A Unified MDL-based Binning and Tensor Factorization Framework for PDF Estimation

Mustafa Musab[1], *Joseph K. Chege*[1], *Arie Yeredor*[2], *and Martin Haardt*[1]
[1]Communications Research Laboratory, Ilmenau University of Technology, Ilmenau, Germany
[2]School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel
Email: {mustafa.musab, joseph.chege, martin.haardt}@tu-ilmenau.de, ariey@tauex.tau.ac.il

***Abstract*—Reliable density estimation is fundamental for numerous applications in statistics and machine learning. In many practical scenarios, data are best modeled as mixtures of component densities that capture complex and multimodal patterns. However, conventional density estimators based on uniform histograms often fail to capture local variations, especially when the underlying distribution is highly nonuniform. Furthermore, the inherent discontinuity of histograms poses challenges for tasks requiring smooth derivatives, such as gradient-based optimization, clustering, and nonparametric discriminant analysis. In this work, we present a novel non-parametric approach for multivariate probability density function (PDF) estimation that utilizes minimum description length (MDL)-based binning with quantile cuts. Our approach builds upon tensor factorization techniques, leveraging the canonical polyadic decomposition (CPD) of a joint probability tensor. We demonstrate the effectiveness of our method on synthetic data and a challenging real dry bean classification dataset.**

***Index Terms*—Probability density function (PDF), minimum description length (MDL), tensor factorization, quantile-based binning, nonparametric density estimation.**

I. Introduction

Accurate density estimation is crucial in many areas of statistics and machine learning, including clustering, classification, and signal processing. In particular, mixture models, where the overall distribution is expressed as a weighted sum of component densities, play a key role in modeling complex, multimodal data. Recovering both the component densities and their corresponding weights is essential for revealing the underlying distribution and making reliable inferences.

Although many natural processes are inherently continuous, conventional methods, such as uniform histograms or probability mass functions (PMFs), often fail to capture local variations in the data. This is especially true in applications involving complex and multimodal distributions. Examples include biological measurements (e.g., gene expression levels in cancer genomics, and multimodal intensity distributions in medical imaging like the BrainWeb MRI dataset [1]) and agricultural applications.

Recent advances in tensor-based density estimation have attracted considerable attention in the statistical community (e.g., Anandkumar *et al.* [2]; Miranda *et al.* [3]; Gottesman *et al.* [4]). In all these works, the focus is on fully parametric models in which the mixture components are assumed to belong to a specific parametric family. In contrast, our method makes no parametric assumption about the underlying density.

Our approach is most similar to that of Kargas *et al.* [5], who proposed a tensor–based method for learning mixtures. In their work, the dataset is discretized using uniform bins, followed by tensor factorization to recover the discretized PDFs, and finally, sinc interpolation is used to reconstruct the continuous PDF. However, uniform bins have been shown to be effective only when the data are approximately uniformly distributed [6]. In contrast, by employing minimum description length (MDL)-based binning with quantile cuts, our method overcomes these limitations.

Much of the existing literature within the MDL framework has focused on the estimation of histogram density, producing discrete PMF models (e.g., [7]). Although such methods yield an accurate discrete representation, they do not directly address PDF estimation. In this work, we extend the classical MDL framework to continuous multivariate settings. Moreover, while previous work [8] represents the joint PDF using a low-rank tensor in the Fourier domain, our method operates directly in the data domain.

II. Problem Formulation

Consider a collection of N continuous random variables $\mathsf{X} = \{X_1, \ldots, X_N\}$. Furthermore, given a realization $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ of X, assume that the joint PDF $f_{\mathsf{X}}(\boldsymbol{x})$ can be written as a weighted sum of R [1] conditional PDFs $f_{\mathsf{X}\,|\,H}$, and that each conditional PDF can be factorized into a product of its marginal densities such that

$$f_{\mathsf{X}}(x_1, \ldots, x_N) = \sum_{r=1}^{R} p_H(r) \prod_{n=1}^{N} f_{X_n|\,H}(x_n \,|\, r). \quad (1)$$

The expression in (1) represents $f_{\mathsf{X}}(\boldsymbol{x})$ as a mixture of product distributions, where H can be interpreted as a latent variable taking R states, while $p_H(r)$ is the prior probability of selecting the r-th product in the mixture (e.g., [5]).

