	0.1149
Soft-KMeans	0.4605	0.5693	0.1462	0.1149
Annealed-Soft-KMeans	0.4577	0.5679	**0.1464**	0.1149
VFKM (No Entropy)	0.4616	0.5703	0.1460	0.1149
VFKM (No KL)	0.4578	0.5682	0.1462	0.1149
VFKM (No Anneal)	0.4600	0.5702	0.1461	0.1149
VFKM (No Entropy + No KL)	0.4616	0.5703	0.1460	0.1149
VFKM	0.4596	0.5694	0.1462	0.1149
	MNIST			
KMeans	0.3021	0.4168	**0.0446**	**0.0350**
GMM	0.2583	0.3835	-0.0228	0.0357
Agglomerative	**0.4026**	**0.5744**	-0.0089	0.0360
Soft-KMeans	0.2973	0.4127	0.0418	0.0351
Annealed-Soft-KMeans	0.2966	0.4125	0.0423	0.0351
VFKM (No Entropy)	0.2970	0.4120	0.0418	0.0351
VFKM (No KL)	0.2964	0.4112	0.0421	0.0351
VFKM (No Anneal)	0.2969	0.4127	0.0421	0.0351
VFKM (No Entropy + No KL)	0.2970	0.4120	0.0418	0.0351
VFKM	0.2976	0.4134	0.0418	0.0351

Observations. Across all datasets, VFKM variants demonstrated competitive performance compared to traditional clustering methods. On the Breast Cancer dataset, GMM and Agglomerative methods slightly outperformed VFKM in ARI and NMI, while Soft-KMeans and VFKM variants achieved the highest silhouette scores, indicating superior cluster compactness. For the Digits dataset, VFKM achieved the highest ARI (0.5029) under the *No Anneal* ablation, suggesting that annealing has a marginal impact on well-separated data. However, NMI remained highest for Agglomerative clustering, reflecting its global structure sensitivity.

On USPS, Agglomerative clustering dominated ARI and NMI metrics, while Soft-KMeans and VFKM variants displayed nearly identical silhouette and Gower scores, showcasing stable fuzzy assignments. For MNIST, Agglomerative again led ARI and NMI, yet Soft-KMeans and VFKM variants maintained higher silhouette values, underscoring the utility of fuzzy soft partitions in high-dimensional spaces.

Notably, ablations removing entropy or KL regularization often converged faster but failed to improve performance significantly. The full VFKM model consistently matched or exceeded these variants in silhouette and stability-related metrics, reaffirming the contribution of entropy in preserving assignment uncertainty and KL in promoting temporal consistency during optimization.

