Tensor decomposition for learning Gaussian mixtures from moments

Rima Khouja, Pierre-Alexandre Mattei, Bernard Mourrain

Tensor decomposition for learning Gaussian mixtures from moments

Rima Khouja, Pierre-Alexandre Mattei, Bernard Mourrain

Inria d'Université Côte d'Azur, 2004 route des Lucioles, B.P. 93, 06902 Sophia Antipolis, France

Abstract

In data processing and machine learning, an important challenge is to recover and exploit models that can represent accurately the data. We consider the problem of recovering Gaussian mixture models from datasets. We investigate symmetric tensor decomposition methods for tackling this problem, where the tensor is built from empirical moments of the data distribution. We consider identifiable tensors, which have a unique decomposition, showing that moment tensors built from spherical Gaussian mixtures have this property. We prove that symmetric tensors with interpolation degree strictly less than half their order are identifiable and we present an algorithm, based on simple linear algebra operations, to compute their decomposition. Illustrative experimentations show the impact of the tensor decomposition method for recovering Gaussian mixtures, in comparison with other state-of-the-art approaches.

1. Introduction

With the relatively recent evolutions of information systems over the last decades, many observations, measurements, data are nowadays available on a variety of subjects. However, too much information can kill the information and one of the main challenges remains to analyse and to model these data, in order to recover and exploit hidden structures.

To tackle this challenge, popular Machine Learning technologies have been developed and used successfully in several application domains (e.g. in image recognition [HZRS16]). These techniques can be grouped in two main classes: Supervised machine learning techniques are approximating a model by optimising the parameters of an enough general model (e.g. a Convolution Neural Network) from training data. Unsupervised machine learning techniques are deducing the parameters characterising a model directly from the given data, using an apriori knowledge on the model. The supervised approach requires annotated data, with a training step that can introduce some bias in the learned model. The unsupervised approach can be applied directly on a given data set avoiding the costly step of annotating data, but the quality of the output strongly depends on the type of models to be recovered.

We consider the latter approach and show how methods from effective algebraic geometry help finding hidden structure in data that can be modelled by mixtures of Gaussian distributions. The algebraic-geometric tool that we consider is tensor decomposition. It consists in decomposing a tensor into a minimal sum of rank-1 tensors. This decomposition generalises the rank decomposition of a matrix, with specific and interesting features. Contrarily to matrix rank decomposition, the decomposition of a tensor is usually unique (up to permutations) when the rank of the tensor, that is the minimal number of rank-1 terms in a decomposition, is small compared to the dimension of the space(s) associated to the tensor. Such a tensor is called *identifiable*. This property is of particular importance when the decomposition is used to recover the parameters of a model. It guaranties the validity of the recovering process and its convergence when the number of data increases.

It has been shown in [COV16] that for symmetric tensors, if the rank of the tensor is strictly less than the rank r_g of a generic tensor of the same size, then the tensor is generically identifiable, except in three cases. We show in Theorem 3.6 a more specific result: for a symmetric tensor T having a decomposition with r points, if the Hankel matrix associated to T in a degree strictly bigger than the degree of interpolation of the r points is of rank r, then the tensor is identifiable. We show in Proposition 3.3, that under some assumption on the spherical gaussian mixtures, a tensor of moments of order 3 of the distribution is identifiable and its decomposition allows to recover the parameters of the Gaussian mixture.

Several types of method have been developed to tackle the difficult problem of tensor decomposition. Direct methods based on simultaneous diagonalisation of matrices built from slices of tensors have been investigated for 3rd order multilinear tensors, e.g. in [Har70, SK90, LRA93, DDL14] or for multilinear tensors of rank smaller than the lowest dimension in [DL06, LA14]. In his proof on lower bounds of tensor ranks, Strassen showed in [Str83, Theorem 4.1] that a 3rd order multilinear tensor is of rank r if it can be embedded into a tensor with slices of rank r matrices, which are simultaneously diagonalised.

For symmetric tensor decomposition, a method based on flat extension of Hankel matrices or commutation of multiplication operators has been proposed in [BCMT10] and extended to multi-symmetric tensors in [BBCM13]. This approach is closely related to the simultaneous diagonalisation of tensor slices, but follows a more algebraic perspective. Eigenvectors of symmetric tensors have been used to compute their decompositions in [OO13]. In [HKM18], Singular Value Decomposition and eigenvector computation are used to decompose a symmetric tensor, when its rank is smaller than the smallest size of its Hankel matrix in degree less than half the order of the tensor. In Section 3, we describe a new algorithm, involving Singular Value Decomposition and simultaneous diagonalisation, to compute the decomposition of an identifiable tensor, which interpolation degree is smaller that half the order of the tensor.

Numerical methods such as homotopy continuation have been applied to tensor decomposition in [HOOS19, BDHM17]. Distance minimisation methods to compute low rank approximations of tensors have also been investigated. Alternating Least Squares (ALS) methods, updating alternately the different factors of the tensor decomposition, is a popular approach (see e.g. [CC70, CHLZ12, Har70, KB09]), but suffers from a slow convergence [EHK15, Usc12]. Other iterative methods such as quasi-Newton methods have been considered to improve the convergence speed. See e.g. [HH82, Paa99, PTC13, SL10, SBL13, TB06, BV18] for multilinear tensors. A Riemannian Newton iteration for symmetric tensors is presented in [KKM22]. In [KP09], a method for decomposing real even-order symmetric tensors, called Subspace Power Method (SPM), and similar to the power method for matrix eigenvector computation, is proposed. In these methods, the choice of the initial decomposition is crucial. In the applications of these algorithms, the initial point is often chosen at random, yielding approximate decompositions which can hardly be controlled. Tensor decomposition methods have numerous applications [KB09]. Some of them were exploited more recently in Machine Learning. In [HK13], symmetric tensor decompositions for moment tensors are studied for spherical Gaussian mixtures. Moment methods have been further investigated for Latent Dirichlet Allocation models, topic or multiview models in [AGH$^+$14, JGKA19]. In [RGL17], a tensor decomposition technique based on Alternate Least Squares (ALS) is used to initialise the Expectation Maximisation (EM) algorithm, for a mixture of discrete distributions (which are not Gaussian distributions).

An overview of tensor decomposition methods in Machine Learning can be found in [RSG17].

After reviewing Gaussian mixtures and moment methods in Section 2, we present in Section 3 an algebraic symmetric tensor decomposition method for identifiable tensors. In Section 4, we apply this algorithm for recovering Gaussian mixtures and show its impact on providing good initialisation point in the EM algorithm, in comparison with other state-of-the-art approaches.

2. Gaussian mixtures and high order moments

In this section, we review Gaussian mixture models and their applications to clustering.

2.1. Gaussian mixtures

Suppose that we wish to deal with some Euclidean data $x \in \mathbb{R}^m$, coming from a population composed of r homogeneous sub-populations (often called *clusters*). A reasonable assumption is then that each sub-population can be modelled using a simple probability distribution (e.g. Gaussian). This idea is at the heart of the notion of *mixture distribution*. The prime example of mixture is the *Gaussian mixture*, whose probability density over $\mathbb{R}^m$ is defined as

$$p_\theta(x) = \sum_{j=1}^{r} \omega_j \mathcal{N}(x|\mu_j, \Sigma_j), \tag{1}$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ denotes the Gaussian density with mean $\mu \in \mathbb{R}^m$ and definite positive covariance matrices $\Sigma \in \mathcal{S}_m^{++}$. The mixture is parametrised by a typically unknown $\theta = (\omega_1, ..., \omega_r, \mu_1, ..., \mu_r, \Sigma_1, ..., \Sigma_r)$, composed of

- $\omega = (\omega_1, ..., \omega_r)$, that belong to the r-simplex and correspond to the cluster proportions,
- μ_j and Σ_j, that correspond respectively to the mean and covariance of each cluster $j \in \{1, ..., r\}$.

Gaussian mixtures are ubiquitous objects in statistics and machine learning, and own their popularity to many reasons. Let us briefly mention a few of these.

Density estimation. If r is allowed to be sufficiently large, it is possible to approximate any probability density using a Gaussian mixture (see e.g. [NCNM20]). This motivates the use of Gaussian mixtures as powerful density estimators that can be subsequently used for downstream tasks such as missing data imputation [DZGL07], supervised classification [HT96], or image classification [SPMV13] and denoising [HBD18].

Clustering. Perhaps the most common use of Gaussian mixtures is *clustering*, also called *unsupervised classification*. The task of clustering consists in uncovering homogeneous groups among the data at hand. Within the context of Gaussian mixtures, each group generally corresponds to a single Gaussian distribution, as in Equation (1). If the parameters of a mixture are known, then each point may be clustered using the posterior probabilities obtained via Bayes's rule:

$$\forall x \in \mathbb{R}^m, k \in \{1, ..., r\}, \ \Pr(x \text{ belongs to cluster } j) = \frac{\omega_j \mathcal{N}(x|\mu_j, \Sigma_j)}{p_\theta(x)}. \tag{2}$$

Detailed reviews on mixture models and their applications, notably to clustering, can be found in [FR02, BCMR19, MLR19].

