Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm

Nikos Kargas
University of Minnesota

Nicholas D. Sidiropoulos
University of Virginia

Abstract

We study the problem of learning a mixture model of non-parametric product distributions. The problem of learning a mixture model is that of finding the component distributions along with the mixing weights using observed samples generated from the mixture. The problem is well-studied in the parametric setting, i.e., when the component distributions are members of a parametric family – such as Gaussian distributions. In this work, we focus on multivariate mixtures of non-parametric product distributions and propose a two-stage approach which recovers the component distributions of the mixture under a smoothness condition. Our approach builds upon the identifiability properties of the canonical polyadic (low-rank) decomposition of tensors, in tandem with Fourier and Shannon-Nyquist sampling staples from signal processing. We demonstrate the effectiveness of the approach on synthetic and real datasets.

1 Introduction

Learning mixture models is a fundamental problem in statistics and machine learning having numerous applications such as density estimation and clustering. In this work, we consider the special case of mixture models whose component distributions factor into the product of the associated marginals. An example is a mixture of axis-aligned Gaussian distributions, an important class of Gaussian Mixture Models (GMMs). Consider a scenario where different diagnostic tests are applied to patients, and test results are assumed to be independent conditioned on the binary disease status of the patient which is the latent variable. The joint Probability Density Function (PDF) of the tests can be expressed as a weighted sum of two components, and each component factors into the product of univariate marginals. Fitting a mixture model to an unlabeled dataset allows us to cluster the patients into two groups by determining the value of the latent variable using the Maximum a Posteriori (MAP) principle.

Most of the existing literature in this area has focused on the fully-parametric setting, where the mixture components are members of a parametric family, such as Gaussian distributions. The most popular algorithm for learning a parametric mixture model is Expectation Maximization (EM) (Dempster et al., 1977). Recently, methods based on tensor decomposition and particularly the Canonical Polyadic Decomposition (CPD) have gained popularity as an alternative to EM for learning various latent variable models (Anandkumar et al., 2014). What makes the CPD a powerful tool for data analysis is its identifiability properties, as the CPD of a tensor is unique under relatively mild rank conditions (Sidiropoulos et al., 2017).

In this work we propose a two-stage approach based on tensor decomposition for recovering the conditional densities of mixtures of smooth product distributions. We show that when the unknown conditional densities are approximately band-limited it is possible to uniquely identify and recover them from partially observed data. The key idea is to jointly factorize histogram estimates of lower-dimensional PDFs that can be easily and reliably estimated from observed samples. The conditional densities can then be recovered using an interpolation procedure. We formulate the problem as a coupled tensor factorization and propose an alternating-optimization algorithm. We demonstrate the effectiveness of the approach on both synthetic and real data.

Notation: Bold, lowercase, $\mathbf{x}$, and uppercase letters, $\mathbf{X}$, denote vectors and matrices respectively. Bold, underlined, uppercase letters, $\underline{\mathbf{X}}$, denote N-way ($N \geq 3$) tensors. We use the notation $\mathbf{x}[i]$, $\mathbf{X}[i,j]$, $\underline{\mathbf{X}}[i,j,k]$ to

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

refer to specific elements of a vector, matrix and tensor respectively. We denote the vector obtained by vertically stacking the columns of the tensor $\underline{\mathbf{X}}$ into a vector by $\text{vec}(\underline{\mathbf{X}})$. Additionally, $\text{diag}(\mathbf{x}) \in \mathbb{R}^{M\times M}$ denotes the diagonal matrix with the elements of vector $\mathbf{x} \in \mathbb{R}^M$ on its diagonal. The set of integers $\{1, \dots, N\}$ is denoted as $[N]$. Uppercase, X, and lowercase letters, x, denote scalar random variables and realizations thereof, respectively.

2 Background

2.1 Canonical Polyadic Decomposition

In this section, we briefly introduce basic concepts related to tensor 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 whose elements are indexed by N indices. A polyadic decomposition expresses $\underline{\mathbf{X}}$ as a sum of rank-1 terms

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

where $\mathbf{A}_n \in \mathbb{R}^{I_n\times R}$, $1 \le r \le R$, $\mathbf{A}_n[:, r]$ denotes the r-th column of matrix $\mathbf{A}_n$ and $\circ$ denotes the outer product. If the number of rank-1 terms is minimal, then Equation (1) 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

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \boldsymbol{\lambda}[r]\mathbf{A}_1[:, r] \circ \mathbf{A}_2[:, r] \circ \cdots \circ \mathbf{A}_N[:, r], \quad (2)$$

where $\|\mathbf{A}_n[:, r]\|_p = 1$ for a certain $p \ge 1$, $\forall\ n, r$, and $\boldsymbol{\lambda} = [\boldsymbol{\lambda}[1], \dots, \boldsymbol{\lambda}[R]]^T$ 'absorbs' the norms of columns. For convenience, we use the shorthand notation $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \dots, \mathbf{A}_N]\!]_R$. We can express the CPD of a tensor in a matricized form. With $\odot$ denoting the Khatri-Rao (columnwise Kronecker) matrix product, it can be shown that the mode-n matrix unfolding of $\underline{\mathbf{X}}$ is given by

$$\mathbf{X}^{(n)} = \left(\bigodot_{\substack{k=1\\k\neq n}}^{N} \mathbf{A}_k\right) \text{diag}(\boldsymbol{\lambda})\mathbf{A}_n^T, \quad (3)$$

where $\bigodot_{\substack{k=1\\k\neq n}}^{N} \mathbf{A}_k = \mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \odot \cdots \odot \mathbf{A}_1$.
The CPD can be expressed in a vectorized form as

$$\text{vec}(\underline{\mathbf{X}}) = \left(\bigodot_{n=1}^{N} \mathbf{A}_n\right) \boldsymbol{\lambda}. \quad (4)$$

It is clear that the rank-1 terms can be arbitrarily permuted without affecting the decomposition. We say that a CPD of a tensor is unique when it is only subject to this trivial indeterminacy.