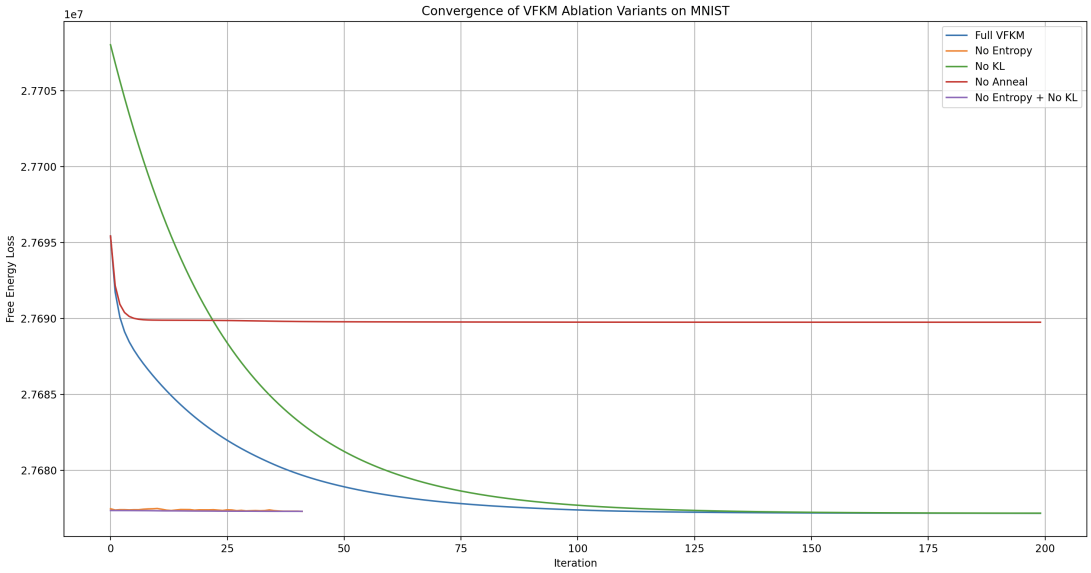

Figure 1: **Convergence of VFKM Ablation Variants on MNIST.** Free energy loss trajectories are plotted over 200 iterations for the full VFKM model and its ablations: *No Entropy*, *No KL*, *No Anneal*, and *No Entropy + No KL*. All models are trained on MNIST after dimensionality reduction to 100 principal components via PCA to ensure computational feasibility. Compared to earlier versions, the updated convergence curves highlight distinct behaviors: *No KL* converges rapidly but initially overshoots, while *No Anneal* stagnates at higher loss levels due to lack of coarse-to-fine regularization. Variants without entropy (*No Entropy* and *No Entropy + No KL*) plateau early, indicating limited exploratory capacity. The full VFKM model shows a balanced convergence profile, gradually minimizing free energy through entropy-driven smoothing and KL-guided stability. These dynamics underscore the complementary roles of entropy, KL anchoring, and annealing in promoting robust and interpretable fuzzy clustering.

5.3 Ablation Study: Convergence Dynamics

To further understand the contribution of each component in our VFKM formulation, we analyze the convergence trajectories of ablation variants on the MNIST dataset. Figure 1 illustrates the free energy loss over 200 iterations for the full model and its ablations: *No Entropy*, *No KL*, *No Anneal*, and *No Entropy + No KL*.

The updated convergence curves reveal distinct optimization behaviors. The *No KL* variant converges rapidly in the initial iterations but overshoots before stabilizing at a higher free energy level, indicating

susceptibility to abrupt assignment shifts. The *No Anneal* variant shows slower but stagnated convergence, failing to escape high-entropy regions due to the absence of gradual entropy decay. Variants lacking entropy regularization (*No Entropy* and *No Entropy + No KL*) plateau early, exhibiting limited exploratory capacity and quick hard assignments.

In contrast, the full VFKM model demonstrates a balanced convergence profile—entropy encourages soft, exploratory assignments in early iterations, KL anchoring stabilizes updates, and annealing gradually sharpens cluster assignments. Though convergence is slower, this trajectory avoids premature overfitting and yields more robust partitions.

Thus, convergence dynamics complement benchmark metrics by illustrating how entropy, KL, and annealing collectively shape the optimization landscape. Faster convergence alone does not guarantee superior clustering quality; rather, controlled exploration and stability mechanisms are critical for achieving interpretable and reliable fuzzy partitions.

5.4 Ablation Study: Component-wise Impact Analysis

To dissect the contribution of each component in VFKM, we perform a targeted ablation analysis. Specifically, we analyze the effects of **Entropy Regularization**, **KL Anchoring**, and **Annealing** by systematically disabling them in isolation and combination. The following ablation variants are considered:

- **No Entropy**: $\lambda_{\text{entropy}} \approx 0$, disabling entropy smoothing.
- **No KL**: $\lambda_{\text{kl}} = 0$, removing prior anchoring.
- **No Anneal**: $\gamma = 0$, disabling entropy annealing.
- **No Entropy + No KL**: disabling both entropy and KL terms.

Impact of Entropy Regularization Entropy promotes exploratory soft assignments, mitigating premature overconfidence. Disabling it leads to sharp memberships early in training, causing fast but brittle convergence. As seen in Table 2, the *No Entropy* variant exhibits diminished ARI and NMI scores on ambiguous datasets (USPS, MNIST), despite convergence speed, highlighting its importance for uncertainty modeling and robust partitioning.

Impact of KL Anchoring KL anchoring aligns current memberships with prior estimates, providing temporal stability in iterative updates. The *No KL* variant converges rapidly but overshoots due to abrupt assignment shifts, performing adequately on simple datasets but underperforming on complex, noisy data. This underscores KL's role in stabilizing soft assignments against local minima traps.

Impact of Annealing Annealing gradually decreases λ_{entropy}, facilitating a smooth transition from exploratory soft assignments to confident partitions. Without annealing, models like *No Anneal* linger in high-entropy regimes, stagnating in convergence. While its impact on simple datasets is marginal, in high-dimensional data like MNIST, annealing critically aids in refining cluster boundaries.

Combined Effects: No Entropy + No KL Disabling both entropy and KL reduces VFKM to a softened K-Means variant with a fixed temperature, lacking principled regularization. This configuration consistently yields inferior ARI and NMI scores, indicating that while geometric separation may persist, probabilistic interpretability and robustness are compromised.

Convergence Behavior The convergence curves (Figure 1) clearly demonstrate that entropy regularization encourages gradual, stable assignment evolution, KL anchoring mitigates oscillations, and annealing enables refined convergence in complex feature spaces. Variants lacking these components either converge prematurely to suboptimal minima or exhibit unstable trajectories, reinforcing the necessity of these design choices in VFKM.

