LEARNING MIXTURES OF SMOOTH PRODUCT DISTRIBUTIONS: IDENTIFIABILITY AND ALGORITHM

Nikos Kargas and Nicholas D. Sidiropoulos

University of Minnesota
Driven to Discover
umn.edu

Introduction

- **Learning mixture models –** fundamental problem in statistics and machine learning.
- **Applications –** density estimation and clustering.

- A PDF is a mixture of R component distributions if it can be expressed as a weighted sum of R multivariate distributions:

$$f_{\mathcal{X}}(x_1,\ldots,x_N) = \sum_{r=1}^{R} w_r f_{\mathcal{X}|H}(x_1,\ldots,x_N|r)$$

- When each conditional PDF factors into the product of its marginal densities we have:

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

- Common assumption: parametric form of the conditional PDFs such as Gaussian distributions.
- Most popular algorithm: Expectation Maximization [Dempster et al., 1977].
- Is it possible to recover mixtures of *non-parametric* product distributions?

Canonical Polyadic Decomposition

- An N-way tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ is a multidimensional array. A polyadic decomposition expresses the tensor as a sum of rank-1 terms:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{A}_1[:,r]\circ\mathbf{A}_2[:,r]\circ\cdots\circ\mathbf{A}_N[:,r]$$

- If the number of rank-1 terms is minimal, then the decomposition is called the CPD of $\underline{\mathbf{X}}$ and R is called the rank of $\underline{\mathbf{X}}$.

- Without loss of generality, we can restrict the columns of $\{\mathbf{A}_n\}_{n=1}^N$ to have unit norm and have the following equivalent expression:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \boldsymbol{\lambda}[r]\mathbf{A}_1[:,r]\circ\mathbf{A}_2[:,r]\circ\cdots\circ\mathbf{A}_N[:,r]$$

Related Work

- **EM-based:**
  – parametric models (Gaussian, Exponential, Laplace, Poisson).
  – non-parametric models [Benaglia et al., 2009, Levine et al., 2011].
    - Kernel-based methods.
    - Lack Identifiability.
- **Tensor-based:**
  – GMMs [Hsu and Kakade, 2013], categorical [Jain and Oh, 2014].
    - Parametric models, algebraic algorithms –> EM for refinement.
- Identifiability for non-parametric mixtures of product distributions [Allman et al., 2009].
    - Identifiability of the conditional PDFs given the true joint PDF, if the conditional PDFs are linearly independent.
    - No estimation procedure.

Approach

- Discretization of each random variable by partitioning its support into uniform intervals $\{\Delta_n^i = (d_n^{i-1}, d_n^i)\}_{1\le i\le I}$.
- Define the probability tensor (histogram):

$$\underline{\mathbf{X}}[i_1,\ldots,i_N] = \text{Pr}\left(X_1\in\Delta_n^{i_1},\ldots,X_N\in\Delta_n^{i_N}\right)$$

given by

$$\underline{\mathbf{X}}[i_1,\ldots,i_N] = \sum_{r=1}^{R} w_r \prod_{n=1}^{N}\int_{\Delta_n^{i_n}} f_{X_n|H}(x_n|r)dx_n = \sum_{r=1}^{R} w_r \prod_{n=1}^{N} \text{Pr}\left(X_n\in\Delta_n^{i_n}\middle| H=r\right).$$

- Is it possible to learn the mixing weights and discretized conditional PDFs from missing/limited data? **Yes!** Joint factorization of histogram estimates of lower-dimensional PDFs.
- Is it possible to recover non-parametric conditional PDFs from their discretized counterparts? **Yes,** if the conditional PDFs are smooth!

Identifiability using Lower-dimensional Statistics

- Realizations of subsets of only three random variables are sufficient to recover $\text{Pr}\left(X_n\in\Delta_n^{i_n}\middle| H=r\right)$ and $\{w_r\}_{r=1}^R$.
- A histogram of any subset of three random variables X_j, X_k, X_ℓ can be written as

$$\underline{\mathbf{X}}_{jk\ell}[i_j,i_k,i_\ell] = \textstyle\sum_{r=1}^{R}\boldsymbol{\lambda}[r]\mathbf{A}_j[i_j,r]\mathbf{A}_k[i_k,r]\mathbf{A}_\ell[i_\ell,r]$$

  which is a CPD of rank R.

- The parameters of the CPD are generically unique for $R \le \frac{(\lfloor\frac{N}{3}\rfloor I+1)^2}{16}$.

