Tensor Decomposition For Learning Gaussian Mixtures From Moments

Rima Khouja

Élie Cartan de Lorraine (IECL) & Centre de Recherche en Automatique de Nancy (CRAN), France

Tensors in statistics, optimization and machine learning, AGATES, November 21st-25th, 2022, IM PAN, Poland

Contact: `rima.khouja@univ-lorraine.fr`

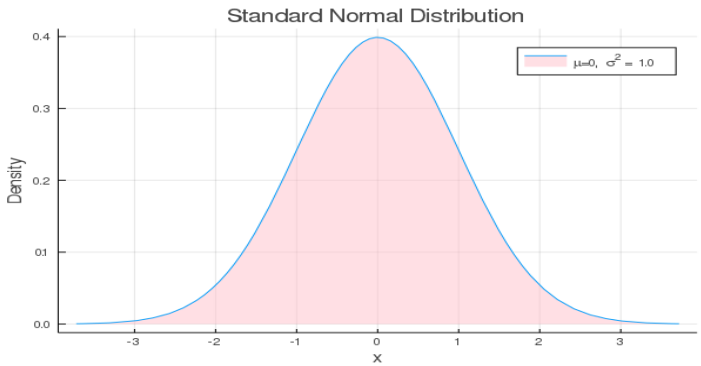
Univariate Gaussian distribution

- In probability theory, a Gaussian distribution is a continuous probability distribution for a real-valued random variable with a probability density function of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

  where μ is the mean of the distribution, σ is its standard deviation and σ^2 is its variance.
- A random variable x with Gaussian distribution of mean μ and standard deviation σ is called normally distributed and denoted by

$$x \sim \mathcal{N}(\mu, \sigma^2).$$



Multivariate Gaussian distribution

- The multivariate Gaussian distribution of a m-dimensional random vector $x = (x_1, \ldots, x_m)^T$ is denoted as follows:

$$x \sim \mathcal{N}(\mu, \Sigma),$$

with m-dimensional mean vector

$$\mu = E[x] = (E[x_1], \ldots, E[x_m])^T,$$

and $m \times m$ covariance matrix

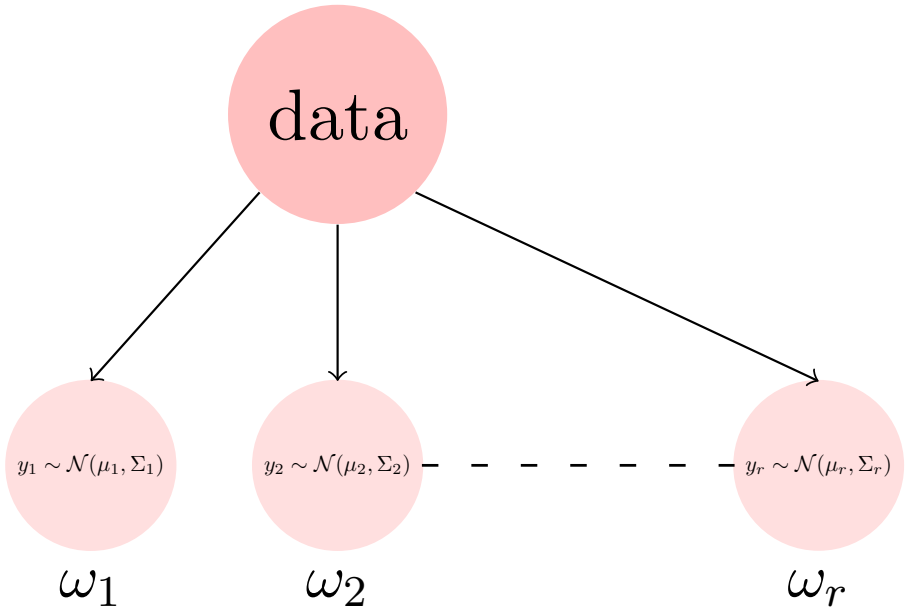
$$\Sigma_{i,j} = E[(x_i - \mu_i)(x_j - \mu_j)] = \mathrm{Cov}(x_i, x_j), \ \forall 1 \leq i, j \leq m.$$

Multivariate Gaussian mixtures

- Suppose that we have a m-dimensional data set $x_1, \ldots, x_n$ (i.e. n observations with m features), coming from a population composed of r homogeneous sub-population (clusters).
- Suppose that each of these sub-population can be modelled using a simple multivariate Gaussian distribution. It follows that the data set can be modelled using *mixture distributions*, in this case the *Gaussian mixture*

$$x \sim \sum_{j=1}^{r} \omega_j \mathcal{N}(\mu_j, \Sigma_j).$$

- The Gaussian mixture is parametrized 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 $j \in \{1, ..., r\}$.



- The probability density is $p_\theta(x) = \sum_{j=1}^{r} \omega_j \mathcal{N}(x|\mu_j, \Sigma_j)$.