2.2 Learning Problem

Let $\mathcal{X} = \{X_n\}_{n=1}^N$ denote a set of N random variables. We say that a PDF $f_{\mathcal{X}}$ is a mixture of R component distributions if it can be expressed as a weighted sum of R multivariate distributions

$$f_{\mathcal{X}}(x_1, \dots, x_N) = \sum_{r=1}^{R} w_r f_{\mathcal{X}|H}(x_1, \dots, x_N|r), \quad (5)$$

where $f_{\mathcal{X}|H}$ are conditional PDFs and $\{w_r\}_{r=1}^R$ are non-negative numbers such that $\sum_{r=1}^R w_r = 1$, called mixing weights. When each conditional PDF factors into the product of its marginal densities we have that

$$f_{\mathcal{X}}(x_1, \dots, x_N) = \sum_{r=1}^{R} w_r \prod_{n=1}^{N} f_{X_n|H}(x_n|r), \quad (6)$$

which can be seen as a continuous extension of the CPD model of Equation (2). A sample from the mixture model is generated by first drawing a component r according to w and then independently drawing samples for every variable $\{X_n\}_{n=1}^N$ from the conditional PDFs $f_{X_n|H}(\cdot|r)$. The problem of learning the mixture is that of finding the conditional PDFs as well as the mixing weights given observed samples.

2.3 Related Work

Mixture models have numerous applications in statistics and machine learning including clustering and density estimation to name a few (McLachlan and Peel, 2000). A common assumption made in multivariate mixture models is a parametric form of the conditional PDFs. For example, when the conditional PDFs are assumed to be Gaussian, the goal is to recover the mean vectors and covariance matrices defining each multivariate Gaussian component and the mixing weights. Other common choices include categorical, exponential, Laplace or Poisson distributions. The most popular algorithm for learning the parameters of the mixture is the EM algorithm (Dempster et al., 1977) which maximizes the likelihood function with respect to the parameters. EM-based methods have been also considered for learning mixture models of non-parametric distributions[1] by parameterizing the unknown conditional PDFs using kernel density estimators (Benaglia et al., 2009; Levine et al., 2011), which lack however theoretical guarantees.

Tensor decomposition methods can be used as an alternative to EM for learning various latent variable models (Anandkumar et al., 2014). High-order moments of several probabilistic models can be expressed

[1]The term non-parametric is used to describe the case in which no assumptions are made about the form of the conditional densities.

using low-rank CPDs. Decomposing these tensors reveals the true parameters of the probabilistic models. In the absence of noise and model mismatch, algebraic algorithms can be applied to compute the CPD under certain conditions, see (Sidiropoulos et al., 2017) and references therein, and (Hsu and Kakade, 2013) for the application to GMMs. Tensor decomposition approaches have been proposed for learning mixture models but are mostly restricted to Gaussian or categorical distributions (Hsu and Kakade, 2013; Jain and Oh, 2014; Gottesman et al., 2018). In practice, mainly due to sampling noise the result of these algorithms may not be satisfactory and EM can be used for refinement (Zhang et al., 2014; Ruffini et al., 2017). In the case of non-parametric mixtures of product distributions, identifiability of the components has been established in (Allman et al., 2009). The authors have shown that it is possible to identify the conditional PDFs given the true joint PDF, if the conditional PDFs of each X_n across different mixture components are linearly independent i.e., the continuous factor "matrices" have linearly independent columns. However, the exact true joint PDF is never given – only samples drawn from it are available in practice, and elements may be missing from any given sample. Furthermore, (Allman et al., 2009) did not provide an estimation procedure, which limits the practical appeal of an interesting theoretical contribution.

In this work, we focus on mixtures of product distributions of continuous variables and do not specify a parametric form of the conditional density functions. We show that it is possible to recover mixtures of *smooth* product distributions from observed samples. The key idea is to first transform the problem to that of learning a mixture of categorical distributions by decomposing lower-dimensional and (possibly coarsely) discretized joint PDFs. Given that the conditional PDFs are (approximately) band-limited (smooth), they can be recovered from the discretized PDFs under certain conditions.

3 Approach

Our approach consists of two stages. We express the problem as a tensor factorization problem and show that if $N \geq 3$, we can recover points of the unknown conditional CDFs. Under a smoothness condition, these points can be used to recover the true conditional PDFs using an interpolation procedure.

3.1 Problem Formulation

We assume that we are given M N-dimensional samples $\{\mathbf{x}_m\}_{m=1}^M$ that have been generated from a mixture of product distributions as in Equation (6). We discretize each random variable X_n by partitioning its support into I uniform intervals $\{\Delta_n^i = \left(d_n^{i-1}, d_n^i\right)\}_{1 \leq i \leq I}$. Specifically, we consider a discretization of the PDF and define the probability tensor (histogram) $\underline{\mathbf{X}}[i_1, \ldots, i_N] \triangleq \mathsf{Pr}\left(X_1 \in \Delta_n^{i_1}, \ldots, X_N \in \Delta_n^{i_N}\right)$ given by

$$\underline{\mathbf{X}}[i_1, \ldots, i_N] = \sum_{r=1}^{R} w_r \prod_{n=1}^{N} \int_{\Delta_n^{i_n}} f_{X_n|H}(x_n|r) dx_n$$
$$= \sum_{r=1}^{R} w_r \prod_{n=1}^{N} \mathsf{Pr}\left(X_n \in \Delta_n^{i_n} \,\middle|\, H = r\right). \quad (7)$$

Let $\mathbf{A}_n[i_n, r] \triangleq \mathsf{Pr}\left(X_n \in \Delta_n^{i_n} \middle| H = r\right)$, $\boldsymbol{\lambda}[r] \triangleq w_r$. Note that $\underline{\mathbf{X}}$ is an N-way tensor and admits a CPD with non-negative factor matrices $\{\mathbf{A}_n\}_{n=1}^N$ and rank R, i.e., $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]_R$. From equation (7) it is clear that the discretized conditional PDFs are identifiable and can be recovered by decomposing the true joint discretized probability tensor, if $N \geq 3$ and R is small enough, by virtue of the uniqueness properties of CPD (Sidiropoulos et al., 2017).