2.2. Learning mixture models

The main statistical question pertaining mixture models is to estimate the parameters $\theta = (\omega_1, ..., \omega_r, \mu_1, ..., \mu_r, \Sigma_1, ..., \Sigma_r)$ based on a data set $x_1, ..., x_n$. Typically, $X_1, ..., X_n$ are assumed to be independent and identically distributed random variables with common density p_{data}. The problem of statistical estimation is then to find some θ such that $p_\theta \approx p_{\text{data}}$. There are many approaches to this question, the most famous one being the *maximum likelihood method*. Maximum likelihood is based on the idea that maximising the *log-likelihood function*

$$\ell(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i), \tag{3}$$

will lead to appropriate values of θ. One heuristic reason of the good behaviour of maximum likelihood is that $\ell(\theta)$ can be seen as a measure of how likely the observed data is, according to the mixture model p_θ. This means that the maximum likelihood estimate will be the value of θ that renders the observed data the likeliest. Another interesting interpretation of maximum likelihood in information-theoretic: when $n \longrightarrow \infty$, maximising the log-likelihood is equivalent to minimising the Kullback-Leibler divergence (an information-theoretic measure of distance between probability distributions) between p_θ and p_{data}, thus giving a precise sense to the statement $p_\theta \approx p_{\text{data}}$ (see e.g. [Bis06, Section 1.6.1]). For more details on the properties of maximum likelihood, see e.g. [VdV98, Section 5.5].

In the specific case of a mixture model, performing maximum-likelihood is however complex for several reasons. Firstly, as shown for instance by [LC90], finding a global maximum is actually often ill-posed in the sense that some problematic values of θ will lead to $\ell(\theta) = \infty$ while being very poor models of the data. While focusing on local rather global maxima will fix this first issue in a sense, iterative optimisation algorithms are likely to pursue these unfortunate global maxima. Because of the peculiarities of mixture likelihoods, the most popular algorithm for maximising $\ell(\theta)$ is the *expectation maximisation* (EM, [DLR77]) algorithm, an iterative algorithm specialised for dealing with log-likelihoods of latent variable models. The EM algorithm is usually preferred to more generic gradient-based optimisation algorithms [XJ96]. In a nutshell, at each iteration, the EM algorithm clusters the data using Equation (2), and then computes the mean and covariance of each cluster. This iterative scheme is related to another popular clustering algorithm known as k-means (the close relationship between the two algorithms is detailed in [Bis06, Section 9]). A key issue when using the EM algorithm for a Gaussian mixture is the choice of initialisation. Indeed, a poor choice may lead to degenerate solutions, extremely slow convergence, or poor local optima (see [BC15] and references therein). We will see in this paper that good initial points can be obtained by using another estimation method called the *method of moments* (as was previously noted by [RGL17] in a context of mixtures of multivariate Bernoulli distributions).

The *method of moments* is a general alternative to maximum likelihood. The idea is to choose several functions $g_1 : \mathbb{R}^m \longrightarrow \mathbb{R}^{q_1}, ..., g_d : \mathbb{R}^m \longrightarrow \mathbb{R}^{q_d}$ called *moments*, and to find θ by attempting to solve the system of equations

$$\begin{cases} \mathbb{E}_{x\sim p_{\text{data}}}[g_1(x)] = \mathbb{E}_{x\sim p_\theta}[g_1(x)] \\ \qquad\qquad \dots \\ \mathbb{E}_{x\sim p_{\text{data}}}[g_d(x)] = \mathbb{E}_{x\sim p_\theta}[g_d(x)]. \end{cases} \tag{4}$$

Of course, since p_{data} is unknown, solving (4) is not feasible. However, one may replace the expected moments by empirical versions, and solve instead

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} g_1(x_i) = \mathbb{E}_{x\sim p_\theta}[g_1(x)] \\ \qquad\qquad \dots \\ \frac{1}{n}\sum_{i=1}^{n} g_d(x_i) = \mathbb{E}_{x\sim p_\theta}[g_d(x)]. \end{cases} \tag{5}$$

A very simple example of this, in the univariate $m = 1$ case, when $g_1(x) = x$, and $g_2(x) = x^2$. Then, solving (4) will ensure that the distributions of the model p_θ and the data p_{data} have the same mean and variance. However, many very different distributions have identical mean and variance! A natural refinement of the previous idea is to consider also higher-order moments $g_3(x) = x^3, g_4(x) = x^4, \dots$. This will considerably improve the estimates found using the method of moments. This approach was pioneered by [Pea94] for learning univariate Gaussian mixtures. In the more general multivariate case $m > 1$, following [HK13], the moments chosen can be tensor products, as we detail in the next section in case of a Gaussian mixture with spherical covariances.

3. Learning structure from tensor decomposition

In this section, we describe the moment tensors revealing the structure of spherical Gaussian mixtures and how it can be decomposed using standard linear algebra operations.

Let $\mathbf{X} = (X_1, \dots, X_m)$ be a set of variables. The ring of polynomials in $\mathbf{X}$ with coefficients in $\mathbb{C}$ is denoted $\mathbb{C}[\mathbf{X}]$. The space of homogeneous polynomials of degree $d \in \mathbb{N}$ is denoted $\mathbb{C}[\mathbf{X}]_d$. We recall that a symmetric tensor T of order d (with real coefficients) can be represented by an homogeneous polynomial of degree d in the variables $\mathbf{X}$ of the form

$$T(\mathbf{X}) = \sum_{|\alpha|=d} T_\alpha \binom{d}{\alpha} \mathbf{X}^\alpha$$

where $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^m$, $|\alpha| = \alpha_1 + \dots + \alpha_m = d$, $T_\alpha \in \mathbb{R}$, $\binom{d}{\alpha} = \frac{d!}{\alpha_1! \cdots \alpha_m!}$, $\mathbf{X}^\alpha = X_1^{\alpha_1} \cdots X_m^{\alpha_m}$.

A decomposition of T as a sum of d^{th} power of linear forms is of the form

$$T(\mathbf{X}) = \sum_{i=1}^{r} \omega_i (\xi_i \cdot \mathbf{X})^d \tag{6}$$

where $\xi_i = (\xi_{i,1}, \dots, \xi_{i,m}) \in \mathbb{C}^m$ and $(\xi_i \cdot \mathbf{X}) = \sum_{j=1}^{m} \xi_{i,j} X_j$. When r is the minimal number of terms in such a decomposition, it is called the rank of T and the decomposition is called a rank decomposition (or a Waring decomposition) of $T(\mathbf{X})$.

We say that the decomposition is unique if the lines spanned by $\xi_1, \dots, \xi_r$ form a unique set of lines with no repetition. In this case, the decomposition of T is unique after normalisation of the vectors ξ_i up to permutation (and sign change when d is even). A tensor T with a unique decomposition is called an *identifiable* tensor. Then the Waring decompositions of T are of the form $T(\mathbf{X}) = \sum_{i=1}^{r} \omega_i \lambda_i^{-d} (\lambda_i \, \xi_i \cdot \mathbf{X})^d$ for $\lambda_i \neq 0$, $i \in [r]$.

Given a random variable $x \in \mathbb{R}^m$, its moments are $T_\alpha = \mathbb{E}[x_1^{\alpha_1} \cdots x_m^{\alpha_m}]$ for $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{N}^m$. The symmetric tensor of all moments of order d of x is

$$\mathbb{E}[(x \cdot \mathbf{X})^d] = \sum_{|\alpha|=d} \mathbb{E}[x_1^{\alpha_1} \cdots x_m^{\alpha_m}] \binom{d}{\alpha} \mathbf{X}^\alpha.$$

3.1. The structure of the moment tensor

We aim at recovering the hidden structure a random variable, from the decomposition of its d^{th} order moment tensor. This is possible in some circumstances, that we detail hereafter.

Assumption 3.1. *The random variable $x \in \mathbb{R}^m$ is a mixture of spherical Gaussians of probability density (1) with parameters $\theta = (\omega_1, ..., \omega_r, \mu_1, ..., \mu_r, \sigma_1^2 I_m, , ..., \sigma_r^2 I_m)$ such that $r \leq m$.*

Theorem 3.2 ([HK13])**.** *Under the previous assumption, let*

- *$\tilde{\sigma}^2$ be the smallest eigenvalue of $\mathbb{E}[(x - \mathbb{E}[x]) \otimes (x - \mathbb{E}[x])]$ and v a corresponding unit eigenvector,*
- *$M_1(\mathbf{X}) = \mathbb{E}[(x \cdot \mathbf{X})(v \cdot (x - \mathbb{E}[x]))^2]$,*
- *$M_2(\mathbf{X}) = \mathbb{E}[(x \cdot \mathbf{X})^2] - \tilde{\sigma}^2 \|\mathbf{X}\|^2$,*
- *$M_3(\mathbf{X}) = \mathbb{E}[(x \cdot \mathbf{X})^3] - 3\,\|\mathbf{X}\|^2 M_1(\mathbf{X})$.*

Then $\tilde{\sigma}^2 = \sum_{i=1}^r \omega_i\,\sigma_i^2$ and

$$M_1(\mathbf{X}) = \sum_{i=1}^r \omega_i\,\sigma_i^2\,(\mu_i \cdot \mathbf{X}), \quad M_2(\mathbf{X}) = \sum_{i=1}^r \omega_i\,(\mu_i \cdot \mathbf{X})^2, \quad M_3(\mathbf{X}) = \sum_{i=1}^r \omega_i\,(\mu_i \cdot \mathbf{X})^3. \tag{7}$$

To analyse the properties of the decomposition (7), we introduce the apolar product on tensors: For two homogeneous polynomials $p(\mathbf{X}) = \sum_{|\alpha|=d} \binom{d}{\alpha} p_\alpha \mathbf{X}^\alpha$ and $q(\mathbf{X}) = \sum_{|\alpha|=d} \binom{d}{\alpha} q_\alpha \mathbf{X}^\alpha$ of degree d, in $\mathbb{C}[\mathbf{X}]_d$, their apolar product is

$$\langle p, q \rangle_d := \sum_{|\alpha|=d} \binom{d}{\alpha} \bar{p}_\alpha q_\alpha.$$

The apolar norm of p is $||p||_d = \sqrt{\langle p, p \rangle_d} = \sqrt{\sum_{|\alpha|=d} \binom{d}{\alpha} \bar{p}_\alpha p_\alpha}$. The apolar product is invariant by a linear change of variables of the unitary group U_m: $\forall u \in U_m, \langle p(u\,\mathbf{X}), q(u\,\mathbf{X}) \rangle_d = \langle p(\mathbf{X}), q(\mathbf{X}) \rangle_d$.