However, note that no explicit parametric form (e.g., Gaussian, etc.) is specified for the conditional PDFs.

The support of each X_n is discretized into I_n bins $\Delta_n^{i_n}$, $i_n = 1, \ldots, I_n$, resulting in a discretized version of (1) which

The authors gratefully acknowledge the support of the German Research Foundation (DFG) under the PROMETHEUS project (reference no. HA 2239/16-1, project no. 462458843).

[1]The choice of the tensor rank R will be discussed in Subsection IV-B.

can be conveniently represented by an N-dimensional tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ where

$$\begin{aligned}\boldsymbol{\mathcal{X}}(i_1, \ldots, i_N) &= \mathsf{Pr}(X_1 \in \Delta_1^{i_1}, \ldots, X_N \in \Delta_N^{i_N}) \\ &= \sum_{r=1}^{R} p_H(r) \prod_{n=1}^{N} \mathsf{Pr}(X_n \in \Delta_n^{i_n} \,|\, H = r). \quad (2)\end{aligned}$$

By defining $\boldsymbol{A}_n(i_n, r) = \mathsf{Pr}(X_n \in \Delta_n^{i_n} \,|\, H = r)$ and $\lambda_r = p_H(r)$, it can be observed that $\boldsymbol{\mathcal{X}}$ admits a rank-R canonical polyadic decomposition (CPD) [9], [10] with factor matrices $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times R}$ and a "loading vector" $\boldsymbol{\lambda} \in \mathbb{R}^R$, subject to a set of probability simplex constraints, i.e.,

$$\begin{aligned}\boldsymbol{\mathcal{X}} &= \sum_{r=1}^{R} \lambda_r \boldsymbol{A}_1(:, r) \circ \boldsymbol{A}_2(:, r) \circ \cdots \circ \boldsymbol{A}_N(:, r) \\ &\text{subject to} \;\; \boldsymbol{\lambda} > \mathbf{0}, \;\; \mathbf{1}^\mathsf{T} \boldsymbol{\lambda} = 1 \\ &\qquad\quad \boldsymbol{A}_n \geq \mathbf{0}, \;\; \mathbf{1}^\mathsf{T} \boldsymbol{A}_n = \mathbf{1}^\mathsf{T}, \;\; n = 1, \ldots, N,\end{aligned} \quad (3)$$

where $\circ$, $(\cdot)^\mathsf{T}$, and $\mathbf{1}$ represent the outer product, the transpose operator, and an all-ones vector, respectively. It has been shown in [11] that (3) corresponds to a naïve Bayes model with a root variable H that takes a finite number R of states.

Given a dataset of T independent and identically distributed realizations $\boldsymbol{x}_t = \{x_{1,t}, \ldots, x_{N,t}\}$ $(t = 1, \ldots, T)$ of X, our objective is to recover the underlying continuous PDF $f_\mathsf{X}(\boldsymbol{x})$ from an estimate of the discretized PDF (joint PMF) $\boldsymbol{\mathcal{X}}$. Leveraging on the naïve Bayes structure of $\boldsymbol{\mathcal{X}}$, we propose to recover $f_\mathsf{X}(\boldsymbol{x})$ by interpolating each discretized marginal distribution $\boldsymbol{A}_n(:, r) = p(X_n \,|\, H = r)$ separately, followed by recombination of the interpolated marginal PDFs to form the joint PDF. A similar approach was considered in [5], where discretization was carried out using uniform bins, while sinc interpolation was adopted to recover the marginal PDFs. However, in the following section, we propose a PDF estimation method that employs nonuniform bins whose width and number are selected to minimize an MDL criterion. The resulting nonuniform bins necessitate the use of a different interpolation strategy. We present motivating examples highlighting the advantages of our approach over uniform binning.

