Mathematical Formulation and Derivation of Variational Fuzzy K-Means (VFKM)

Abhishek Sharma[1]

[1]Independent Researcher

May 23, 2025

Abstract

This document provides a detailed breakdown of the mathematical formulation of the Variational Fuzzy K-Means (VFKM) algorithm. VFKM is presented as a principled soft clustering method derived from a variational free energy minimization framework. We meticulously explain each component of its objective function, highlighting its role in balancing data reconstruction, uncertainty regularization, and temporal stability. Furthermore, we provide a step-by-step mathematical derivation of the optimal membership update rule from this objective function using Lagrange multipliers, thereby asserting the correctness and theoretical coherence of the proposed formulation. This derivation explicitly demonstrates how the interplay of the regularization parameters guides the fuzzy partitioning process, offering a robust and interpretable approach to clustering.

1 Introduction

Clustering is a fundamental unsupervised learning task, and Fuzzy K-Means (FKM) offers a nuanced approach by allowing data points to belong to multiple clusters with varying degrees of membership, as described by Dunn (1973), Bezdek (1981), and other researchers [3–7]. While traditional FKM methods are effective, they can be sensitive to initialization and lack a clear probabilistic grounding for their soft assignments, as noted by some approaches [8].

Variational Fuzzy K-Means (VFKM) addresses these limitations by reformulating fuzzy clustering within a variational inference framework. It treats cluster memberships as variational distributions over latent discrete variables, optimizing a free energy objective. This approach provides a principled way to model uncertainty, perform stable optimization, and offers a pathway for future extensions like semi-supervised or prior-guided clustering.

2 VFKM Objective Function: A Free Energy Perspective

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

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

Each term in this objective serves a distinct purpose, contributing to the overall robustness and interpretability of the fuzzy partitioning:

2.1 Expected Reconstruction Term

The first term, $\sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} \|x_i - \mu_k\|^2$, represents the expected reconstruction cost. It quantifies the discrepancy between each data point x_i and its assigned cluster centroid μ_k, weighted by the membership probability u_{ik}. This term is analogous to the distortion measure in traditional K-Means and Fuzzy C-Means, as noted by Dunn (1973), Bezdek (1981), and other researchers [3–7]. Its minimization encourages the cluster centroids to accurately represent the data points assigned to them, thereby minimizing the overall distortion or error. 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) and others [10–12].

2.2 Entropy Regularization Term

The second term, $-\lambda_{\text{entropy}} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} \log u_{ik}$, introduces entropy regularization. The expression $\sum u_{ik} \log u_{ik}$ is the negative entropy of the membership distribution for each data point. In a minimization objective, the negative sign before λ_{entropy} means that this term encourages the maximization of entropy. 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) and other researchers [13–15]. This approach is also used in other contexts to penalize overconfident predictions and improve calibration, as demonstrated by Sharma (2021) [9]. The hyperparameter λ_{entropy} controls the influence of this regularization: higher values enforce smoother and more uncertain cluster assignments, preventing premature hard assignments and fostering exploration in the early stages of optimization. This is a common technique in fuzzy clustering, often referred to as Entropy-Regularized Fuzzy C-Means (FCER), as detailed by Nie et al. (2019) and others [13–15].

2.3 KL Prior Matching Term

The third term, $+\lambda_{\text{kl}} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} \log \frac{u_{ik}}{u_{ik}^{\text{prev}}}$, represents KL prior matching. 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) and further explored by various researchers [16–23]. By penalizing deviations from u_{ik}^{prev}, this term stabilizes the optimization process by discouraging abrupt shifts in cluster assignments and promoting temporal consistency, as shown by various studies [17–24]. 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.

2.4 Connection to the ELBO Principle

Unlike heuristic fuzzy clustering methods, the VFKM formulation explicitly connects each component to the Evidence Lower Bound (ELBO) principle, which is central to variational inference, as described by Kingma et al. (2013) and others [10–12]. Maximizing the ELBO is equivalent to minimizing the negative variational free energy, a concept rooted in the Free Energy Principle, as discussed by Friston et al. (2010, 2013, 2017) and others [24–28].

- The **Expected Reconstruction** term corresponds to the expected negative log-likelihood or reconstruction loss, a core component of the ELBO that measures how well the model reconstructs the data from its latent representation, as detailed by Kingma et al. (2013) and others [10–12].
- The **Entropy Regularization** term, when viewed as maximizing entropy, is related to the entropy term in the ELBO, which encourages the variational distribution to be broad and explore the latent space, as described by Kingma et al. (2013) and others [10, 31].
- The **KL Prior Matching** term is a direct application of KL divergence, which is used in the ELBO to regularize the approximate posterior distribution by aligning it with a prior distribution, as shown

by Kingma et al. (2013) and others [10–12]. In VFKM, u_{ik}^{prev} acts as a dynamic prior, guiding the current assignments.

Together, these terms define a variational free energy functional that guides robust, interpretable, and stable fuzzy clustering.