In practice we do not observe $\underline{\mathbf{X}}$ but we have to deal with perturbed versions. Based on the observed samples, we can compute an approximation of the probability tensor $\underline{\mathbf{X}}$ by counting how many samples fall into each bin and normalizing the tensor by dividing with the total number of samples. The size of the probability tensor grows exponentially with the number of variables and therefore the estimate will be highly inaccurate even when the number of discretization intervals is small. More importantly, datasets often contain missing data and therefore its impossible to construct such tensor. On the other hand, it may be possible to estimate low-dimensional discretized joint PDFs of subsets of the random variables which correspond to low-order tensors. For example, in the clustering example given in the introduction some patients may be tested on a subset of the available tests. Finally, the model of Equation (7) is just an approximation of our original model, as our ultimate goal is to recover the true conditional PDFs. To address the aforementioned challenges we have to answer the following two questions

1. Is it possible to learn the mixing weights and discretized conditional PDFs from missing/limited data?
2. Is it possible to recover non-parametric conditional PDFs from their discretized counterparts?

Regarding the first question, it has been recently shown that a joint Probability Mass Function (PMF) of a set of random variables can be recovered from

lower-dimensional joint PMFs if the joint PMF has low enough rank (Kargas et al., 2018). This result allows us to recover the discretized conditional PDFs from low-dimensional histograms but cannot be extended to the continuous setting in general because of the loss of information induced from the discretization step. We further discuss and provide conditions under which we can overcome these issues.

3.2 Identifiability using Lower-dimensional Statistics

In this section we provide insights regarding the first issue. It turns out that realizations of subsets of only three random variables are sufficient to recover $\mathsf{Pr}\left(X_n \in \Delta_n^{i_n} \middle| H = r\right)$ and $\{w_r\}_{r=1}^R$. Under the mixture model (6), a histogram of any subset of three random variables X_j, X_k, X_ℓ denoted as $\underline{\mathbf{X}}_{jk\ell}$, with $\underline{\mathbf{X}}_{jk\ell}[i_j, i_k, i_\ell] = \mathsf{Pr}(X_j \in \Delta_j^{i_j}, X_k \in \Delta_k^{i_k}, X_\ell \in \Delta_\ell^{i_\ell})$ can be written as $\underline{\mathbf{X}}_{jk\ell}[i_j, i_k, i_\ell] = \sum_{r=1}^R \boldsymbol{\lambda}[r]\mathbf{A}_j[i_j, r]\mathbf{A}_k[i_k, r]\mathbf{A}_\ell[i_\ell, r]$, which is a third-order tensor of rank R. A fundamental result on the uniqueness of tensor decomposition of third-order tensors was given by in (Kruskal, 1977). The result states that if $\underline{\mathbf{X}}$ admits a decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]_R$, with $k_{\mathbf{A}_1} + k_{\mathbf{A}_2} + k_{\mathbf{A}_3} \geq 2R + 2$ then $\text{rank}(\underline{\mathbf{X}}) = R$ and the decomposition of $\underline{\mathbf{X}}$ is unique. Here, $k_{\mathbf{A}}$ denotes the Kruskal rank of the matrix $\mathbf{A}$ which is equal to the largest integer such that every subset of $k_{\mathbf{A}}$ columns are linearly independent. When the rank is small and the decomposition is exact, the parameters of the CPD model can be computed exactly via Generalized Eigenvalue Decomposition (GEVD) and related algebraic algorithms (Leurgans et al., 1993; Domanov and Lathauwer, 2014; Sidiropoulos et al., 2017).

Theorem 1 *(Leurgans et al., 1993) Let $\underline{\mathbf{X}}$ be a tensor that admits a polyadic decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]_R$, $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times R}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times R}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times R}$, $\boldsymbol{\lambda} \in \mathbb{R}^R$ and suppose that $\mathbf{A}_1$, $\mathbf{A}_2$ are full column rank and $k_{\mathbf{A}_3} \geq 2$. Then rank($\underline{\mathbf{X}}$) $= R$, the decomposition of $\underline{\mathbf{X}}$ is unique and can be found algebraically.*

More relaxed uniqueness conditions from the field of algebraic geometry have been proven in recent years.

Theorem 2 *(Chiantini and Ottaviani, 2012) Let $\underline{\mathbf{X}}$ be a tensor that admits a polyadic decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, where $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$, $I_1 \leq I_2 \leq I_3$. Let α, β be the largest integers such that $2^\alpha \leq I_1$ and $2^\beta \leq I_2$. If $F \leq 2^{\alpha+\beta-2}$ then the decomposition of $\underline{\mathbf{X}}$ is essentially unique almost surely.*