It also satisfies the following properties. For $v \in \mathbb{C}^m$, $v(\mathbf{X})^d = (v \cdot \mathbf{X})^d = (v_1 X_1 + \cdots + v_m X_m)^d, p \in \mathbb{C}[\mathbf{X}]_d, q \in \mathbb{C}[\mathbf{X}]_{d-1}$, we have :

- $\langle (v \cdot \mathbf{X})^d, p \rangle_d = p(\bar{v})$,
- $\langle p, X_i q \rangle_d = \frac{1}{d} \langle \partial_{X_i} p, q \rangle_{d-1}$.

For an homogeneous polynomial T of degree $d \in \mathbb{N}$ (or equivalently a symmetric tensor of order d), we define the *Hankel* operator of T in degree $k \leq d$ as the map

$$H_T^{k,d-k} : p \in \mathbb{C}[\mathbf{X}]_{d-k} \mapsto [\langle T, \mathbf{X}^\alpha\, p \rangle_d]_{|\alpha|=k} \in \mathbb{C}^{s_k}$$

where $s_k = \binom{m+k-1}{k} = \dim \mathbb{C}[\mathbf{X}]_k$ is the number of monomials of degree k in $\mathbf{X}$. The matrix of $H_T^{k,d-k}$ in the basis $(\mathbf{X}^\beta)_{|\beta|=d-k}$ is

$$H_T^{k,d-k} = (\langle T, \mathbf{X}^{\alpha+\beta} \rangle_d)_{|\alpha|=k, |\beta|=d-k}.$$

From the properties of the apolar product, we see that $H_T^{1,d-1} : p \mapsto \frac{1}{d}[\langle \partial_{X_i} T, p \rangle_{d-1}]_{1 \leq i \leq m}$. For $\xi \in \mathbb{C}^m$ and $k \in \mathbb{N}$, let $\xi^{(k)} = (\xi^\alpha)_{|\alpha|=k}$. We also check that if $T = (\xi \cdot \mathbf{X})^d$ with $\xi \in \mathbb{C}^m$, then $H_{(\xi \cdot \mathbf{X})^d}^{k,d-k} = \bar{\xi}^{(k)} \otimes \bar{\xi}^{(d-k)}$ is of rank 1 and its image is spanned by the vector $\bar{\xi}^{(k)}$.

Proposition 3.3. *Assume that $r \leq m$, $w_i > 0$ for $i \in [r]$ and $\mu_1, \ldots, \mu_r \in \mathbb{R}^m$ are linearly independent. The symmetric tensor $M_3(\mathbf{X})$ is identifiable, of rank r and has a unique Waring decomposition satisfying* (7).

Proof. Assume that $M_3(\mathbf{X})$ has a decomposition of the form (7). Since the vector $\mu_1, \ldots, \mu_r$ are linearly independent, by a linear change of coordinates in Gl_m, we can further assume that $\mu_1 = e_1, \ldots, \mu_r = e_r$ are the first r vectors of the canonical basis of $\mathbb{R}^m$. In this coordinate system, $M_3(\mathbf{X}) = \sum_{i=1}^r X_i^3$ and the matrix $H_{M_3}^{1,2}$ in a convenient basis has a $r \times r$ identity block and zero elsewhere. Thus $H_{M_3}^{1,2}$ is of rank r. Its kernel of dimension $\frac{1}{2}\, m\,(m+1) - r$ is spanned by the polynomials $X_i X_j$ with $(i,j) \neq (k,k)$ for $k \in [r]$. The kernel of $H_{M_3}^{1,2}$ is thus the space of homogeneous polynomials of degree 2, vanishing at $e_1, \ldots, e_r \in \mathbb{R}^n$.

If $M_3(\mathbf{X})$ can be decomposed as $M_3(\mathbf{X}) = \sum_{i=1}^{r'} \omega_i' \,(\mu_i' \cdot \mathbf{X})^3$ with $\omega_i' \in \mathbb{C}$, $\mu_i' \in \mathbb{C}^m$ and $r' < r$, then $H_{M_3}^{1,2}$, as a sum of $r' < r$ matrices $\omega_i' H^{1,2}_{(\mu_i' \cdot \mathbf{X})^3}$ of rank 1, would be of rank smaller than $r' < r$, which is a contradiction. Thus a minimal decomposition of $M_3(\mathbf{X})$ is of length r and r is the rank of $M_3(\mathbf{X})$.

Let us show that the decomposition (7) of $M_3(\mathbf{X})$ is unique up to a scaling of the vector μ_i, i.e. that $M_3(\mathbf{X})$ is identifiable. For any Waring decomposition $M_3(\mathbf{X}) = \sum_{i=1}^r \omega_i' \,(\mu_i' {\cdot} \mathbf{X})^3$, the vectors $\mu_1', \ldots, \mu_r'$ are linear independant, since μ_i' spans $\mathrm{im} H^{1,2}_{(\mu_i' \cdot \mathbf{X})^3}$ and $H_{M_3}^{1,2} = \sum_{i=1}^r \omega_i' H^{1,2}_{(\mu_i' \cdot \mathbf{X})^3}$ is of rank r. As $\mu_1', \ldots, \mu_r'$ can be transformed into $e_1, \ldots, e_r$ by a linear change of variables, $\ker H_{M_3}^{1,2}$ is also the vector space of homogeneous polynomials of degree 2, vanishing at $\mu_1', \ldots, \mu_r' \in \mathbb{C}^m$. Therefore, the set of $\{\mu_1', \ldots, \mu_r'\}$ coincides, up to a scaling, with the set of points $\{\mu_1, \ldots, \mu_r\}$ of another Waring decomposition of $M_3(\mathbf{X}) = \sum_{i=1}^r \omega_i \,(\mu_i \cdot \mathbf{X})^3$. This shows that $M_3(\mathbf{X})$ is identifiable.

Therefore, a Waring decomposition of $M_3(\mathbf{X})$ is of the form $M_3(\mathbf{X}) = \sum_{i=1}^r \tilde{\omega}_i \,(\tilde{\mu}_i \cdot \mathbf{X})^3$ with $\tilde{\omega}_i = \lambda^{-3} \omega_i$, $\tilde{\mu}_i = \lambda_i \mu_i$ and $\lambda_i \neq 0$ for $i \in [r]$. As $\tilde{\mu}_1, \ldots, \tilde{\mu}_r$ are linearly independent, the homogeneous polynomials $(\tilde{\mu}_1 \cdot \mathbf{X})^2, \ldots, (\tilde{\mu}_r \cdot \mathbf{X})^2$ are also linearly independant in $\mathbb{C}[\mathbf{X}]_2$ (by a linear change of variables, they are equivalent to $X_1^2, \ldots, X_r^2$). Consequently, the relation

$$M_2(\mathbf{X}) = \sum_{i=1}^r \omega_i (\mu_i \cdot \mathbf{X})^2 = \sum_{i=1}^r \lambda_i \tilde{\omega}_i (\tilde{\mu}_i \cdot \mathbf{X})^2$$

defines uniquely $\lambda_1, \ldots, \lambda_r$, and $M_3(\mathbf{X})$ has a unique Waring decomposition, which satisfies the relations (7). □

Under Assumption 3.1, the hidden structure of the random variable x can thus be recovered using Algorithm 1.

This yields the parameters $\omega_i \in \mathbb{R}_+, \mu_i \in \mathbb{R}^m$, $\sigma_i \in \mathbb{R}_+$ for $i \in [r]$ of the Gaussian mixture x.

In the experimentation, the moments involved in the tensors M_i will be approximated by empirical moments and we will compute an approximate decomposition of the empirical moment tensor $\hat{M}_3(\mathbf{X})$.

3.2. Decomposition of identifiable tensors

We describe now an important step of the approach, which is computing a Waring decomposition of a tensor. In this section, we consider a tensor $T \in \mathbb{C}[\mathbf{X}]_d$ of order $d \in \mathbb{N}$ with a Waring decomposition of the form $T = \sum_{i=1}^r \omega_i \,(\xi_i \cdot \mathbf{X})^d$ with $\omega_i \in \mathbb{C}, \xi_i \in \mathbb{C}^m$, that we recover by linear algebra techniques, under some hypotheses.