III. Proposed Methodology

We propose a PDF estimation framework that addresses the limitations of uniform binning and sinc interpolation. Our approach proceeds in three main steps. We first employ an MDL-based strategy to learn the histogram of each marginal distribution. By determining both the number and locations of the bin edges in a data-driven manner, this step captures the inherent structure of the data. As a result, continuous variables are effectively transformed into categorical ones. Next, we recover the complete discretized PDF by applying a maximum likelihood PMF estimation algorithm (SQUAREM-PMF, [12]) within a coupled nonnegative tensor factorization framework. Finally, we employ spline interpolation to obtain a smooth PDF from the discretized joint PDF estimate. Although this description emphasizes PDF estimation, the same adaptive binning and smoothing procedure naturally extends to nonparametric mixture models.

A. MDL binning with quantile cuts

Many histogram density estimation methods rely on uniform binning, which can be suboptimal if the underlying distribution is strongly nonuniform or multimodal. Intuitively speaking, wider bins in regions with sparse data help to reduce noise from sampling randomness, whereas narrower bins in dense regions capture fine details more effectively. Therefore, adapting the bin widths and locations to the data can significantly improve estimation quality. In the MDL framework [7], the goal is to select the simplest possible model that sufficiently explains the observed data by determining both the optimal number of bins and their locations. This dual optimization is formalized via the *normalized maximum likelihood* (NML), which provides strong theoretical guarantees [2].

Let $\mathcal{M} = \{ f(\cdot \mid \theta) : \theta \in \Theta \}$ denote a histogram model class (for example, all histograms with K bins). Each $\theta \in \Theta$ represents a specific choice of bin edges, thereby defining a particular histogram within this class. We use $f(\boldsymbol{x} \mid \theta)$ to denote the density given specific parameters, and $f(\boldsymbol{x} \mid \theta, \mathcal{M})$ explicitly highlights the density given these parameters within the chosen histogram model class $\mathcal{M}$. For a univariate sample $\boldsymbol{x} = \{x_1, \ldots, x_T\} \subset X$ of length T, the maximum-likelihood estimate is $\hat{\theta}(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} f(\boldsymbol{x} \mid \theta)$.

The NML density [13] is defined by

$$f_{\text{NML}}(\boldsymbol{x} \mid \mathcal{M}) = \frac{f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x}), \mathcal{M})}{\mathcal{R}_\mathcal{M}}, \quad (4)$$

where the normalizing constant $\mathcal{R}_\mathcal{M}$, known as the *parametric complexity*, quantifies the intrinsic complexity of the model:

$$\mathcal{R}_\mathcal{M} = \int_{\boldsymbol{x} \in X} f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x}), \mathcal{M}) \, \mathrm{d}\boldsymbol{x}. \quad (5)$$

The *stochastic complexity* of $\boldsymbol{x}$ under $\mathcal{M}$ is then given by

$$\text{SC}(\boldsymbol{x} \mid \mathcal{M}) = -\log f_{\text{NML}}(\boldsymbol{x} \mid \mathcal{M}),$$

and the MDL principle prescribes selecting model $\mathcal{M}^\star$ and the corresponding $\hat{\theta}(\boldsymbol{x})$ that minimizes this quantity [15].

An essential step in constructing the MDL-optimal histogram is the definition of candidate cut points for the optimization process. The authors in [7] describe two approaches for selecting candidate cuts. The first, known as the midpoint approach, places a cut at the midpoint between each pair of consecutive data points. While this method ensures that every bin has at least one observation, it does not allow for empty bins, which is a disadvantage when large gaps are present. An alternative approach involves placing two cut points between each pair of consecutive data values, positioned as close as possible to the data values. Although this method improves adaptability to data gaps, it significantly enlarges the candidate-cut set, increasing the computational cost.