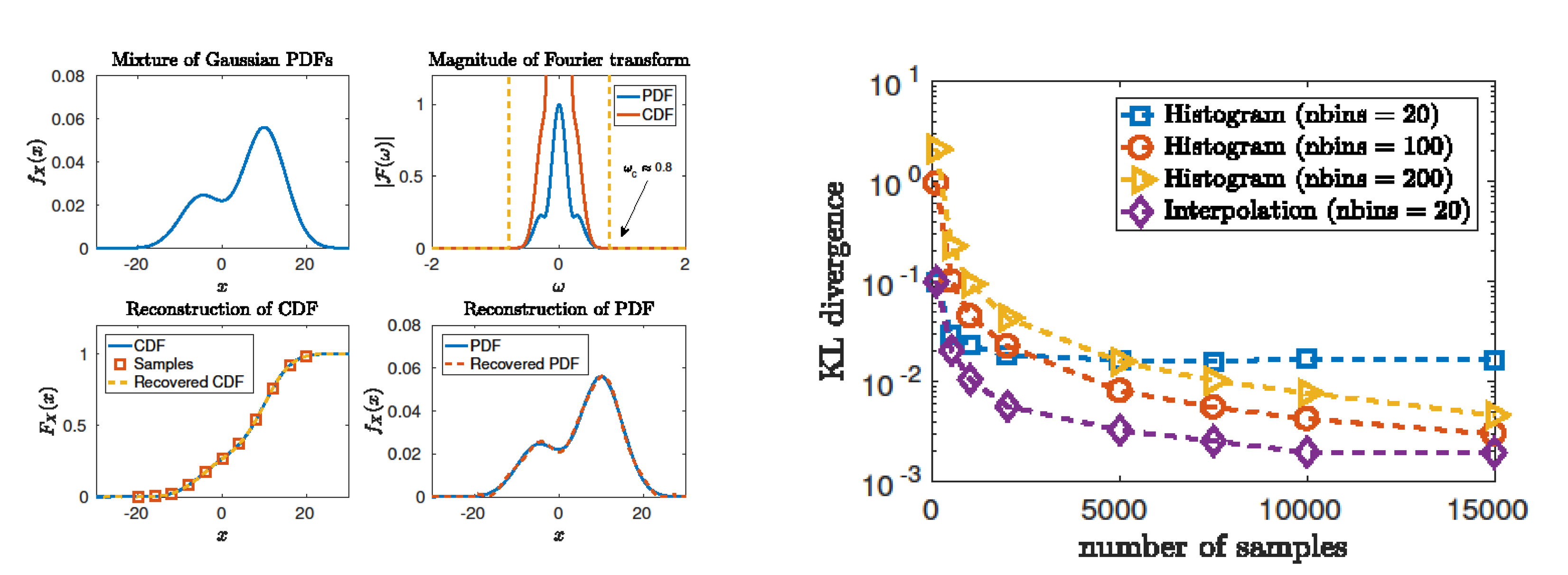
Remarks:

1. Finer discretization can lead to improved identifiability results. Many samples to reliably estimate these histograms!
2. Histograms of subsets of two variables correspond to Non-negative Matrix Factorization which is not identifiable in general!

Recovery of the Conditional PDFs

- **Proposition**: *A PDF that is (approximately) band-limited with cutoff frequency ω_c can be recovered from uniform samples of the associated CDF taken π/ω_c apart.*

Toy Example



Algorithm

- **Optimization problem:**

$$\begin{aligned}\min_{\{\mathbf{A}_n\}_{n=1}^N,\boldsymbol{\lambda}} \quad & \sum_{j=1}^{N}\sum_{k>j}^{N}\sum_{\ell>k}^{N} \text{D}\left(\widehat{\underline{\mathbf{X}}}_{jk\ell}, [\![\boldsymbol{\lambda},\mathbf{A}_j,\mathbf{A}_k,\mathbf{A}_\ell]\!]_R\right)\\ \text{s.t.} \quad & \boldsymbol{\lambda}\ge\mathbf{0}, \mathbf{1}^T\boldsymbol{\lambda}=1\\ & \mathbf{A}_n \ge \mathbf{0},\ n=1,\ldots,N\\ & \mathbf{1}^T\mathbf{A}_n = \mathbf{1}^T,\ n=1,\ldots,N\end{aligned}$$

- **Alternating optimization approach:**
  Cyclically update the variables while keeping all but one fixed.

$$\min_{\mathbf{A}_j\in\mathcal{C}} \sum_{k\neq j}\sum_{\substack{l\neq j\\ l>k}} \text{D}\left(\underline{\mathbf{X}}_{jk\ell}^{(1)}, (\mathbf{A}_\ell\odot\mathbf{A}_k)\text{diag}(\boldsymbol{\lambda})\mathbf{A}_j^T\right)$$

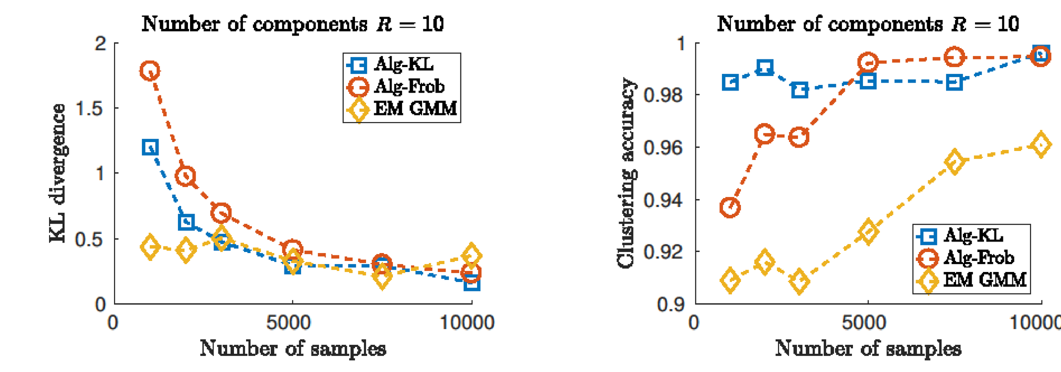
  solved via Exponentiated Gradient. The update rule becomes:

$$\mathbf{A}_j^{\tau} = \mathbf{A}_j^{\tau-1} \circledast \exp\left(-\eta_\tau\nabla f\left(\mathbf{A}_j^{\tau-1}\right)\right)$$

  Similarly for $\boldsymbol{\lambda}$.

Experiments

- **Conditional PDFs: Gaussian**



- **Conditional PDFs: Gamma**