Algorithm 1 Recovering the hidden structure of a Gaussian mixture

```
Input: The moment tensors $M_1(\mathbf{X}), M_2(\mathbf{X}), M_3(\mathbf{X})$.
  • Compute a Waring decomposition of $M_3(\mathbf{X})$ to get $\tilde{\omega}_i \in \mathbb{R}, \tilde{\mu}_i \in \mathbb{R}^m$, $i \in [r]$ such that $M_3(\mathbf{X}) = \sum_{i=1}^{r} \tilde{\omega}_i\, (\tilde{\mu}_i \cdot \mathbf{X})^3$.
  • Solve the system $\sum_{i=1}^{r} \tilde{\omega}_i\, (\tilde{\mu}_i \cdot \mathbf{X})^2 \lambda_i = M_2(\mathbf{X})$ to get $\lambda_i \in \mathbb{R}$ and $\omega_i = \lambda_i^3 \tilde{\omega}_i \in \mathbb{R}_+$, $\mu_i = \lambda_i^{-1} \tilde{\mu}_i \in \mathbb{R}^m$ such that $M_3(\mathbf{X}) = \sum_{i=1}^{r} \omega_i\, (\mu_i \cdot \mathbf{X})^3$ and $M_2(\mathbf{X}) = \sum_{i=1}^{r} \omega_i\, (\mu_i \cdot \mathbf{X})^2$.
  • Solve the system $\sum_{i=1}^{r} \omega_i (\mu_i \cdot \mathbf{X}) \sigma_i^2 = M_1(\mathbf{X})$ to get $\sigma_i^2 \in \mathbb{R}_+$.
Output: $\omega_i \in \mathbb{R}_+, \mu_i \in \mathbb{R}^n$, $\sigma_i^2 \in \mathbb{R}_+$ for $i \in [r]$.
```

Definition 3.4. The interpolation degree $\iota(\Xi)$ of $\Xi = \{\xi_1, \dots, \xi_r\} \subset \mathbb{C}^m$ is the smallest degree k of a family of homogenous interpolation polynomials $u_1, \dots, u_r \in \mathbb{C}[\mathbf{X}]_k$ at the points Ξ ($u_i(\xi_j) = \delta_{i,j}$ for $i, j \in [r]$).

For any $d \geq \iota(\Xi)$, there exists a family $(\tilde{u}_i)_{i \in [r]}$ of interpolation polynomials of degree d, obtained from an interpolation family $(u_i)_{i \in [r]}$ in degree $\iota(\Xi)$ as $\tilde{u}_i = \frac{(\lambda \cdot \mathbf{X})^{d-\iota(\Xi)}}{(\lambda \cdot \xi_i)^{d-\iota(\Xi)}} u_i$ for a generic $\lambda \in \mathbb{C}^m$ such that $\lambda \cdot \xi_i \neq 0$ for $i \in [r]$.

Notice that if the points $\Xi = \{\xi_1, \dots, \xi_r\}$ are linearly independent (and therefore $r \leq m$), then $\iota(\Xi) = 1$ since a family of linear forms interpolating Ξ can be constructed.

If $k \geq \iota(\Xi)$, then the evaluation map $\mathbf{e}_\Xi^{(k)} : p \in \mathbb{C}[\mathbf{X}]_k \mapsto (p(\xi_1), \dots, p(\xi_r)) \in \mathbb{C}^r$ is surjective. Its kernel is the space of homogeneous polynomials of degree k vanishing at Ξ. Any supplementary space admits a basis $u_1, \dots, u_r$, which is an interpolating family for Ξ in degree k. A property of the interpolation degree is the following:

Lemma 3.5. *For $k > \iota(\Xi)$, the common roots of $\ker \mathbf{e}_\Xi^{(k)}$ is the union $\cup_{i=1}^{r} \mathbb{C}\, \xi_i$ of lines spanned by $\xi_1, \dots, \xi_r \in \mathbb{C}^m$.*

Proof. As $\iota(\Xi) + 1$ is the Castelnuovo-Mumford regularity of the vanishing ideal $I(\Xi) = \{p \in \mathbb{C}[\mathbf{X}] \mid p$ homogeneous, $p(\xi) = 0$ for $\xi \in \Xi\}$ [Eis05][Ch.4], it is generated in degree $k > \iota(\Xi)$ and the common roots of $\ker \mathbf{e}_\Xi^{(k)} = I(\Xi)_k$ is $\cup_{i=1}^{r} \mathbb{C}\, \xi_i$. □

Hereafter, we show that tensors T such that $\operatorname{rank} H_T^{k,d-k} = r$ for $k > \iota(\Xi) + 1$ are identifiable and we describe a numerically robust algorithm to compute their Waring decomposition.

Let $U = (U_{\alpha,j})_{|\alpha|=k, j \in [r]} \in \mathbb{C}^{s_k \times r}$ be such that $\operatorname{im} U = \operatorname{im} H_T^{k,d-k}$ and $U_i = (U_{e_i+\alpha,j})_{|\alpha|=k-1, j \in [r]}$ be the submatrices of U with the rows indexed by the monomials divisible by X_i for $i \in [m]$.

Theorem 3.6. *Let $T \in \mathbb{C}[\mathbf{X}]_d$ with a decomposition $T = \sum_{i=1}^{r} \omega_i\, (\xi_i \cdot \mathbf{X})^d$ with $\omega_i \in \mathbb{C}$ and $\xi_i = (\xi_{i,1}, \dots, \xi_{i,n}) \in \mathbb{C}^m$ such that $\operatorname{rank} H_T^{k,d-k} = r$ for some $k \in [\iota(\xi_1, \dots, \xi_r) + 1, d]$. Then T is identifiable of rank r and there exist invertible matrices $E \in \mathbb{C}^{s_k \times s_k}$, $F \in \mathbb{C}^{r \times r}$ such that*

$$E^t\, U_i\, F = \begin{bmatrix} \Delta_i \\ 0 \end{bmatrix} \tag{8}$$

with $\Delta_i = \operatorname{diag}(\bar{\xi}_{1,i}, \dots, \bar{\xi}_{r,i})$ for $i \in [m]$. For any pair (E, F), which diagonalises simultaneously $[U_1, \dots, U_m]$ as in (8), there exist unique $\omega'_1, \dots, \omega'_r \in \mathbb{C}$ such that $T = \sum_{i=1}^{r} \omega'_i\, (\xi'_i \cdot \mathbf{X})^d$ with $\bar{\xi}'_i = ((\Delta_1)_{i,i}, \dots, (\Delta_m)_{i,i})$.

Proof. From the decomposition of T, we have for $k \le d$ that

$$H_T^{k,d-k} = \sum_{i=1}^{r} \omega_i \, \bar{\xi}_i^{(k)} \otimes \bar{\xi}_i^{(d-k)}$$

is a linear combination of r Hankel matrices $\bar{\xi}_i^{(k)} \otimes \bar{\xi}_i^{(d-k)}$ of rank 1. If T is of rank $r' < r$, then using its decomposition of rank r', $H_T^{k,d-k}$ would be of rank $\le r' < r$, which is a contradiction. This shows that T is of rank r.

As rank $H_T^{k,d-k} = r$, we deduce that the image of $H_T^{k,d-k}$ is spanned by $\bar{\xi}_1^{(k)}, \ldots, \bar{\xi}_r^{(k)}$ and there exists an invertible matrix $F \in \mathbb{C}^{r\times r}$ such that

$$U\,F = [\bar{\xi}_1^{(k)}, \ldots, \bar{\xi}_r^{(k)}]$$

For any polynomial $p \in \mathbb{C}[\mathbf{X}]_k$, which coefficient vector in the monomial basis $(\mathbf{X}^\alpha)_{|\alpha|=k}$ is denoted $[p]$, we have $[p]^t U F = [p(\bar{\xi}_1), \ldots, p(\bar{\xi}_r)]^t$. This shows that $U^\perp = \{p \in \mathbb{C}[\mathbf{X}] \mid [p]^t U = 0\}$ is $\ker \mathbf{e}_{\bar{\Xi}}^{(k)}$. By Lemma 3.5 since $k \ge \iota(\bar{\Xi})$, the common roots of the homogeneous polynomials in $\ker \mathbf{e}_{\bar{\Xi}}^{(k)}$ are the scalar multiples of $\bar{\Xi}$. Consequently, the set of lines spanned by the vectors Ξ of a Waring decomposition of T is uniquely determined as the conjugate of the zero locus of $U^\perp \subset \mathbb{C}[\mathbf{X}]_k$ and T is identifiable.

For any $p \in \mathbb{C}[\mathbf{X}]_{k-1}$ represented by its coefficient vector $[p]$ in the monomial basis $(\mathbf{X}^\alpha)_{|\alpha|=k-1}$, we have

$$[p]^t U_i F = [x_i p]^t U F = [\bar{\xi}_{1,i}\, p(\bar{\xi}_1), \ldots, \bar{\xi}_{r,i}\, p(\bar{\xi}_r)]^t. \tag{9}$$

Let E be the coefficient matrix of a basis $u_1, \ldots, u_r, v_{r+1}, \ldots, v_{s_{k-1}}$ of $\mathbb{C}[\mathbf{X}]_{k-1}$, such that $u_1, \ldots, u_r$ is an interpolating family for $\bar{\Xi} = \{\bar{\xi}_1, \ldots, \bar{\xi}_r\}$ and $v_{r+1}, \ldots, v_{s_{k-1}}$ is a basis of $\ker \mathbf{e}_{\bar{\Xi}}^{(k-1)}$. The matrix E is invertible by construction, and we deduce from (9) that

$$E^t U_i F = \left[\begin{array}{c} \operatorname{diag}(\bar{\xi}_{1,i}, \ldots, \bar{\xi}_{r,i}) \\ 0 \end{array} \right].$$

Let us show conversely that for any pair of matrices (E', F'), which diagonalises simultaneously $[U_1, \ldots, U_m]$ as in (8) with $\Delta_i = \operatorname{diag}(\bar{\xi}'_{1,i}, \ldots, \bar{\xi}'_{r,i})$, there exist unique $\omega'_1, \ldots, \omega'_r \in \mathbb{C}$ such that $T = \sum_{i=1}^r \omega'_i \, (\xi'_i \cdot \mathbf{X})^d$.