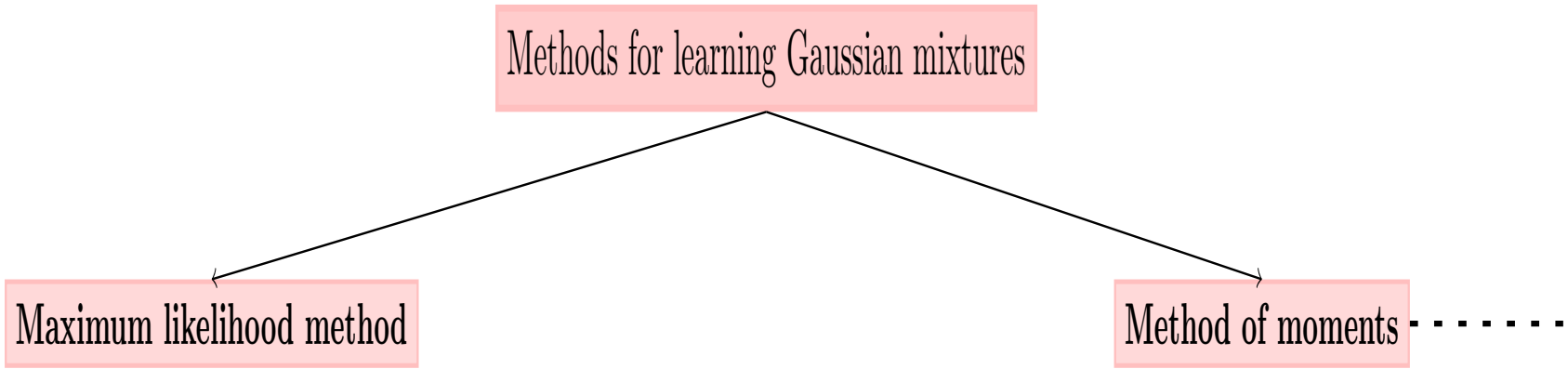
Clustering

- One of the most commom use of Gaussian mixtures is clustering.
- It consists of recovering homogeneous groups (clusters) among the data at hand.
- Within the context of Gaussian mixtures, each cluster corresponds to a single multivariate Gaussian distribution.
- If the parameter θ of the Gaussian mixture is known, then each point may be clustered using the posterior probabilities obtained via Bayes's rule:

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

Learning Gaussian mixtures

- Learning Gaussian mixtures is to estimate $\theta = (\omega_1, ..., \omega_r, \mu_1, ..., \mu_r, \Sigma_1, ..., \Sigma_r)$ based on the data at hand.
- Typically, $x_1, \ldots, x_n$ are assumed to be independent and identically distributed random variables with common density p_{data}.
- Herein, the problem of statistical estimation is to find some θ such that $p_\theta \approx p_{\text{data}}$.



Maximum likelihood method

- The maximum likelihood method consists of maximizing the *log-likelihood function*

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

- The function $\ell(\theta)$ can be seen as a measure of how likely the observed data is, according to the mixture model p_θ.
- Consequently, the *maximum likelihood estimate* (MLE) will be the value of θ that renders the observed data the likeliest.

The expectation maximization algorithm

- The most popular algorithm for maximizing $\ell(\theta)$ is the *expectation maximization* (EM) algorithm.
- Briefly, at each iteration, the EM algorithm clusters the data using

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

  and then computes the means and covariances of each cluster.
- The choice of the initial point when using the EM algorithm for a Gaussian mixture is crucial, in the sense that a poor choice may lead to degenerate solutions, extremely slow convergence, or poor local optima.
- In this case, using another estimation method such as the *method of moments* could provides a *good initial choice* for the EM algorithm ([BC15] and references therein).

The method of moments

- 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

$$\begin{cases} \mathbb{E}_{x\sim p_{\text{data}}}[g_1(x)] = \mathbb{E}_{x\sim p_\theta}[g_1(x)] \\ \dots \\ \mathbb{E}_{x\sim p_{\text{data}}}[g_d(x)] = \mathbb{E}_{x\sim p_\theta}[g_d(x)] \end{cases} \tag{1}$$

- Since p_{data} is unknown, solving (1) is not feasible. However, one may replace the expected moments by empirical versions, and solve instead

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n g_1(x_i) = \mathbb{E}_{x\sim p_\theta}[g_1(x)] \\ \dots \\ \frac{1}{n}\sum_{i=1}^n g_d(x_i) = \mathbb{E}_{x\sim p_\theta}[g_d(x)] \end{cases} \tag{2}$$

- **Example:** in the univariate case $m = 1$, when $g_1(x) = x$, and $g_2(x) = x^2$. Then, solving (1) will ensure that the distributions of the model p_θ and the data p_{data} have the same mean and variance.

- **Example:** in the univariate case $m = 1$, when $g_1(x) = x$, and $g_2(x) = x^2$. Then, solving (1) 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 $g_3(x) = x^3, g_4(x) = x^4, ...$

- In the more general multivariate case $m > 1$, the moments chosen can be tensor products.
- Given a random variable $x \in \mathbb{R}^m$, its moments of order d are $T_\alpha = \mathbb{E}[x_1^{\alpha_1} \cdots x_m^{\alpha_m}]$ for $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m$, $|\alpha| = d$.
- The symmetric tensor of all moments of order d of x is

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

- Herein, **the objective** is to use the decomposition of the symmetric tensor

$$T(\mathbf{X}) = \sum_{|\alpha|=d} \binom{d}{\alpha} \mathbb{E}[x^\alpha] \mathbf{X}^\alpha,$$

to recover the structure of the Gaussian mixture.