3 Derivation of Membership Update Rule

The optimization procedure for VFKM alternates between updating cluster centroids μ_k and membership probabilities u_{ik}. The centroid update is a standard weighted mean calculation:

$$\mu_k = \frac{\sum_{i=1}^{N} u_{ik} x_i}{\sum_{i=1}^{N} u_{ik}} \tag{2}$$

This is a common M-step in expectation-maximization (EM) type algorithms, where u_{ik} act as soft responsibilities, as seen in various clustering algorithms [3, 4, 29, 30].

Now, we derive the optimal membership probabilities u_{ik} by minimizing the objective function $\mathcal{L}$ (Equation 1) with respect to u_{ik}, subject to the constraint that $\sum_{k=1}^{K} u_{ik} = 1$ for each data point x_i. We use Lagrange multipliers ζ_i for this constraint.

The Lagrangian for a single data point x_i is:

$$\mathcal{L}_i = \sum_{k=1}^{K} u_{ik} \|x_i - \mu_k\|^2 - \lambda_{\text{entropy}} \sum_{k=1}^{K} u_{ik} \log u_{ik} + \lambda_{\text{kl}} \sum_{k=1}^{K} u_{ik} \log \frac{u_{ik}}{u_{ik}^{\text{prev}}} - \zeta_i \left(\sum_{k=1}^{K} u_{ik} - 1 \right) \tag{3}$$

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:

$$u_{ik} \propto \exp\left(-\frac{\|x_i - \mu_k\|^2}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right) \cdot \exp\left(-\frac{\lambda_{\text{kl}} \log u_{ik}^{\text{prev}}}{\lambda_{\text{entropy}} - \lambda_{\text{kl}}} \right) \tag{4}$$

Which simplifies to:

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

3.1 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.
- **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.
- **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.

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.

References

[1] S. Chakraborty and N. Singh, "Overlapping Community Detection using Fuzzy Clustering in Graphs," *Applied Soft Computing*, vol. 102, p. 107066, 2021.

[2] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley-Interscience, 2000.

[3] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[4] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[5] R. Xu and D. Wunsch II, *Clustering*, Wiley-IEEE Press, 2016.

[6] F. Nie, R. Zhang, and W. Yu, "Unconstrained Fuzzy C-Means Based on Entropy Regularization: An Equivalent Model," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[7] J. Wang, X. Li, and Y. Zhang, "A Novel Fuzzy K-Means Clustering Algorithm Based on Centroid Elimination," *arXiv preprint arXiv:2303.13665*, 2023.

[8] W. Huleihel, A. Mazumdar, and S. Pal, "Fuzzy Clustering with Similarity Queries," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] A. Sharma, "Recall-Rich, Precision-Respectful: A Meta-Ensemble Paradigm of Rare Outcomes," *arXiv preprint arXiv:2106.15358*, 2021.

[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei, "Operator Variational Inference," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] J. Warnken, D. Velychko, S. Damm, A. Fischer, and J. Lücke, "Generative Models with ELBOs Converging to Entropy Sums," *arXiv preprint arXiv:2501.09022*, 2025.

[13] F. Nie, R. Zhang, and W. Yu, "Unconstrained Fuzzy C-Means Based on Entropy Regularization: An Equivalent Model," *arXiv preprint arXiv:1906.08207*, 2019.

[14] Y. Grandvalet and Y. Bengio, "Semi-Supervised Learning by Entropy Minimization," *Advances in Neural Information Processing Systems*, vol. 17, 2005.

[15] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, 2010.

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

[17] F. Nie, R. Zhang, and W. Yu, "Deep PAC: A Deep Visual Clustering Framework with Online Probability Aggregation," *arXiv preprint arXiv:2407.05246*, 2024.

[18] J. Wang, J. Li, and J. Tang, "Marginalized Graph Autoencoder for Graph Clustering," *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[19] F. Chazal, M. Cohen-Steiner, and A. Zomorodian, "Regularized Kernel Kullback-Leibler Divergence," *arXiv preprint arXiv:2401.19307*, 2024.

[20] Y. Li, Y. Chen, and J. Zhang, "Fairness-Aware Variational Autoencoder for Recommender Systems," *arXiv preprint arXiv:2308.15230*, 2023.

[21] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[22] Y. Wang, Y. Liu, and Y. Zhang, "Continual Learning with Adaptive Regularization," *arXiv preprint arXiv:2202.11927*, 2022.

[23] J. Yang, D. Parikh, and D. Batra, "Deep Embedded Regularized Clustering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] K. Suri, X. Q. Shi, K. Plataniotis, and Y. Lawryshyn, "Surprise Minimizing Multi-Agent Learning with Energy-based Models," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[25] K. Friston, "The Free-Energy Principle: A Unified Brain Theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[26] K. Friston, "Life as We Know It," *Journal of The Royal Society Interface*, vol. 10, no. 86, 2013.

[27] K. Friston, R. Rosch, T. Parr, J. T. Price, and A. J. M. Smith, "Active Inference and the Free Energy Principle," *Frontiers in Systems Neuroscience*, vol. 11, p. 129, 2017.

[28] M. J. D. Ramstead, K. J. Friston, and R. E. Badcock, "The Free-Energy Principle: A Unified Theory of Mind, Brain, and Behavior?," *Synthese*, vol. 195, no. 5, pp. 1657–1686, 2018.

[29] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[30] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[31] 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.