Let $u'_1, \ldots, u'_r, v'_{r+1}, \ldots, v'_{s_{k-1}} \in \mathbb{C}[\mathbf{X}]$ be the polynomials corresponding to the columns of E'. Then for a generic $\lambda = (\lambda_1, \ldots, \lambda_r) \in \mathbb{C}^m$, we have

$$\begin{aligned}
\operatorname{diag}((\lambda\cdot\bar{\xi}'_1), \ldots, (\lambda\cdot\bar{\xi}'_r)) &= \sum_{i=1}^m \lambda_i [u'_1, \ldots, u'_r]^t U_i F' = \sum_{i=1}^m \lambda_i [u'_1, \ldots, u'_r]^t U_i F (F^{-1}F') \\
&= [(\lambda\cdot\bar{\xi}_j)\, u'_i(\bar{\xi}_j)]_{i,j\in[r]} F^{-1} F' \\
&= \operatorname{diag}((\lambda\cdot\bar{\xi}_1), \ldots, (\lambda\cdot\bar{\xi}_r))\, [u'_i(\bar{\xi}_j)]_{i,j\in[r]} F^{-1} F'.
\end{aligned}$$

As $\lambda \in \mathbb{C}^m$ is generic and $\lambda \cdot \bar{\xi}_i \neq 0$ for $i \in [r]$, we deduce that $\Delta = [u'_i(\bar{\xi}_j)]_{i,j\in[r]} F^{-1} F'$ is a diagonal and invertible matrix and that $\xi'_i = \bar{\Delta}_{i,i} \xi_i$ with $\Delta_{i,i} \neq 0$.

Then we have $(\xi'_i \cdot \mathbf{X})^d = \bar{\Delta}_{i,i}^d \, (\xi_i \cdot \mathbf{X})^d$ and $T = \sum_{i=1}^r \omega'_i \, (\xi'_i \cdot \mathbf{X})^d$ with $\omega'_i = \bar{\Delta}_{i,i}^{-d} \omega_i$, which concludes the proof of the theorem. □

This leads to Algorithm 2 to compute a Waring decomposition of an identifiable tensor T.

Algorithm 2 Decomposition of an identifiable tensor

```
Input: $T \in \mathbb{C}[\mathbf{X}]_d$, which admits a decomposition with $r$ points $\Xi = \{\xi_1, \dots, \xi_r\}$ and $k > \iota(\Xi)$.
    • Compute the Singular Value Decomposition of $H_T^{k,d-k} = U\,S\,V^t$;
    • Deduce the rank $r$ of $H_T^{k,d-k}$, take the first $r$ columns of $U$ and build the submatrices $U_i$ with rows indexed by the monomials $(X_i \mathbf{X}^\alpha)_{|\alpha|=k-1}$ for $i \in [n]$;
    • Compute a simultaneous diagonalisation of the pencil $[U_1 \dots, U_m]$ as $E^t U_i F = \begin{bmatrix} \mathrm{diag}(\bar{\xi}_{1,i}, \dots, \bar{\xi}_{r,i}) \\ 0 \end{bmatrix}$ and deduce the points $\xi_i = (\xi_{i,1}, \dots, \xi_{i,m}) \in \mathbb{C}^m$ for $i \in [r]$;
    • Compute the weights $\omega_1, \dots, \omega_r$ by solving the linear system $T = \sum_{i=1}^r \omega_i\, (\xi_i \cdot \mathbf{X})^d$;
Output: $\omega_i \in \mathbb{C}$, $\xi_i \in \mathbb{C}^m$ s.t. $T = \sum_{i=1}^r \omega_i\, (\xi_i \cdot \mathbf{X})^d$.
```

4. Numerical experimentations

The model used in this section is the Gaussian Mixture Model (GMM) with differing spherical covariance matrices. Recall that if $x = (x_1, \dots, x_n)$ is a sample of n independent observations from r multivariate Gaussian mixture with differing spherical covariance matrices of dimension m, and $h = (h_1, h_2, \dots, h_n)$ is the latent variable that determine the component from which the observation originates, then:

$$x_i \mid (h_i = k) \sim \mathcal{N}_m(\mu_k, \sigma_k^2 I_m) \text{ where,}$$

$$\Pr(h_i = k) = \omega_k, \text{ for } k \in [r], \text{ such that } \sum_{k=1}^{r} \omega_k = 1.$$

The aim of statistical inference is to find the unknown parametrs μ_k, σ_k^2 and w_k, for $k \in [r]$ from the data x. This can be done by finding the maximum likelihood estimation (MLE) i.e. finding the optimal maximum of the likelihood function associated to this model. The expectation maximisation algorithm (EM) [DLR77], usually used for finding MLEs, is an iterative algorithm in which the initialisation i.e. the initial estimation of the latent parameters is crucial, since various initialisations can lead to different local maxima of the likelihood function, consequently, yielding different clustering partition. Thus, in this section we compare the clustering results obtained by different initialisation of the EM algorithm against the initialisation by the method of moments through examples of simulated (subsection 4.1) and real (subsection 4.2) datasets. We fix a maximum of 100 iterations of the EM algorithm. The different initialisation considered in this section are the following:

- The k-means method [Mac67] according to the following strategy:
  The best partition obtained out of 50 runs of the k-means algorithm.

- The method of moments, where Algorithm 1 is applied to build the moments and Algorithm 2 is applied to the empirical moment tensor corresponding to $M_3(\mathbf{X})$ (see Theorem 3.2), with less than 5 Riemannian Newton iterations [KKM22] to reduce the distance between the empirical moment tensor and its decomposition.

- The Model-based hierarchical agglomerative clustering algorithm (MBHC) [VD00, Fra98].
- The emEM strategy [BCG03] as in [LIL+15] which makes 5 iterations for each of 50 short runs of EM, and follows the one which maximises the log-likelihood function by a long run of EM.

The k-means, MBHC and emEM are common strategies for initialising the EM algorithm for GMMs. The comparison among the different EM initialisation strategies is based on three measures: The Bayesian Information Criterion (BIC) [Sch78, FR98], the Adjusted Rand Index (ARI) [HA85], and the error rate (errorRate). The BIC is a penalized-likelihood criterion given by the following formula

$$\text{BIC} = -2\ell(\hat{\theta}) + \log(n)\nu,$$

where ℓ is the log-likelihood function , $\hat{\theta}$ is the MLE which maximises the log-likelihood function and ν is the number of the estimated parameters. This criterion measures the quality of the model such that for comparing models the one with the largest BIC value among the other models is the most fitted to the studied dataset. The ARI criterion measures the similarity between the estimated clustering obtained by the applied model and the exact true clustering. Its value is bounded between 0 and 1. The more this measure is close to 1 the more the estimated clustering is accurate. The error rate measure can be viewed as an alternative of the ARI. In fact this criterion measures the minimum error between the predicted clustering and the true clustering, and thus low error rate means high agreement between the estimated and the true clustering. The former criteria as well as the EM algorithm are used from the tools of the package `mclust` [SFMR16] in `R` programming language.

4.1. Simulation

We performed 100 simulations from each of the two models described in examples 4.1 and 4.2. We counted the instances where each of the considered initialising strategies for the EM could find throughout the 100 simulated data and among the other initialisation methods the largest BIC, the highest ARI, ARI$\geq$ 0.99 (as in this case the clustering obtained is the most accurate) and the lowest errorRate. The values of the BIC, ARI, errorRate and consumed time of the different considered initialisation strategies for one dataset sampled according to the model of Example 4.1 (resp. 4.2) are presented in Table 1 (resp. 3), and Figure 1 (resp. 2) shows a two-dimensional visualisation of the observations according to the first four features, the observations in the upper panels are labeled according to the actual clustering, while they are labeled in the lower panels according to the clustering obtained by the EM algorithm initialised by the method of moments. In order to have an estimation about the numerical stability of the obtained results, we repeat the same numerical experiment for each example 20 times and we compute the means (Table 2, 4) and the variances (values in parentheses in Table 2, 4) of the 20 percentages obtained of each of the BIC, ARI, ARI$\geq$ 0.99 and errorRate values for the different initialising strategies.