Unlike [7], we adopt a *quantile*-based strategy: candidate cuts are the empirical quantiles, giving equal-frequency bins that mirror the underlying density. This data-adaptive

[2]The NML criterion provides two important theoretical guarantees: (i) it uniquely solves Shtarkov's minimax problem [13], (ii) it also minimizes the expected worst-case regret in code length among all universal codes [14].

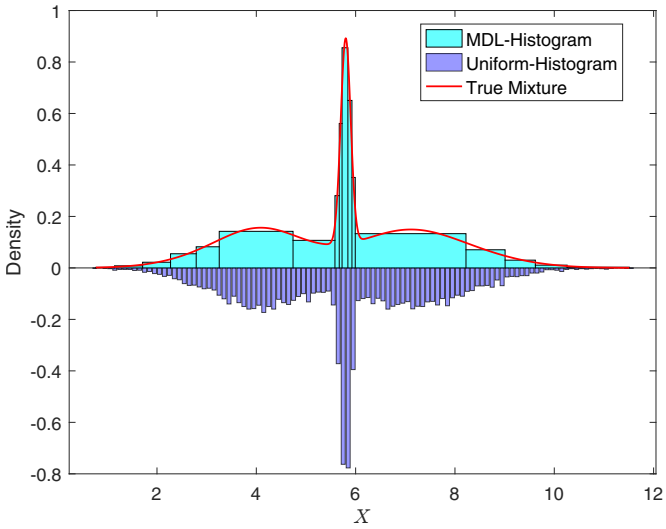

Fig. 1: The generating density, the MDL-optimal histogram (19 bins), and the uniform 100-bin histogram (mirrored).

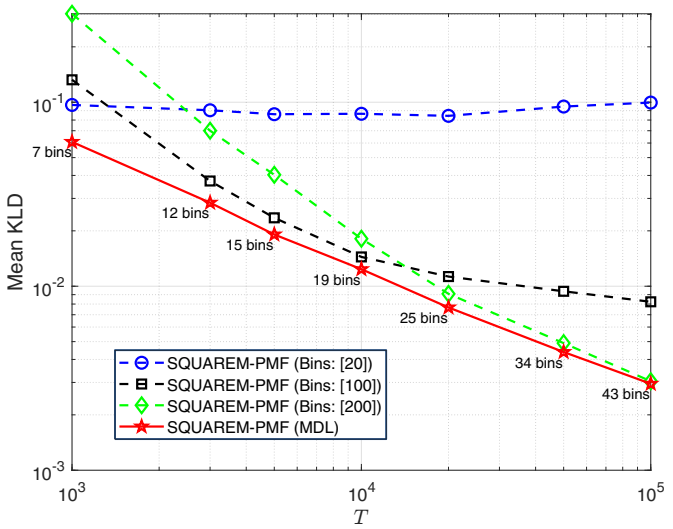

Fig. 2: KLD between the true density and its estimates obtained with the MDL histogram and uniform histograms of 20, 100, and 200 bins.

placement captures structure more faithfully while remaining computationally efficient. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(T)}$ denote an ordered sample of size T. The empirical cumulative distribution function (eCDF) is defined as:

$$\hat{F}_T(y) \;=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{x_{(t)} \leq y\},$$

where $\mathbb{1}(\cdot)$ is the indicator function. For a given probability $p \in [0,1]$, the empirical quantile $Q(p)$ is defined as the smallest y such that $\hat{F}_t(y) \geq p$. In our application, to partition the support of the data into E equal-frequency bins, we define the set of candidate interior cuts as:

$$\tilde{\mathcal{C}} = \left\{ c_j = Q\left(\frac{j}{E}\right) = \hat{F}_T^{-1}\left(\frac{j}{E}\right) \right\}_{j=1}^{E-1}$$