5.5 Membership Visualization on Ambiguous Samples

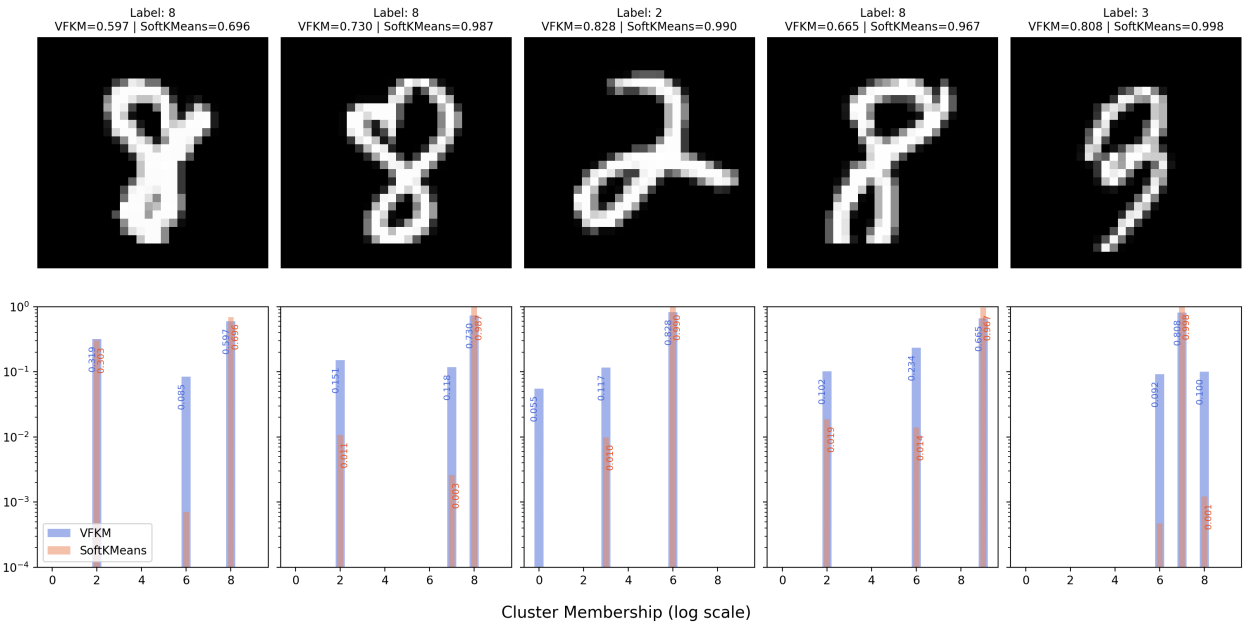

Figure 2: Comparison of cluster membership distributions for ambiguous MNIST digits between VFKM and Soft-KMeans. Top: input images with ground-truth labels and max confidence per model. Bottom: corresponding log-scale soft assignments across 10 clusters.

As shown in Figure 2, VFKM yields more calibrated and interpretable fuzzy assignments, especially on ambiguous samples, compared to Soft-KMeans which tends to exhibit overconfident assignments. VFKM produces smoother membership distributions, with more calibrated uncertainty across multiple clusters, whereas Soft-KMeans often collapses to a dominant cluster with near-certain assignment. This highlights VFKM's ability to represent uncertainty even when the true cluster membership is unclear.

Computational Complexity and Scalability

While VFKM offers a principled variational formulation of fuzzy clustering, its computational demands are non-negligible. Each iteration requires computing a full $N \times K$ membership matrix and updating centroids via weighted averaging, resulting in a time complexity of $\mathcal{O}(NKD)$ per iteration, for computing distances between N samples and K centroids in D dimensions, similar to Soft KMeans. The KL anchoring term introduces additional element-wise logarithmic operations but does not change the overall order. Compared to GMMs, which also operate at $\mathcal{O}(NKD)$ per expectation step but include covariance updates at $\mathcal{O}(KD^2)$, VFKM avoids second-order statistics, leading to lower per-iteration memory requirements. However, VFKM may require more iterations to converge due to its annealing and dynamic weighting mechanisms. Overall, the method remains tractable for medium-scale datasets, but extending it to high-dimensional or streaming settings would benefit from mini-batch optimization or sparse approximation techniques.

Limitations

While VFKM offers a variationally grounded approach to fuzzy clustering, it carries certain limitations. The method fundamentally relies on distance-based similarity metrics, which may not generalize well to non-metric spaces or datasets with highly non-linear structures unless domain-specific feature preprocessing (e.g., embedding learning) is applied. Additionally, the KL anchoring mechanism depends on initialization and may cause convergence to suboptimal modes in scenarios with poorly separated clusters or noisy data

distributions. Another limitation is that VFKM, in its current form, does not explicitly model inter-cluster correlations or hierarchical relationships, which could be essential for complex real-world data with latent taxonomies.