As we mentioned before the initialisation strategies considered in this comparison against the method of moments are common and have, in general, good numerical behavior. Nevertheless, we cannot expect all the initialisation strategies that exist for the EM algorithm to work well in all the cases [BCG03, MM10]. Hereafter, two examples are chosen in such a way to present some cases where the common initialisation strategies k-means, MBHC and emEM have some difficulties to provide a good initialisation to the EM algorithm for the GMMs with differing spherical covariance matrices, or in other words where the initialisation by the method of moments outperforms the other considered initialisations. For instance, we put in each of these two examples one cluster of small size (the blue cluster in Figure 1, the red cluster in Figure 2), we want to make the clusters overlap, since these initialisation strategies could misscluster the dataset if the clusters are intersecting. We notice that this choice of the mean vectors and the different variances in each of the two examples yields a dataset with the expected clustering characteristic.

Example 4.1. In the first simulation example, a multivariate dataset (m=6) of n=1000 observations generated with r=4 clusters according to the following parameters:

- The probability vector: $\omega = (0.2782, 0.0139, 0.3324, 0.3756)^T$.
- The mean vectors: $\mu_1 = (-5.0, -9.0, 8.0, 8.0, 2.0, 5.0)^T$, $\mu_2 = (-7.0, 6.0, -1.0, 6.0, -8.0, -10.0)^T$, $\mu_3 = (-4.0, -10.0, -5.0, 1.0, 5.0, 4.0)^T$, $\mu_4 = (-6.0, 6.0, 5.0, 4.0, -1.0, -1.0)^T$.
- The variances: $\sigma_1^2 = 1.5$, $\sigma_2^2 = 2.5$, $\sigma_3^2 = 5.0$, $\sigma_4^2 = 15.0$.

Table 1: Numerical results of one data set of Example 4.1

Method	BIC	ARI	errorRate	time(s)
em_km	-29590.48	0.8281	0.168	**0.045**
em_mom	**-29492.11**	**1.0**	**0.0**	0.547
em_mbhc	-29594.97	0.8574	0.099	0.287
em_emEM	-29593.18	0.8366	0.132	0.171

Table 2: Estimation of the stability of Example 4.1 results

Method	BIC	ARI	ARI ≥ 0.99	errorRate
em_km	38.35% (37.82)	47.6% (21.41)	48.85% (21.61)	47.6% (21.2)
em_mom	**74.8%** (41.01)	**88.75%** (15.36)	**83.4%** (18.36)	**88.60%** (14.46)
em_mbhc	10.75% (12.41)	15.9% (17.57)	15.55% (22.99)	15.9% (19.46)
em_emEM	7.3% (8.43)	14.5% (8.05)	12.6% (17.83)	14.95% (7.52)

Example 4.2. In the second simulation example, a multivariate dataset (m=5) of n=1000 observations generated with r=3 clusters according to the following parameters:

- The probability vector: $\omega = (0.0930, 0.2151, 0.6918)^T$.
- The mean vectors: $\mu_1 = (7.0, -4.0, -4.0, -6.0, -4.0)^T$, $\mu_2 = (2.0, -4.0, -6.0, -10.0, -3.0)^T$, $\mu_3 = (4.0, -4.0, -5.0, 6.0, 1.0)^T$.
- The variances: $\sigma_1^2 = 5.0$, $\sigma_2^2 = 10.0$, $\sigma_3^2 = 15.0$.

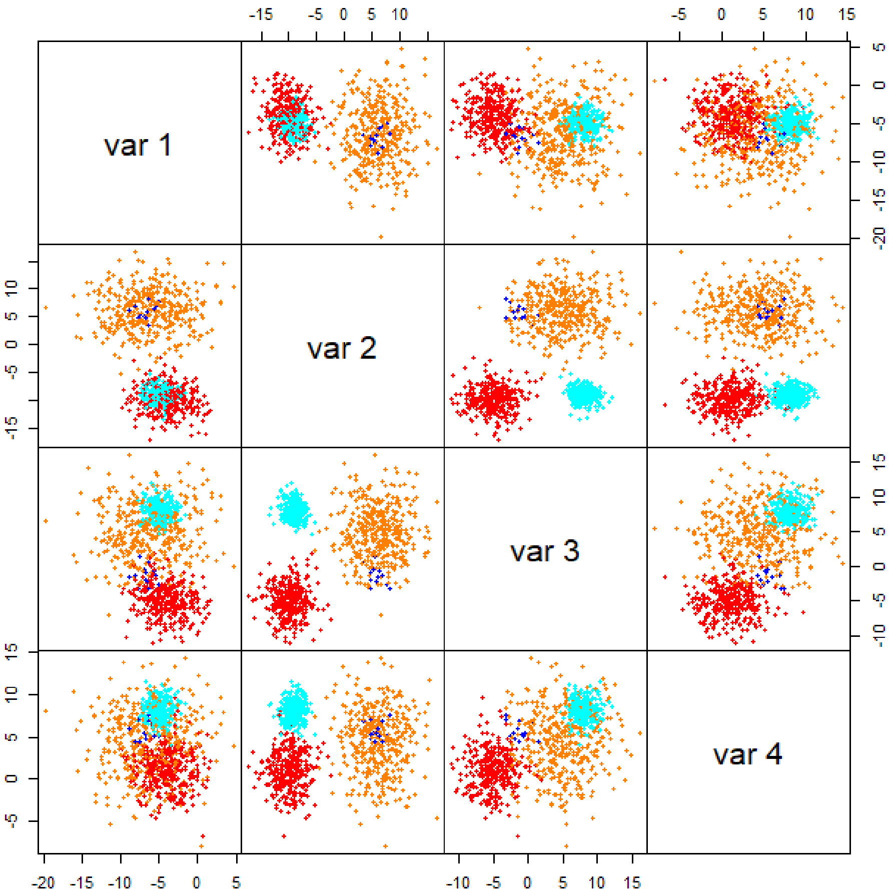

Figure 1: Scatterplot matrix for the sampled dataset of Example 4.1 projected onto the first four variables (features): upper panels show scatterplots for pairs of variables in the original clustering; lower panels show the clustering obtained by applying the EM algorithm initialised by the method of moments.

The Table 2, 4 show that in Example 4.1, 4.2 the best results among the considered initialising strategies are for the method of moments. In fact, in the former two tables we see that the method of moments found throughout the 100 simulated datasets, in average (by runing the numerical experiment 20 times), the largest BIC, highest ARI, ARI$\geq$ 0.99 and lowest errorRate among the other initialisation strategies in more instances than all the other considered initialisation method, implying in this context marked outperformance for the moments initialisation method. Note that the consumed time (see. Table 1, 3) tends to be higher in the method of moments than in the other initialisation strategies. This is expected since stochastic approaches (to which the methods k-means, MBHC and emEM belong) outperform the deterministic approaches (as the method of moments) in this term.

Table 3: Numerical results of one data set of Example 4.2

Method	BIC	ARI	errorRate	time(s)
em_km	-28360.30	0.4352	0.309	**0.051**
em_mom	**-28246.02**	**0.9498**	**0.03**	0.504
em_mbhc	-28358.67	0.3197	0.384	0.292
em_emEM	-28360.42	0.4408	0.296	0.141

Table 4: Estimation of the stability of Example 4.2 results

Method	BIC	ARI	ARI ≥ 0.99	errorRate
em_km	0.45% (0.576)	0.05% (0.05)	0.0% (0.0)	0.1%(0.095)
em_mom	**50.0%** (18.63)	**92.35%** (9.82)	**0.0%** (0.0)	**92.1%** (7.46)
em_mbhc	49.35% (19.82)	2.45% (3.63)	0.0% (0.0)	2.45% (2.58)
em_emEM	0.3% (0.326)	5.2% (4.48)	0.0% (0.0)	5.9% (5.36)

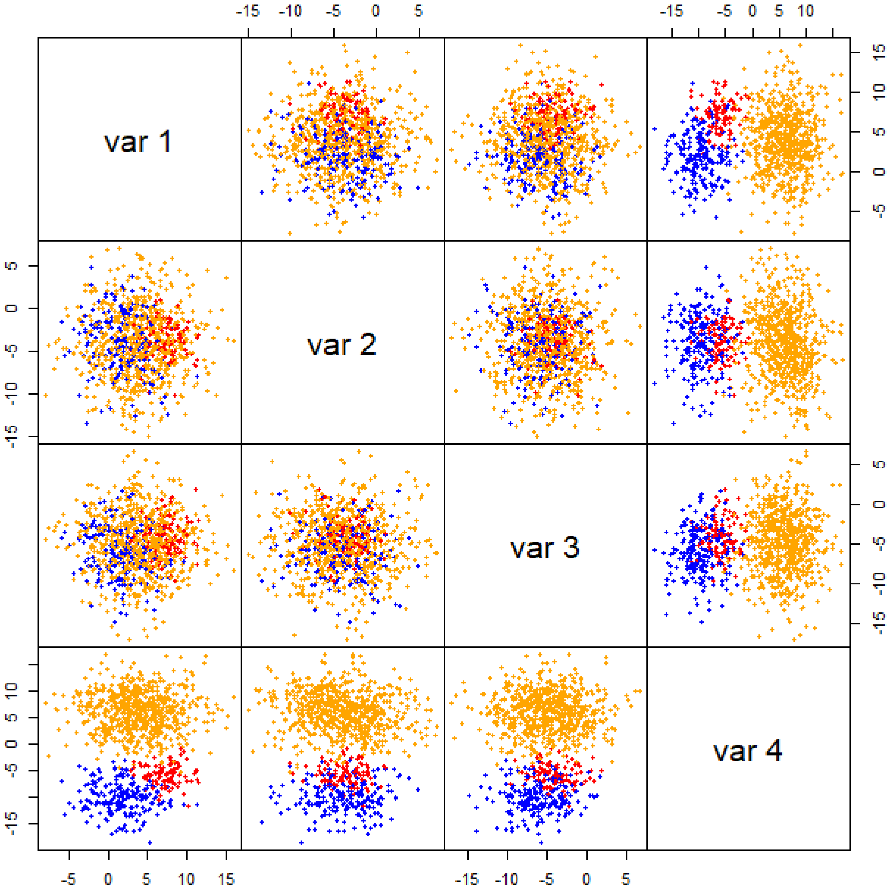

Figure 2: Scatterplot matrix for the sampled dataset of Example 4.2 projected onto the first four variables (features): upper panels show scatterplots for pairs of variables in the original clustering; lower panels show the clustering obtained by applying the EM algorithm initialised by the method of moments.