The goal is then to select a K-bin subset $C \subseteq \tilde{C}$ that minimizes the MDL criterion:

$$B(\boldsymbol{x} \mid E, K, C) = \mathrm{SC}(\boldsymbol{x} \mid C) \;+\; \log\binom{E}{K-1}, \qquad (6)$$

where $\mathrm{SC}(\boldsymbol{x} \mid C)$ is the stochastic complexity (negative log NML), measuring how well the model fits the observed data. The second term, $\log\binom{E}{K-1}$, represents the model complexity penalty. It measures the description length needed to specify which $K-1$ cut points are chosen from the E possible candidates. As detailed in [7], the SC is computed recursively, and the optimal cuts can be found by dynamic programming in $\mathcal{O}(E^2 \cdot K_{\max})$ time, where $K_{\max}$ is the maximum number of bins considered during optimization.

To demonstrate the advantage of the MDL histogram, we consider a toy example of five-component univariate Gaussian mixture. In our experiment, samples are drawn from the mixture depicted in Fig. 1. Fig. 2 shows the mean Kullback–Leibler divergence (KLD) (over 50 trials) between the true and estimated densities. Despite using significantly fewer bins, the MDL histogram consistently achieves a lower KLD compared to its uniform counterparts. As the sample size grows, the performance of the 100 and 200-bin uniform histograms gradually converges to that of the MDL. Interestingly, at a sample size of 10^4, the 19-bin MDL histogram significantly outperforms the uniform 20-bin, despite the nearly identical bin count. This highlights that the placement of the bins is as important as the bin count in capturing the underlying distribution.

B. Estimation of the Discretized PDF

We estimate the joint PMF (discretized PDF) by applying a low-rank factorization to the discretized data. Specifically, we adopt the SQUAREM-PMF algorithm proposed in [12], which extends the expectation-maximization (EM) algorithm proposed in [16] with a squared iterative methods (SQUAREM) acceleration step, thereby improving convergence speed even when the data are only partially observed. In this framework, the joint PMF tensor $\boldsymbol{\mathcal{X}}$ is constrained to have a CPD of rank R. The discretized observations are used in a maximum-likelihood (ML) setting [16]:

$$\begin{aligned} \min_{\{\boldsymbol{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \quad & -\sum_{t=1}^{T} \log\left(\sum_{r=1}^{R} \lambda_r \prod_{n=1}^{N} \boldsymbol{A}_n(x_{n,t}, r)\right) \\ \text{subject to} \quad & \boldsymbol{\lambda} > \mathbf{0}, \quad \mathbf{1}^\mathsf{T}\boldsymbol{\lambda} = 1 \\ & \boldsymbol{A}_n \geq \mathbf{0}, \quad \mathbf{1}^\mathsf{T}\boldsymbol{A}_n = \mathbf{1}^\mathsf{T}, \quad n = 1, \ldots, N \end{aligned} \qquad (7)$$

The EM updates for $\{\boldsymbol{A}_n\}$ and $\boldsymbol{\lambda}$ are derived from the objective (7). However, EM typically exhibits slow convergence.

SQUAREM-PMF addresses this limitation by squaring the standard EM fixed-point updates. At each iteration, it computes two successive EM steps, then performs a polynomial extrapolation in parameter space to accelerate convergence–while preserving the monotonic likelihood increase guaranteed by EM. Numerical evidence in [12] shows that SQUAREM-PMF can substantially reduce the iteration count and run