Theorem 2 is a generic uniqueness result i.e, all non-identifiable parameters form a set of Lebesgue measure zero. To see how the above theorems can be applied in our setup, consider the joint decomposition of the probability tensors $\underline{\mathbf{X}}_{jk\ell}$. Let $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ denote disjoint ordered subsets of $[N]$, with cardinality $c_1 = |\mathcal{S}_1|$, $c_2 = |\mathcal{S}_2|$, and $c_3 = |\mathcal{S}_3|$, respectively. Let $\underline{\mathbf{Y}}$ be the $c_1 \times c_2 \times c_3$ block tensor whose (j, k, ℓ)-th block is the tensor $\underline{\mathbf{X}}_{jk\ell}$, $j \in \mathcal{S}_1$, $k \in \mathcal{S}_2$, $\ell \in \mathcal{S}_3$. It is clear that the tensor $\underline{\mathbf{Y}}$ admits a CPD $\underline{\mathbf{Y}} = [\![\boldsymbol{\lambda}, \widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2, \widehat{\mathbf{A}}_3]\!]_R$ where $\widehat{\mathbf{A}}_1 = [\mathbf{A}_{\mathcal{S}_1(1)}^T, \cdots, \mathbf{A}_{\mathcal{S}_1(c_1)}^T]^T$, $\widehat{\mathbf{A}}_2 = [\mathbf{A}_{\mathcal{S}_2(1)}^T, \cdots, \mathbf{A}_{\mathcal{S}_2(c_2)}^T]^T$, $\widehat{\mathbf{A}}_3 = [\mathbf{A}_{\mathcal{S}_3(1)}^T, \cdots, \mathbf{A}_{\mathcal{S}_3(c_3)}^T]^T$. By considering the joint decomposition of lower-dimensional discretized PDFs, we have constructed a single virtual non-negative CPD model and therefore uniqueness properties hold. For example, by setting $\mathcal{S}_1 = \{1, \ldots, \lfloor \frac{N-1}{2} \rfloor - 1\}$, $\mathcal{S}_2 = \{\lfloor \frac{N-1}{2} \rfloor, \ldots, N-1\}$, $\mathcal{S}_3 = \{N\}$ we have that

$$\underline{\mathbf{Y}}^{(1)} = \left(\begin{bmatrix} \mathbf{A}_{\lfloor \frac{N-1}{2} \rfloor} \\ \vdots \\ \mathbf{A}_{N-1} \end{bmatrix} \odot \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_{\lfloor \frac{N-1}{2} \rfloor - 1} \end{bmatrix} \right) \text{diag}(\boldsymbol{\lambda}) \mathbf{A}_N^T.$$

According to Theorem 1, the CPD can be computed exactly if $R \leq (\lfloor \frac{N-1}{2} \rfloor - 1)I$. Similarly, it is easy to verify that by setting $c_1 = c_2 = \lfloor \frac{N}{3} \rfloor I$, i.e., $\alpha = \lfloor \log_2(\lfloor \frac{N}{3} \rfloor I) \rfloor$, the CPD of $\underline{\mathbf{Y}}$ is generically unique for $R \leq 2^{2(\alpha-1)}$ according to Theorem 2. The later inequality is implied by $R \leq \frac{(\lfloor \frac{N}{3} \rfloor I + 1)^2}{16}$ which shows that the bound is quadratic in N and I.

Remark 1: The previous discussion suggests that finer discretization can lead to improved identifiability results. The number of hidden components may be arbitrarily large and we may still be able to identify the discretized conditional PDFs by increasing the dimensions of the sub-tensors i.e., the discretization intervals of the random variables. The caveat is that one will need many more samples to reliably estimate these histograms. Ideally, one would like to have the minimum number of intervals that can guarantee identifiability of the conditional PDFs.

Remark 2: The factor matrices can be recovered by decomposing the lower-order probability tensors of dimension $N \geq 3$. It is important to note that histograms of subsets of two variables correspond to Non-negative Matrix Factorization (NMF) which is not identifiable unless additional conditions such as sparsity are assumed for the latent factors (Fu et al., 2018). Therefore, second-order distributions are not sufficient for recovering dense latent factor matrices.

3.3 Recovery of the Conditional PDFs

In the previous section we have shown that given lower-dimensional discretized PDFs, we can uniquely

identify and recover discretized versions of the conditional PDFs via joint tensor decomposition. Recovering the true conditional PDFs from the discretized counterparts can be viewed as a signal reconstruction problem. We know that this is not possible unless the signals have some smoothness properties. We will use the following result.

Proposition 1 *A PDF that is (approximately) band-limited with cutoff frequency ω_c can be recovered from uniform samples of the associated CDF taken $\frac{\pi}{\omega_c}$ apart.*

Proof: Assume that the PDF f_X is band-limited with cutoff frequency ω_c i.e., its Fourier transform $\mathcal{F}(\omega) = 0, \forall\ |\omega| \geq \omega_c$. Let F_X denote the CDF of f_X, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. We can express the integration as a convolution of the PDF with a unit step function, i.e., $F_X(x) = \int_{-\infty}^{\infty} f_X(t)u(x-t)d\tau$. The Fourier transform of a convolution is the point-wise product of Fourier transforms. Therefore, we can express the Fourier transform $\mathcal{G}(\omega)$ of the CDF as

$$\mathcal{G}(\omega) = \pi\delta(\omega)\mathcal{F}(0) + \frac{\mathcal{F}(\omega)}{j\omega}, \tag{8}$$

where $\delta(\cdot)$ is the Dirac delta. From Equation (8), it is clear that the CDF obeys the same band-limit as the PDF . From Shannon's sampling theorem we have that

$$F_X(x) = \sum_{n=-\infty}^{\infty} F_X(nT)\ \mathrm{sinc}\left(\frac{x-nT}{T}\right), \tag{9}$$

where $T = \frac{\pi}{\omega_c}$. The PDF can then be determined by differentiation, which amounts to linear interpolation of the CDF samples using the derivative of the sinc kernel. Note that for exact reconstruction of f_X an infinite number of data points are needed. In signal processing practice we always deal with finite support signals which are only approximately band-limited; the point is that the bandlimited assumption is accurate enough to afford high-quality signal reconstruction. In our present context, a good example is the Gaussian distribution: even though it is of infinite extent, it is not strictly bandlimited (as its Fourier transform is another Gaussian); but it is approximately bandlimited, and that is good enough for our purposes, as we will see shortly.