4.2. Real data

In this subsection we present four examples of real datasets, for which we know already their number of clusters, and we report the different BIC, ARI and errorRate values as well as the consumed time attained by the EM algorithm initialised by the different considered initialisation strategies and used with the GMM of different spherical covariance matrices. The explored real data are: The famous iris data [Fis36, DT17] widely used as an example of clustering to test the algorithms, Diabetes [RM79], olive oil [AM14], and MNIST [Den12].

Example 4.3 (Iris)**.** The iris dataset contains four physical measurements (length and width of sepals and petals) for 50 samples of three species of iris (setosa, virginica and versicolor). The number of features is $m = 4$ and the number of clusters is $r = 3$.
The four initialisation strategies yield the same BIC value. The ARI and the errorRate values are slightly better with the moment initialisation among the other considered initialisation strategies. On the other hand, the consumed time is clear higher in the moment method initialisation.

Table 5: Numerical results of Example 4.3

Method	BIC	ARI	errorRate	time(s)
em_km	-1227.6656	0.6199	0.167	**0.007**
em_mom	**-1227.6676**	**0.6410**	**0.153**	0.203
em_mbhc	-1227.6696	0.6199	0.167	**0.007**
em_emEM	-1227.6495	0.6302	0.160	0.045

Example 4.4 (Diabetes)**.** The Diabete dataset [RM79] contains three measurements: glucose, insulin and sspg; made on 145 non-obese adult patients classified into three types of diabetes: Normal, Overt, and Chemical. Herein, in this example $m = r = 3$. We apply the different initialisation strategies for the EM algorithm, the Table 6 shows the results.

Table 6: Numerical results of Example 4.4

Method	BIC	ARI	errorRate	time(s)
em_km	-5363.06	0.3371	0.289	**0.007**
em_mom	-5222.11	**0.6355**	**0.144**	0.380
em_mbhc	**-5221.32**	**0.6355**	**0.144**	0.008
em_emEM	-5221.33	0.6207	0.151	0.049

Despite the fact that k-means method is the fastest method in this example, the ARI and the BIC are noticeably lower than in the other methods. Concerning the method of moments, it succeeds to have quite similar scores to the other methods in this example, but with a bigger computation time.

Example 4.5 (Olive oil)**.** The olive oil data set contains the chemical composition (8 chemical properties) of 572 olive oils. They are derived from three different macro-areas in Italy (South, Sardinia and Centre North). The dataset contains nine regions from which the olive oils were taken in Italy. Thus we can cluster this dataset according to the macro-areas ($r = 3$) or the region ($r = 9$). As the number of features in this dataset is $m = 8$, we choose $r = 3$, so that the

Table 7: Numerical results of Example 4.5

Method	BIC	ARI	errorRate	time(s)
em_km	-10948.64	0.4018	0.262	**0.021**
em_mom	-10946.46	0.4532	0.210	0.508
em_mbhc	**-10625.59**	**0.5003**	**0.185**	0.080
em_emEM	-10948.72	0.4040	0.260	0.087

condition $r \leq m$ for the method of moment is verified.
The results show that the MBHC initialisation strategy yields the largest BIC, the highest ARI and the lowest errorRate values among the other initialisation strategies. Nevertheless, the initialisation by the moment method comes in second position after the MBHC strategy in terms of the BIC, ARI and errorRate values, while the K-means and the emEM initialisation strategies attain almost the same values of the previously mentionned criteria.

This shows that for these datasets which are not well fitted by the mixture of spherical Gaussians, the moment method can still give good initialisations for the EM algorithm, in comparision with the common initialisation strategies.

Example 4.6 (MNIST digit image database)**.** The MNIST digit image database [Den12] is a large database that contains images of 28 × 28 pixels for handwritten digits (0 to 9). Each pixel contains an integer between 0 and 255 that represents the grayscale levels. The number of features is $28 \times 28 = 784$. We choose the MNIST digit image dataset which contains 60000 images. We take a subset of this dataset that contains the images of label 0 or 1. The size of the subset is 12665 images. Since the number of features is quite large (784), and we aim to test a spherical Gaussian mixture model, a good practice in this case is to apply one of the dimensionality reduction strategies. Roughly speaking, the dimensionality reduction strategies aim to reduce the number of features such that a high percentage of the information within the dataset is conserved. In other words, the performance in term of accuracy of the clustering methods will not be noticeably affected by this reduction, and on the other hand this will reduce considerably the time of computation. For this purpose, we choose to apply the Principal Component Analysis transformation (PCA) [F.R01, Jol11]. We conserve the first five variables given by this transformation (see Figure 3). The dataset that we consider in this example contains 12665 observations, the number of clusters is $r = 2$, and the number of features is $m = 5$. We apply the different initialisation strategies and we report the results in Table 8.
As we can see, the results given by the method of moments in Table 8 are very satisfactory in comparison with the other initialisation strategies with ARI= 0.9308. In particular, the method of moments clearly outperforms MBHC method in this regard, in term of accuracy and the time of computation. In fact, the MBHC takes 543.4 seconds without reaching a *good* ARI score.

Table 8: Numerical results of Example 4.5

Method	BIC	ARI	errorRate	time(s)
em_km	-384977.3	0.9304	**0.017**	**0.537**
em_mom	-384978.2	**0.9308**	**0.017**	1.87
em_mbhc	**-382746.2**	0.2445	0.252	543.4
em_emEM	-384977.6	0.9301	0.0177655	1.80

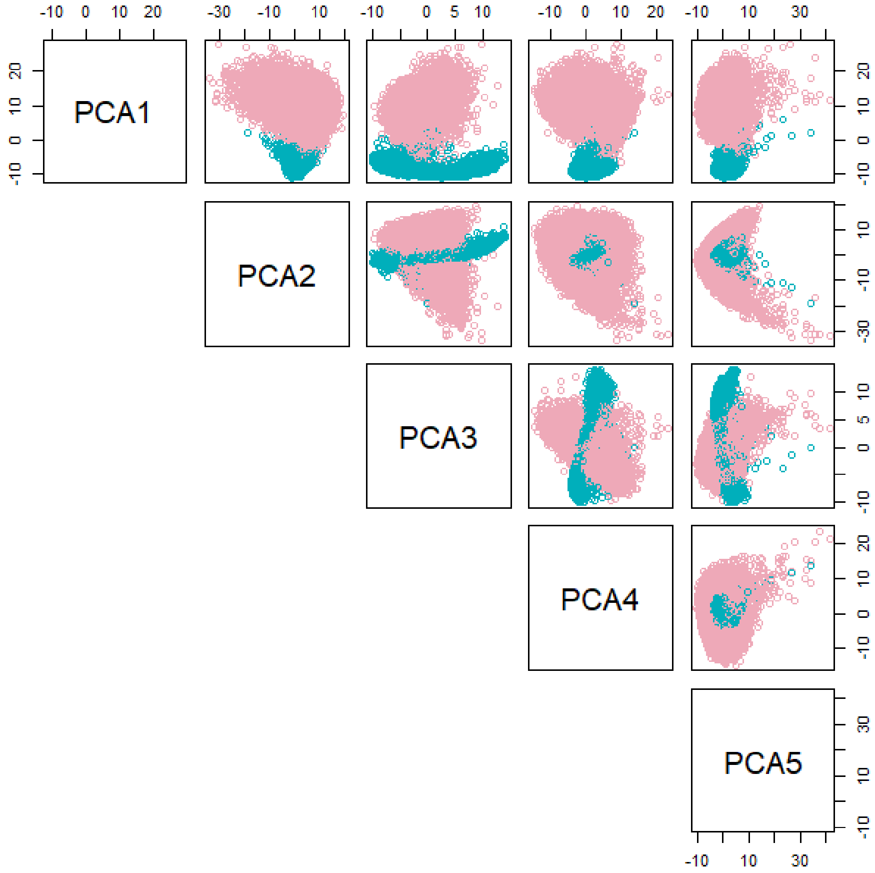

Figure 3: Scatterplot for pairs of variables: upper panels show the first five features obtained by applying the PCA transformation on the dataset of Example 4.6. The graphs points marked according to the true two classes 0 and 1.

This example sheds some light on the performance of the method of moments. The large number of samples (in this example equal to 12665) does not have a high impact on the computation time, which is not the case, for the MBHC method, where this factor increases significantly its computation time. Moreover, it is true that a large number of features could have a negative impact on the computation time of the method of moments, but it is not a sever limitation since as we saw in this example, this can be efficiently remedied by applying one of the dimensionality reduction techniques. In this regard, some recent work [PKK22] studies how the computation complexity of the moment method can be reduced while conserving its desirable high accuracy property. Conducting more research in this direction, we believe that the method of moments will have more sophisticated and competitive (in term of computation time) developments in the future.