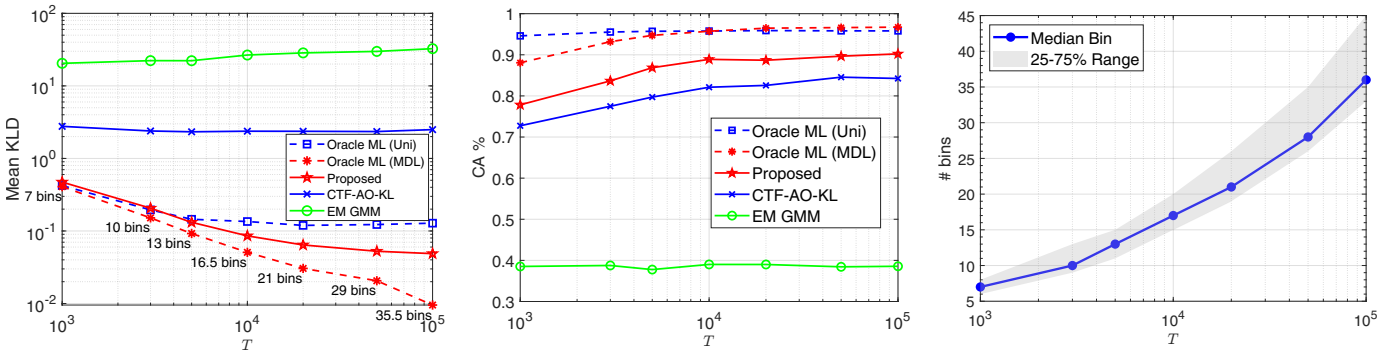

Fig. 3: Performance versus sample size: (a) KLD, (b) clustering accuracy (CA), and (c) median number of MDL-selected bins.

time compared to both plain EM and other factorization-based approaches, making it an efficient choice for the PMF estimation step in our proposed methodology.

C. PDF Reconstruction using Spline Interpolation

In the classical Shannon-sampling framework, a bandlimited signal can be perfectly reconstructed from uniform samples via sinc interpolation. In fact, Kargas *et al.* [5] demonstrated that a PDF which is (approximately) bandlimited with a cutoff frequency ω_c can be reconstructed from uniformly spaced samples of its corresponding CDF, provided the sampling interval $\Delta t \leq \frac{\pi}{\omega_c}$. However, extending this result to our settings is problematic because the MDL-estimated bins are nonuniform; consequently, the underlying assumption of uniform sampling (and hence a time-invariant reconstruction kernel) does not hold. Although generalized non-uniform sampling theory provides conditions under which sinc interpolation can be applied to nonuniform samples (see, e.g., [17]), these require that the average sampling rate satisfies the Nyquist criterion, something we cannot ensure a-priori with MDL-chosen cuts. This limitation motivates the use of cubic spline interpolation, which naturally accommodates nonuniform spacing and produces smooth, continuous curves with continuous first and second derivatives (i.e., the C^2 class) [18]. Let the r-th column $\boldsymbol{A}_n(:, r) \in \mathbb{R}^{I_n}$ of the n-th factor matrix $\boldsymbol{A}_n$ represent the discretized conditional PDF $p_{X_n \mid H}(x_n \mid r)$ of the n-th random variable. The corresponding CDF points are obtained via the cumulative sum of that column, yielding $\{F_i\}_{i=1}^{I_n}$. For each bin $\Delta^i = [i, i+1]$, a cubic polynomial is fitted:

$$S_i(x) = a_i + b_i(x - i) + c_i(x - i)^2 + d_i(x - i)^3. \quad (8)$$

This yields a piecewise definition of the continuous CDF

$$F_{X|H}(x_n \mid r) = \sum_{i=1}^{I_n - 1} S_i(x_n) \cdot \mathbb{1}_{[i,i+1)}(x_n), \quad (9)$$

where

$$\mathbb{1}_{[i,i+1)}(x_n) = \begin{cases} 1, & \text{if } x_n \in [i, i+1), \\ 0, & \text{otherwise.} \end{cases}$$

We enforce zero end-slope conditions (i.e., the first derivative of the spline is zero at the boundaries). This ensures that the reconstructed PDF smoothly approaches zero at the boundaries and prevents unrealistic extrapolation at the edges.