In section 3.2, we saw how lower-dimensional histograms can be used to obtain estimates of the discretized conditional PDFs. Now, consider the conditional PDF of the n-th variable given the r-th component. The corresponding column of factor matrix $\mathbf{A}_n$ is

$$\mathbf{A}_n[:, r] = [F_{X_n|H}(d_n^1|r) - F_{X_n|H}(d_n^0|r), \ldots, 1 - F_{X_n|H}(d_n^{I-1}|r)]^T.$$

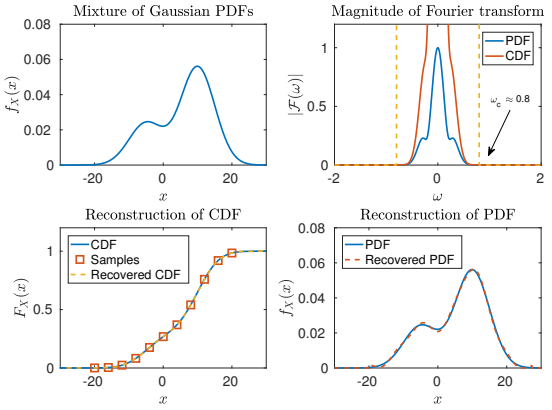

Figure 1: Illustration of the key idea on a univariate Gaussian mixture. The CDF can be recovered from its samples if $T_s \leq \frac{\pi}{0.8}$.

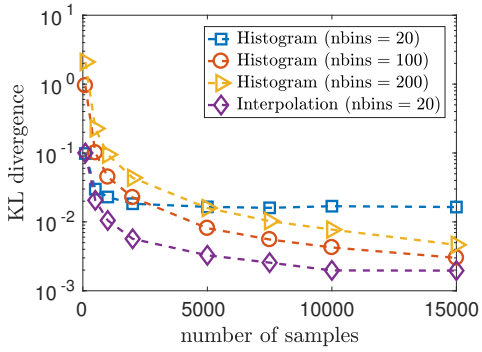

Figure 2: KL divergence between the true mixture of Gaussians and different approximations.

Since, $F_{X_n|H}(d_n^0|r) = 0$, we can compute $F_{X_n|H}(d_n^i|r)$, $\forall i \in [I-1], n \in [N]$. We also know that $F_{X_n|H}(x_n|r) = 1, \forall x_n \geq d_n^I$. Therefore, we can recover the conditional CDFs using the interpolation formula

$$F_{X|H}(x_n|r) = \sum_{k=-L}^{L} F_{X_n|H}(kT|r)\ \mathrm{sinc}\left(\frac{x-kT}{T}\right), \tag{10}$$

where $T = d_n^i - d_n^{i-1}$ and L a large integer. The conditional PDF $f_{X_n|H}$ can then be recovered via differentiation.

3.4 Toy example

An example to illustrate the idea is shown in Figure 1. Assume that the PDF of a random variable is a mixture of two Gaussian distributions with means $\mu_1 = -6$, $\mu_2 = 10$ and standard deviations $\sigma_1 = \sigma_2 = 5$. It is clear from Figure 1 that $\mathcal{F}(\omega) \approx 0$ for $|\omega| \geq \omega_c = 0.8$ and therefore the PDF is approximately band-limited. The CDF has the same band-limit, thus, it can be

recovered from points being $T = \frac{\pi}{\omega_c} \approx 4$ apart. In this example we have used only 10 discretization intervals as they suffice to capture 99% of the data. We use the finite sum formula of Equation (10) to recover the CDF and then we recover the PDF by differentiating the CDF. The recovered PDF essentially coincides with the true PDF given a few exact estimates of the CDF as shown in Figure 1.

Figure 2 shows the approximation error for different methods when we do not have exact points of the CDF but estimate them from randomly drawn samples. We know that a histogram converges to the true PDF as the number of samples grows and the bin width goes to 0 at appropriate rate. However, when the conditional PDF is smooth, the interpolation procedure using a few discretization intervals leads to a lower approximation error compared to plain histogram estimates as illustrated in the figure.

4 Algorithm

In this section we develop an algorithm for recovering the latent factors of the CPD model given the histogram estimates of lower-dimensional PDFs (Alg. 1). We define the following optimization problem

$$\begin{aligned} \min_{\{\mathbf{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \quad & \sum_{j=1}^{N}\sum_{k>j}^{N}\sum_{\ell>k}^{N} \mathrm{D}\left(\widehat{\underline{\mathbf{X}}}_{jk\ell}, [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_\ell]\!]_R\right) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{1}^T\boldsymbol{\lambda} = 1 \\ & \mathbf{A}_n \geq \mathbf{0},\ n = 1 \dots N \\ & \mathbf{1}^T\mathbf{A}_n = \mathbf{1}^T,\ n = 1 \dots N \end{aligned} \tag{11}$$

where $\mathrm{D}(\cdot,\cdot)$ is a suitable metric. The Frobenious norm and Kullback-Leibler (KL) divergence are considered in this work. For probability tensors $\underline{\mathbf{X}}, \underline{\mathbf{Y}}$ we define

$$\mathrm{D}_{\mathrm{KL}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) \triangleq \sum_{i_1,i_2,i_3} \underline{\mathbf{X}}[i_1,i_2,i_3] \log \frac{\underline{\mathbf{X}}[i_1,i_2,i_3]}{\underline{\mathbf{Y}}[i_1,i_2,i_3]}$$

$$\mathrm{D}_{\mathrm{FRO}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) \triangleq \sum_{i_1,i_2,i_3} \left(\underline{\mathbf{X}}[i_1,i_2,i_3] - \underline{\mathbf{Y}}[i_1,i_2,i_3]\right)^2.$$