5. Conclusion

In the context of unsupervised machine learning, the type of models to be recovered plays an important role. For Gaussian mixture models, where iterative methods such as Expectation Maximisation algorithms are applied, the choice of the initialisation is also crucial to recover an accurate model of a given dataset. We demonstrated in the experimentation that tensor decomposition techniques can provide a good initial point for the EM algorithm, and that the moment tensor method outperforms the other state-of-the-art strategies, when datasets are well represented by spherical Gaussian mixture models. For that purpose, we presented a new tensor decomposition algorithm adapted to the decomposition of identifiable tensors with low interpolation degree, which applies to a 3^{rd} order moment tensors associated to the data distribution as we have shown.

Acknowledgement. We would like to thank the anonymous reviewers for their valuable remarks that helped us improving the paper.

References

[AGH$^+$14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AM14] Adelchi Azzalini and Giovanna Menardi. Clustering via nonparametric density estimation: The R package pdfcluster. *Journal of Statistical Software, Articles*, 57(11):1–26, 2014.

[BBCM13] Alessandra Bernardi, Jérome Brachat, Pierre Comon, and Bernard Mourrain. General tensor decomposition, moment matrices and applications. *Journal of Symbolic Computation*, 52:51–71, May 2013.

[BC15] Jean-Patrick Baudry and Gilles Celeux. Em for mixtures. *Statistics and computing*, 25(4):713–726, 2015.

[BCG03] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Comput. Stat. Data Anal.*, 41(3–4):561–575, January 2003.

[BCMR19] Charles Bouveyron, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

[BCMT10] Jerome Brachat, Pierre Comon, Bernard Mourrain, and Elias Tsigaridas. Symmetric tensor decomposition. *Linear Algebra and its Applications*, 433(11-12):1851–1872, December 2010.

[BDHM17] Alessandra Bernardi, Noah S. Daleo, Jonathan D. Hauenstein, and Bernard Mourrain. Tensor decomposition and homotopy continuation. *Differential Geometry and its Applications*, 55:78–105, December 2017.

[Bis06] Christopher M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[BV18] Paul Breiding and Nick Vannieuwenhoven. A Riemannian trust region method for the canonical tensor rank approximation problem. *SIAM Journal on Optimization*, 28(3):2435–2465, 2018.

[CC70] J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika*, 35(3):283–319, Sep 1970.

[CHLZ12] Bilian Chen, Simai He, Zhening Li, and Shuzhong Zhang. Maximum block improvement and polynomial optimization. *SIAM Journal on Optimization*, 22(1):87–107, 6 2012.

[COV16] Luca Chiantini, Giorgio Ottaviani, and Nick Vannieuwenhoven. On generic identifiability of symmetric tensors of subgeneric rank. *Transactions of the American Mathematical Society*, 369(6):4021–4042, Nov 2016.

[DDL14] Ignat Domanov and Lieven De Lathauwer. Canonical Polyadic Decomposition of Third-Order Tensors: Reduction to Generalized Eigenvalue Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 35(2):636–660, January 2014.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[DL06] Lieven De Lathauwer. A Link between the Canonical Decomposition in Multilinear Algebra and Simultaneous Matrix Diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 28(3):642–666, January 2006.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[DT17] Dua Dheeru and E. Karra Taniskidou. UCI machine learning repository. https://archive.ics.uci.edu/ml/index.php, 2017.

[DZGL07] Marco Di Zio, Ugo Guarnera, and Orietta Luzi. Imputation through finite Gaussian mixture models. *Computational Statistics & Data Analysis*, 51(11):5305–5316, 2007.

[EHK15] Mike Espig, Wolfgang Hackbusch, and Aram Khachatryan. On the convergence of alternating least squares optimisation in tensor format representations. *arXiv preprint arXiv:1506.00062*, 2015.

[Eis05] David Eisenbud. *The Geometry of Syzygies: A Second Course in Commutative Algebra and Algebraic Geometry.* Springer, 2005. OCLC: 249751633.

[Fis36] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[F.R01] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[FR98] Chris Fraley and Adrian E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[FR02] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[Fra98] Chris Fraley. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[Har70] Richard Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[HBD18] Antoine Houdard, Charles Bouveyron, and Julie Delon. High-dimensional mixture models for unsupervised image denoising (HDMI). *SIAM Journal on Imaging Sciences*, 11(4):2815–2846, 2018.

[HH82] Chikio Hayashi and Fumi Hayashi. A new algorithm to solve parafac-model. *Behaviormetrika*, 9(11):49–60, Jan 1982.

[HK13] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 11–20, New York, NY, USA, January 2013. Association for Computing Machinery.

[HKM18] Jouhayna Harmouch, Houssam Khalil, and Bernard Mourrain. Structured low rank decomposition of multivariate Hankel matrices. *Linear Algebra and Applications*, 542:161–185, April 2018.

[HOOS19] Jonathan D. Hauenstein, Luke Oeding, Giorgio Ottaviani, and Andrew J. Sommese. Homotopy techniques for tensor decomposition and perfect identifiability. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2019(753):1–22, August 2019.

[HT96] Trevor Hastie and Robert Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):155–176, 1996.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JGKA19] Majid Janzamin, Rong Ge, Jean Kossaifi, and Anima Anandkumar. Spectral Learning on Matrices and Tensors. *Foundations and Trends® in Machine Learning*, 12(5-6):393–536, 2019.

[Jol11] Ian Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[KB09] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[KKM22] Rima Khouja, Houssam Khalil, and Bernard Mourrain. Riemannian newton optimization methods for the symmetric tensor approximation problem. *Linear Algebra and its Applications*, 637:175–211, 2022.

[KP09] Joe Kileel and João M. Pereira. Subspace power method for symmetric tensor decomposition and generalized PCA. 2019-12-09.

[LA14] Xavier Luciani and Laurent Albera. Canonical Polyadic Decomposition based on joint eigenvalue decomposition. *Chemometrics and Intelligent Laboratory Systems*, 132:152–167, March 2014.

[LC90] Lucien Le Cam. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171, 1990.

[LIL$^+$15] Rémi Lebret, Serge Iovleff, Florent Langrognet, Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software*, 67(6):1–29, 2015.

[LRA93] Sue E. Leurgans, Robert T. Ross, and R. B. Abel. A Decomposition for Three-Way Arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, October 1993.

[Mac67] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[MLR19] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

[MM10] Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4(none):80 – 116, 2010.

[NCNM20] TrungTin Nguyen, Faicel Chamroukhi, Hien D. Nguyen, and Geoffrey J. McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *arXiv preprint arXiv:2008.09787*, 2020.

[OO13] Luke Oeding and Giorgio Ottaviani. Eigenvectors of tensors and algorithms for Waring decomposition. *Journal of Symbolic Computation*, 54:9–35, July 2013.

[Paa99] Pentti Paatero. The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, 8(4):854–888, 1999.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[PKK22] João M Pereira, Joe Kileel, and Tamara G Kolda. Tensor moments of gaussian mixture models: Theory and applications. *arXiv preprint arXiv:2202.06930*, 2022.

[PTC13] Anh-Huy Phan, Petr Tichavský, and Andrzej Cichocki. Low complexity damped gauss–newton algorithms for candecomp/parafac. *SIAM Journal on Matrix Analysis and Applications*, 34(1):126–147, 2013.

[RGL17] Matteo Ruffini, Ricard Gavalda, and Esther Limón. Clustering patients with tensor decomposition. In *Machine Learning for Healthcare Conference*, pages 126–146. PMLR, 2017.

[RM79] Gerald M. Reaven and Rachel G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16:17–24, 1979.

[RSG17] Stephan Rabanser, Oleksandr Shchur, and Stephan Günnemann. Introduction to Tensor Decompositions and their Applications in Machine Learning. *arXiv:1711.10781 [cs, stat]*, November 2017. Comment: 13 pages, 12 figures.

[SBL13] Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in $rank-(l_r, l_r, 1)$ terms, and a new generalization. *SIAM Journal on Optimization*, 23(2):695–720, 2013.

[Sch78] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, July 1978.

[SFMR16] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016.

[SK90] Eugenio Sanchez and Bruce R. Kowalski. Tensorial resolution: A direct trilinear decomposition. *undefined*, 1990.

[SL10] Berkant Savas and Lek-Heng Lim. Quasi-newton methods on grassmannians and multilinear approximations of tensors. *SIAM Journal on Scientific Computing*, 32(6):3352–3393, 2010.

[SPMV13] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the Fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[Str83] Volker Strassen. Rank and optimal computation of generic tensors. *Linear Algebra and its Applications*, 52-53:645–685, July 1983.

[TB06] Giorgio Tomasi and Rasmus Bro. A comparison of algorithms for fitting the parafac model. *Comput. Stat. Data Anal.*, 50(7):1700–1734, April 2006.

[Usc12] André Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM Journal on Matrix Analysis and Applications*, 33(2):639–652, 2012.

[VD00] Shivakumar Vaithyanathan and Byron Dom. Model-based hierarchical clustering. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, page 599–608, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[VdV98] Aad W. Van der Vaart. *Asymptotic statistics.* Cambridge university press, 1998.

[XJ96] Lei Xu and Michael I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.