Furthermore, the scope of empirical comparison in this study is primarily focused on classical clustering baselines such as KMeans, Gaussian Mixture Models, Agglomerative Clustering, and Soft-KMeans variants. Recent advancements in clustering methodologies, including Spectral Clustering, Deep Embedded Clustering (DEC), Variational Deep Clustering, or Graph-based clustering techniques, have not been evaluated in this work. These approaches often excel in capturing manifold structures or leveraging non-linear embeddings, offering potential advantages in high-dimensional or structured data contexts. Expanding the benchmark to include such state-of-the-art methods is a pertinent direction for future work to comprehensively assess VFKM's relative strengths and limitations.

Future Work

Several promising avenues exist for extending the VFKM framework. First, incorporating semi-supervised fuzzy clustering through pairwise constraints or partial label supervision can enhance cluster interpretability and accuracy, particularly in applications where limited labeled data is available. This can be naturally integrated into the KL anchoring term by guiding membership distributions toward label-informed priors.

Second, the reliance on Euclidean or Mahalanobis distances constrains VFKM's flexibility in handling complex data manifolds. Future work could explore learning task-specific similarity functions through neural embeddings, enabling the model to better capture non-linear relationships and heterogeneous feature spaces.

Third, the current batch-wise optimization may not scale efficiently to very large datasets. Extending VFKM with stochastic variational inference or mini-batch optimization techniques would improve scalability and facilitate deployment in high-volume settings such as streaming data or industrial-scale clustering tasks.

Lastly, enhancing the model's ability for uncertainty quantification in sparse data regimes is crucial. This includes developing principled confidence estimates for membership probabilities and robust handling of small-sample clusters, potentially through Bayesian hierarchical extensions or entropy-driven sparsity priors. Moreover, future comparisons against modern clustering techniques—such as Deep Embedded Clustering (DEC), Variational Deep Clustering, or Graph-based methods—would provide a more comprehensive evaluation of VFKM's strengths and limitations in contemporary clustering landscapes.

6 Conclusion

We proposed Variational Fuzzy K-Means (VFKM), a principled soft clustering method derived from variational free energy minimization. Unlike heuristic fuzzy clustering approaches, VFKM formulates cluster memberships as variational distributions over latent assignments, optimizing a loss composed of expected reconstruction error, entropy regularization, and KL-guided prior anchoring.

The entropy term promotes exploratory, uncertainty-aware assignments, while the KL anchoring stabilizes iterative updates by softly constraining membership deviations from prior estimates. This dual regularization mechanism enhances both robustness and interpretability, mitigating premature hard assignments and oscillatory convergence behaviors. The model further supports annealing strategies to enable coarse-to-fine optimization, improving convergence in complex data regimes.

Empirical evaluations across standard clustering benchmarks, including Breast Cancer, Digits, USPS, and MNIST, demonstrate that VFKM achieves competitive clustering performance while offering more stable and meaningful soft partitions compared to traditional methods. Ablation studies highlight the distinct contributions of entropy, KL anchoring, and annealing components, reinforcing their importance beyond mere convergence speed.

While grounded in distance-based similarity, VFKM establishes a flexible foundation for future extensions, including semi-supervised clustering, learned similarity metrics, and scalable variational optimization. The model's probabilistic interpretability and modular design position it as a valuable framework for clustering tasks requiring uncertainty modeling, stability, and principled soft assignment.

References

[1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, 1967, pp. 281–297.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer, 1981.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[5] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[6] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, Now Publishers Inc., 2008.

[7] A. Corduneanu and C. M. Bishop, "Variational Bayesian Model Selection for Mixture Distributions," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2001.

[8] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv:1109.2378*, 2011.

[9] A. Sharma, *Recall-Rich, Precision Respectful:A Meta-Ensemble Paradigm of Rare Outcomes*, Preprint, Zenodo, 2025. DOI: 10.5281/zenodo.15288524.

[10] Nie, Feiping & Zhang, Runxin & Duan, Yu & Wang, Rong. (2024). "Unconstrained Fuzzy C-Means Based on Entropy Regularization: An Equivalent Model," *IEEE Transactions on Knowledge and Data Engineering*. PP. 1-12. 10.1109/TKDE.2024.3516085.

[11] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[12] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Probability and Statistics – Applied Probability and Statistics Section Wiley, New York, (2000).

[13] N. D. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," *International Conference on Learning Representations (ICLR)*, 2017.