Optimization problem (11) is non-convex and NP-hard in its general form. Nevertheless, sensible approximation algorithms can be derived, based on well-appreciated nonconvex optimization tools. The idea is to cyclically update the variables while keeping all but one fixed. By fixing all other variables and optimizing with respect to $\mathbf{A}_j$ we have

$$\min_{\mathbf{A}_j \in \mathcal{C}} \sum_{k\neq j}\sum_{\substack{l\neq j \\ l>k}} \mathrm{D}\left(\underline{\mathbf{X}}^{(1)}_{jk\ell}, (\mathbf{A}_\ell \odot \mathbf{A}_k)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_j^T\right), \tag{12}$$

where $\mathcal{C} = \{\mathbf{A} \mid \mathbf{A} \geq 0, \mathbf{1}^T\mathbf{A} = \mathbf{1}^T\}$. Problem (12) is convex and can be solved efficiently using Exponentiated Gradient (EG) (Kivinen and Warmuth, 1997)

Algorithm 1 Proposed Algorithm

Input: A dataset $\mathbf{D} \in \mathbb{R}^{M\times N}$
Estimate $\underline{\mathbf{X}}_{jk\ell}\ \forall j,k,\ell \in [N],\ \ell > k > j$ from data.
Initialize $\{\mathbf{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$.
repeat
  for all $n \in [N]$ **do**
    Solve opt. problem (12) via mirror descent.
  end for
  Solve opt. problem (14) via mirror descent.
until convergence criterion is satisfied
for all $n \in [N]$ **do**
  Recover $f_{X_n|H}$ by differentiation using Eq. (10)
end for

– which is a special case of mirror descent (Beck and Teboulle, 2003). At each iteration τ of mirror descent we update $\mathbf{A}_j^\tau$ by solving

$$\mathbf{A}_j^\tau = \arg\min_{\mathbf{A}_j \in \mathcal{C}} \langle\ \nabla f\big(\mathbf{A}_j^{\tau-1}\big), \mathbf{A}_j\rangle + \frac{1}{\eta_\tau} B_\Phi\big(\mathbf{A}_j, \mathbf{A}_j^{\tau-1}\big)$$

where $B_\Phi(\mathbf{A}, \widehat{\mathbf{A}}) = \Phi(\mathbf{A}) - \Phi(\widehat{\mathbf{A}}) - \langle\ \mathbf{A} - \widehat{\mathbf{A}}, \nabla\Phi(\widehat{\mathbf{A}})\rangle$ is a Bregman divergence. Setting Φ to be the negative entropy $\Phi(\mathbf{A}) = \sum_{i,j} \mathbf{A}(i,j)\log\mathbf{A}(i,j)$, the update becomes

$$\mathbf{A}_j^\tau = \mathbf{A}_j^{\tau-1} \circledast \exp\left(-\eta_\tau \nabla f\left(\mathbf{A}_j^{\tau-1}\right)\right), \tag{13}$$

where $\circledast$ is the Hadamard product, followed by column normalization $\mathbf{A}_j^\tau[:,r] = \frac{\mathbf{A}_j[:,r]}{\mathbf{1}^T\mathbf{A}[:,r]}$. The optimization problem with respect to $\boldsymbol{\lambda}$ is the following

$$\min_{\boldsymbol{\lambda}\in\mathcal{C}} \sum_{j,k,\ell} \mathrm{D}\left(\mathrm{vec}(\underline{\mathbf{X}}_{jk\ell}), (\mathbf{A}_\ell \odot \mathbf{A}_k \odot \mathbf{A}_j)\boldsymbol{\lambda}\right). \tag{14}$$

The update rules for $\boldsymbol{\lambda}$ are similar

$$\boldsymbol{\lambda}^\tau = \boldsymbol{\lambda}^{\tau-1} \circledast \exp\left(-\eta_\tau \nabla f\left(\boldsymbol{\lambda}^{\tau-1}\right)\right). \tag{15}$$

The step η_τ can be computed by the Armijo rule (Bertsekas, 1999).

5 Experiments

5.1 Synthetic Data

In this section, we employ synthetic data simulations to showcase the effectiveness of the proposed algorithm. Experiments are conducted on synthetic datasets $\{\mathbf{x}_m\}_{m=1}^M$ of varying sample sizes, generated from R component distributions. We set the number of variables to $N = 10$, and vary the number of components $R \in \{5, 10\}$. We run the algorithms using 5 different random initializations and for each algorithm keep the model that yields the lowest cost. We

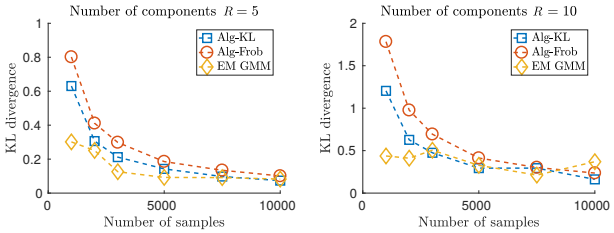

Figure 3: KL divergence (Gaussian).

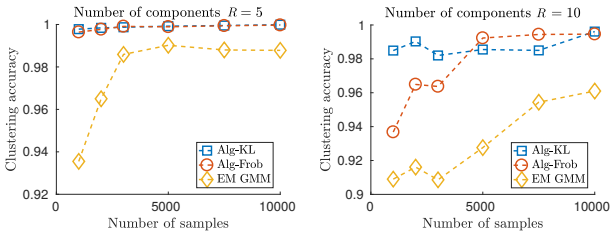

Figure 4: Clustering accuracy (Gaussian).

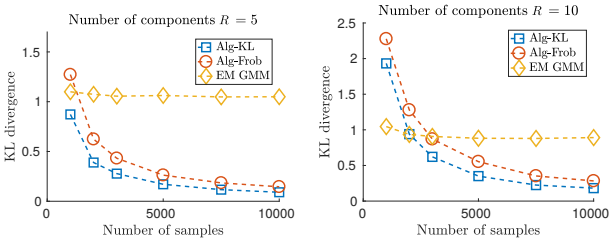

Figure 5: KL divergence (GMM).