The final PDF is obtained by differentiating the piecewise CDF approximation:

$$\hat{f}_{X_n|H}(x_n \mid r) = \frac{\mathrm{d}}{\mathrm{d}x_n} F_{X_n|H}(x_n \mid r)$$

IV. Simulation Results

A. Synthetic Data

We first investigate how the candidate-cutting strategy affects the MDL algorithm. A six-component univariate Gaussian mixture is sampled with $T = 20\,000$ observations, the experiment is repeated 50 times. Each method starts with the same upper limit of $K_{\max} = 50$ bins and searches for the MDL-optimal bins. Fig. 4 displays box-plots for the proposed *quantile* method, the *two-cuts*, and *mid-points* heuristics of [7]. We report (i) the bins count, (ii) the *MDL score* [3], (iii) the runtime (in minutes), and (iv) the negative log-likelihood (NLL). The MDL-score and NLL box-plots almost entirely overlap across the three strategies; medians and inter-quartile ranges coincide, indicating statistically similar fit quality. The runtime boxes, however, are spread over two orders of magnitude: the quantile method finishes in under 1 min, the mid-points grid centers around 20 min, and the two-cuts exceeds 70 min. These results indicate that restricting the candidate set via quantiles does not impose a significant loss in MDL optimality or likelihood fit, yet drastically improve computational efficiency.

Next, we evaluate our MDL-based PDF framework against a uniform discretization baseline following the set-up of [5]. Here, data are generated from five-dimensional Gaussian mixture ($N = 5$, $R = 6$), and the accuracy is measured by the KLD between the true and estimated densities, averaged over 100 Monte-Carlo trials. We compare the performance of our algorithm to that of the classical Gaussian mixture model based on EM (EM GMM), coupled tensor factorization algorithm based on alternating optimization and a KLD loss criterion (CTF-AO-KL) [5], and the "Oracle" method, which assumes that the labels (latent states) are known and serves as an empirical lower bound for the KLD. As shown in

[3] We refer to (6) as the MDL score which quantifies the quality of the MDL histogram.

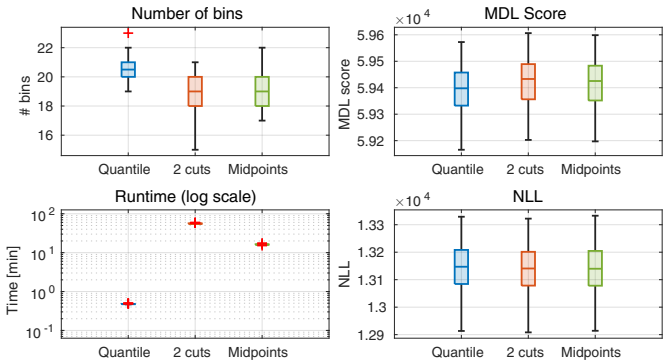

Fig. 4: Performance of Cuts Placement Strategies.

Fig. 3 (a),(b), our approach outperforms the uniformly binned (20 bins) CTF-AO-KL in both KLD and clustering accuracy. Fig. 3 (c) demonstrates how the histogram grows with T. When T is small, adding many cuts would leave only a few points per bin, so the NML term cannot compensate for the MDL penalty. As more data become available, each prospective bin contains enough points to estimate its probability reliably; the NML term now rewards a finer partition, making higher-resolution histogram both feasible and favorable. Although the EM GMM remains a practical choice for modeling Gaussian mixtures, its performance can be compromised if the discretization is not well adapted to the underlying density.

B. Real Data

We further evaluate our proposed approach on a real dry bean dataset [19]. In this experiment, each sample is characterized by features from 7 different dry bean varieties, and the objective is to accurately predict the bean class. Initially, we estimate the model rank using variational Bayesian inference (VB-PMF) [20]; given an upper bound for the rank R (e.g. the maximum rank for which the Kruskal identifiability condition for the CPD is satisfied [10]), this algorithm automatically prunes irrelevant components to determine the effective rank. The estimated rank ($R = 48$) is then used to train our method and other competing algorithms. We randomly split the data into training set ($80\,\%$) and testing set ($20\,\%$). We calculate the classification accuracy (defined as the proportion of correctly classified samples to the total number of samples) and report the results averaged over 50 random data splits. In general, as demonstrated in Table I, our method achieves higher classification accuracy while also significantly reducing the computational cost.