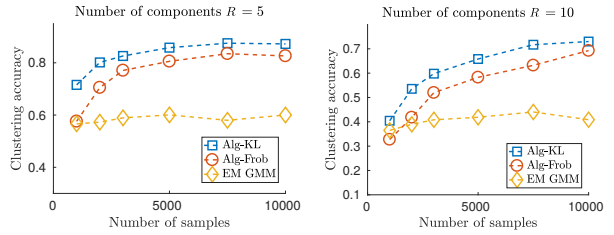

Figure 6: Clustering Accuracy (GMM).

evaluate the performance of the algorithms by calculating the KL divergence between the true and learned model, which is approximated using Monte Carlo integration. Specifically, we generate $\{\mathbf{x}_{m'}\}_{m'=1}^{M'}$ test points, $M' = 1000$ drawn from the mixture and approximate the KL divergence between the true and learned model by

$$\mathrm{D}_{\mathrm{KL}}\left(f_{\mathcal{X}}, \widehat{f}_{\mathcal{X}}\right) \approx \frac{1}{M'} \sum_{m'=1}^{M'} \log f_{\mathcal{X}}(\mathbf{x}_{m'}) / \widehat{f}_{\mathcal{X}}(\mathbf{x}_{m'}).$$

We also compute the clustering accuracy on the test points as follows. Each data point $\mathbf{x}_{m'}$ is first assigned to the component yielding the highest posterior probability $\widehat{c}_m = \arg\max_c f_{H|\mathcal{X}}(c|\mathbf{x}_m)$. Due to the label permutation ambiguity, the obtained components are aligned with the true components using the Hungarian algorithm (Kuhn, 1955). The clustering accuracy is then calculated as the ratio of wrongly labeled data points over the total number of data points.For each scenario, we repeat 10 Monte Carlo simulations and report the average results. We explore the following settings for the conditional PDFs: (1) Gaussian (2) GMM with two components (3) Gamma and (4) Laplace. The mixing weights are drawn from a Dirichlet distribution $\omega \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$ with $\alpha_r = 10\ \forall r$. We emphasize that our approach does not use any knowledge of the parametric form of the conditional PDFs; it only assumes smoothness.

Gaussian Conditional Densities: In the first experiment we assume that each conditional PDF is a Gaussian. For cluster r and random variable X_n we set $f_{X_n|H}(x_n|r) = \mathcal{N}(\mu_{nr}, \sigma_{nr}^2)$. Mean and variance are drawn from uniform distributions, $\mu_{nr} \sim \mathcal{U}(-5, 5)$, $\sigma_{nr}^2 \sim \mathcal{U}(1, 2)$. We compare the performance of our algorithms to that of EM (EM GMM). Figure 3 shows the KL divergence between the true and the learned model for various dataset sizes and different number of components. We see that the performance of our methods converges to that of EM despite the fact that we do not assume a particular model for the conditional densities. Interestingly, our approach performs better in terms of clustering accuracy as shown in Figure 4. We can see that although the joint distribution learned by EM is closer to the true in terms of the KL divergence, EM may fail to identify the true parameters of every component.

GMM Conditional Densities: In the second experiment we assume that each conditional PDF is itself a mixture model of two univariate Gaussian distributions. More specifically, we set $f_{X_n|H}(x_n|r) = \frac{1}{2}\mathcal{N}\left(\mu_{nr}^{(1)}, \sigma_{nr}^{(1)2}\right) + \frac{1}{2}\mathcal{N}\left(\mu_{nr}^{(2)}, \sigma_{nr}^{(2)2}\right)$. Means and variances are drawn from uniform distributions $\mu_{nr}^{(1)} \sim \mathcal{U}(0, 7)$, $\sigma_{nr}^{(1)2} \sim \mathcal{U}(1, 4)$, $\mu_{nr}^{(2)} \sim \mathcal{U}(-7, 0)$, $\sigma_{nr}^{(2)2} \sim \mathcal{U}(1, 4)$. Our method is able to learn the mixture model in contrast to the EM GMM which exhibits poor performance, due to the model mismatch, as shown in Figures 5, 6.

Gamma Conditional Densities: Another example of a smooth distribution is the shifted Gamma distribution. We set $f_{X_n|H}(x_n|r) = \frac{1}{\beta^\alpha \Gamma(\alpha)}(x - \mu_{nr})^{\alpha-1} \exp(-\frac{x-\mu_{nr}}{\beta})$ with $\alpha = 5$, $\mu_{nr} \sim \mathcal{U}(-5, 0)$, $\beta_{nr} \sim \mathcal{U}(0.1, 0.5)$. As the number of samples grows our method exhibits better performance, significantly outperforming EM GMM as shown in Figures 7, 8.

Laplace Conditional Densities: In the last

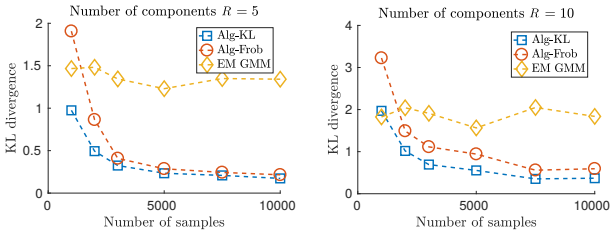

Figure 7: KL divergence (Gamma).