V. Conclusion

In this paper, we introduced a unified framework that combines MDL and tensor factorization for non-parametric PDF estimation. By integrating a quantile-based cutting strategy, we improved the MDL computation time without compromising the performance. Experimental results on both synthetic and real datasets demonstrated the advantage of our approach compared to conventional uniform binning methods.

TABLE I: Models Performance in Multiclass classification of dry beans (mean ± std).

Model	Class. Acc.	Runtime [min]
VB-PMF	86.85±0.62	8.67±0.56
Proposed	**87.72±0.67**	**2.49±1.24**
CTF-AO-KL	87.04±0.59	75.3±10.9

References

[1] C. Cocosco, V. Kollokian, R. K.-S. Kwan, and A. C. Evans, "BrainWeb: Online Interface to a 3D MRI Simulated Brain Database," *NeuroImage*, 1997.

[2] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, M. Telgarsky *et al.*, "Tensor Decompositions for Learning Latent Variable Models." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.

[3] M. F. Miranda, H. Zhu, J. G. Ibrahim, A. D. N. Initiative *et al.*, "TPRM: Tensor Partition Regression Models with Applications in Imaging Biomarker Detection," *The Annals of Applied Statistics*, vol. 12, no. 3, pp. 1422–1450, 2018.

[4] O. Gottesman, W. Pan, and F. Doshi-Velez, "Weighted Tensor Decomposition for Learning Latent Variables with Partial Data," in *Proc. 21st International Conference on Artificial Intelligence and Statistics*, 2018.

[5] N. Kargas and N. D. Sidiropoulos, "Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm," in *Proc. 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[6] J. Rissanen, T. P. Speed, and B. Yu, "Density Estimation by Stochastic Complexity," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 315–323, 1992.

[7] P. Kontkanen and P. Myllymäki, "MDL Histogram Density Estimation," in *Artificial Intelligence and Statistics*. PMLR, 2007, pp. 219–226.

[8] M. Amiridi, N. Kargas, and N. D. Sidiropoulos, "Low-Rank Characteristic Tensor Density Estimation Part I: Foundations," *IEEE Transactions on Signal Processing*, vol. 70, pp. 2654–2668, 2022.

[9] T. G. Kolda and B. W. Bader, "Tensor Decompositions and Applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[10] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor Decomposition for Signal Processing and Machine Learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[11] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, Learning, and "Kolmogorov Extension" for Finite-Alphabet Random Vectors," *IEEE Transactions on Signal Processing*, vol. 66, no. 18, pp. 4854–4868, 2018.

[12] J. K. Chege, M. J. Grasis, A. Manina, A. Yeredor, and M. Haardt, "Efficient Probability Mass Function Estimation from Partially Observed Data," in *Proc. 56th Asilomar Conference on Signals, Systems, and Computers*, 2022.

[13] Y. M. Shtarkov, "Universal Sequential Coding of Single Messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, 1987.

[14] J. Rissanen, "Strong optimality of the normalized ml models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, 2002.

[15] J. Rissanen, "Stochastic Complexity and Modeling," *The Annals of Statistics*, pp. 1080–1100, 1986.

[16] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-rank Probability Mass Tensor from Partial Observations," *IEEE Signal Processing Letters*, vol. 26, no. 10, pp. 1551–1555, 2019.

[17] S. Maymon and A. V. Oppenheim, "Sinc Interpolation of Nonuniform Samples," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4745–4758, 2011.

[18] C. de Boor, *A Practical Guide to Splines*. New York: Springer Verlag, 1978.

[19] M. Koklu and I. A. Ozkan, "Multiclass Classification of Dry Beans using Computer Vision and Machine Learning Techniques," *Computers and Electronics in Agriculture*, vol. 174, p. 105507, 2020.

[20] J. K. Chege, M. J. Grasis, A. Yeredor, and M. Haardt, "Bayesian Estimation of a Probability Mass Function Tensor with Automatic Rank Detection," in *Proc. 9th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Herradura, Costa Rica, 2023.