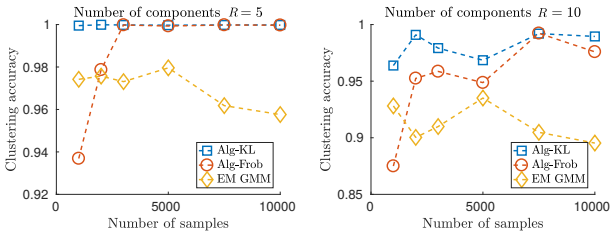

Figure 8: Clustering accuracy (Gamma).

simulated experiment we assume that each conditional PDF is a Laplace distribution with mean μ_{nr} and standard deviation σ_{nr} i.e., $f_{X_n|H}(x_n|r) = \frac{1}{\sqrt{2}\sigma_{nr}} \exp\left(\frac{\sqrt{2}|x_n-\mu_{nr}|}{\sigma_{nr}}\right)$. A Laplace distribution in contrast to the previous cases is not smooth (at its mean). Parameters are drawn from uniform distributions, $\mu_{nf} \sim \mathcal{U}(-5,5)$, $\sigma_{nf}^2 \sim \mathcal{U}(5,10)$. We compare the performance of our methods to that of the EM GMM and an EM algorithm for a Laplace mixture model (EM LMM). The proposed method approaches the performance of EM LMM and exhibits better performance in terms of KL and clustering accuracy compared to the EM GMM for higher number of data samples, as shown in Figures 9, 10.

5.2 Real Data

Finally, we conduct several real-data experiments to test the ability of the algorithms to cluster data. We selected 7 datasets with continuous variables suitable for classification or regression tasks from the UCI repository. For each labeled dataset we hide the label and treat it as the latent component. For datasets that contained a continuous variable as a response, we discretized the response into R uniform intervals and treated it as the latent component. For each dataset we repeated 10 Monte Carlo simulations by randomly splitting the dataset into three sets; 70% was used as a training set, 10% as a validation set and 20% as a test set. The validation set was used to select the number of discretization intervals which was either 5 or 10. We compare our methods against the EM GMM with diagonal covariance, EM GMM with full-covariance and the K-means algorithm in terms of clustering accuracy.

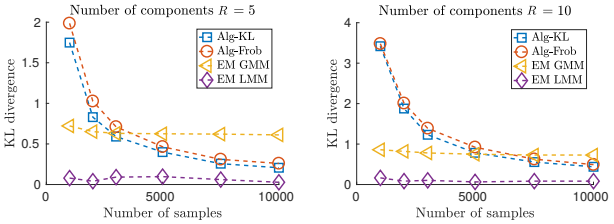

Figure 9: KL divergence (Laplace).

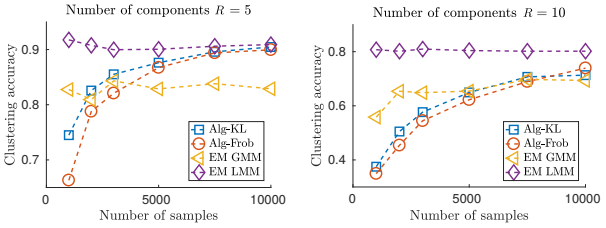

Figure 10: Clustering accuracy (Laplace).

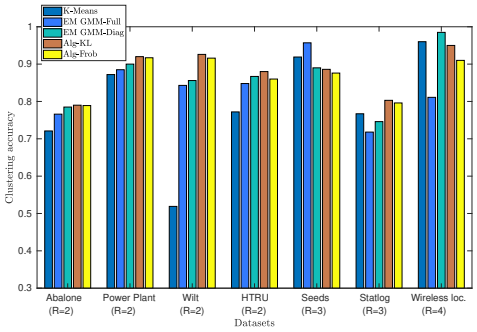

Figure 11: Clustering accuracy on real datasets.

Note that although the conditional independence assumption may not actually hold in practice, almost all the algorithms give satisfactory results in the tested datasets. The proposed algorithms perform well, outperforming the baselines in 5 out of 7 datasets while performing reasonably well in the remaining.

6 Discussion and Conclusion

We have proposed a two-stage approach based on tensor decomposition and signal processing tools for recovering the conditional densities of mixtures of smooth product distributions. Our method does not assume a parametric form for the unknown conditional PDFs. We have formulated the problem as a coupled tensor factorization and proposed an alternating-optimization algorithm. Experiments on synthetic data have shown that when the underlying conditional PDFs are indeed smooth our method can recover them with high accuracy. Results on real data have shown satisfactory performance on data clustering tasks.

References

E. S. Allman, C. Matias, J. A. Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, Aug. 2014.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

T. Benaglia, D. Chauveau, and D. R. Hunter. An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

L. Chiantini and G. Ottaviani. On generic identifiability of 3-tensors of small rank. *SIAM Journal on Matrix Analysis and Applications*, 33(3):1018–1037, 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

I. Domanov and L. D. Lathauwer. Canonical polyadic decomposition of third-order tensors: reduction to generalized eigenvalue decomposition. *SIAM Journal on Matrix Analysis and Applications*, 35(2):636–660, 2014.

X. Fu, K. Huang, and N. D. Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, Mar. 2018.

O. Gottesman, W. Pan, and F. Doshi-Velez. Weighted tensor decomposition for learning latent variables with partial data. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1664–1672, Apr. 2018.

D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

P. Jain and S. Oh. Learning mixtures of discrete product distributions using spectral decompositions. In *Conference on Learning Theory*, pages 824–856, 2014.

N. Kargas, N. D. Sidiropoulos, and X. Fu. Tensors, learning, and kolmogorov extension for finite-alphabet random vectors. *IEEE Transactions on Signal Processing*, 66(18):4854–4868, Sept 2018.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

S. Leurgans, R. Ross, and R. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

M. Levine, D. R. Hunter, and D. Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, pages 403–416, 2011.

G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, 2000.

M. Ruffini, R. Gavalda, and E. Limon. Clustering patients with tensor decomposition. In *Machine Learning for Healthcare Conference*, pages 126–146, 2017.

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*, 65 (13):3551–3582, July 2017.

Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pages 1260–1268, 2014.