Symmetric tensors

- Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a set of variables. The space of homogeneous polynomials of degree $d \in \mathbb{N}$ is denoted $\mathbb{C}[\mathbf{X}]_d$.
- 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 $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^m$, $|\alpha| = \alpha_1 + \cdots + \alpha_m = d$, $T_\alpha \in \mathbb{R}$, $\binom{d}{\alpha} = \frac{d!}{\alpha_1! \cdots \alpha_m!}$, $\mathbf{X}^\alpha = X_1^{\alpha_1} \ldots X_m^{\alpha_m}$.

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

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

where $\xi_i = (\xi_{i,1}, \ldots, \xi_{i,m}) \in \mathbb{C}^m$ 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 $T(\mathbf{X})$.
- 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 $||p||_d = \sqrt{\langle p, p \rangle_d} = \sqrt{\sum_{|\alpha|=d} \binom{d}{\alpha} \bar{p}_\alpha p_\alpha}$.

Identifiability of symmetric tensors

- We say that the decomposition is unique if the lines spanned by $\xi_1, \ldots, \xi_r$ 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 $\lambda_i \neq 0$, $i \in [r]$.

Learning spherical Gaussian mixture from symmetric tensor decomposition

Assumption: The random variable $x \in \mathbb{R}^m$ is a mixture of spherical Gaussians 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 ([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*
$M_1(\mathbf{X}) = \sum_{i=1}^r \omega_i\, \sigma_i^2\, (\mu_i \cdot \mathbf{X})$, $M_2(\mathbf{X}) = \sum_{i=1}^r \omega_i\, (\mu_i \cdot \mathbf{X})^2$, $M_3(\mathbf{X}) = \sum_{i=1}^r \omega_i\, (\mu_i \cdot \mathbf{X})^3$.

Proposition

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 the Theorem.

Algorithm 1 Recovering the hidden structure of a Gaussian mixture

---

Input: The moment tensors $M_1(\mathbf{X}), M_2(\mathbf{X}), M_3(\mathbf{X})$.

1: 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 $M_3(\mathbf{X}) = \sum_{i=1}^r \tilde{\omega}_i \, (\tilde{\mu}_i \cdot \mathbf{X})^3$.

2: 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 $M_2(\mathbf{X}) = \sum_{i=1}^r \omega_i \, (\mu_i \cdot \mathbf{X})^2$.

3: 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

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

Definition

The *Hankel* operator of $T \in \mathbb{C}[\mathbf{X}]_d$ in degree $k \leq d$ is 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

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

- 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

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 $\xi_i = (\xi_{i,1}, \ldots, \xi_{i,n}) \in \mathbb{C}^m$ such that $\operatorname{rank} H_T^{k,d-k} = r$ for some $k \in [\iota(\xi_1, \ldots, \xi_r) + 1, d]$. 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

$$E^t U_i F = \begin{bmatrix} \Delta_i \\ 0 \end{bmatrix} \tag{3}$$

with $\Delta_i = \operatorname{diag}(\bar{\xi}_{1,i}, \ldots, \bar{\xi}_{r,i})$ for $i \in [m]$. For any pair (E, F), which diagonalises simultaneously $[U_1, \ldots, U_m]$ as in (3), there exist unique $\omega'_1, \ldots, \omega'_r \in \mathbb{C}$ such that $T = \sum_{i=1}^r \omega'_i (\xi'_i \cdot \mathbf{X})^d$ with $\bar{\xi}'_i = ((\Delta_1)_{i,i}, \ldots, (\Delta_m)_{i,i})$.

Algorithm 2 Decomposition of an identifiable tensor

Input: $T \in \mathbb{C}[\mathbf{X}]_d$, which admits a decomposition with r points $\Xi = \{\xi_1, \ldots, \xi_r\}$ and $k > \iota(\Xi)$.

1: Compute the Singular Value Decomposition of $H_T^{k,d-k} = U\,S\,V^t$;

2: 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 $i \in [m]$;

3: Compute a simultaneous diagonalisation of the pencil $[U_1 \ldots, U_m]$ as $E^t U_i F = \begin{bmatrix} \mathrm{diag}(\bar{\xi}_{1,i}, \ldots, \bar{\xi}_{r,i}) \\ 0 \end{bmatrix}$ and deduce the points $\xi_i = (\xi_{i,1}, \ldots, \xi_{i,m}) \in \mathbb{C}^m$ for $i \in [r]$;

4: Compute the weights $\omega_1, \ldots, \omega_r$ 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. $T = \sum_{i=1}^r \omega_i\,(\xi_i \cdot \mathbf{X})^d$.

Illustrative scheme

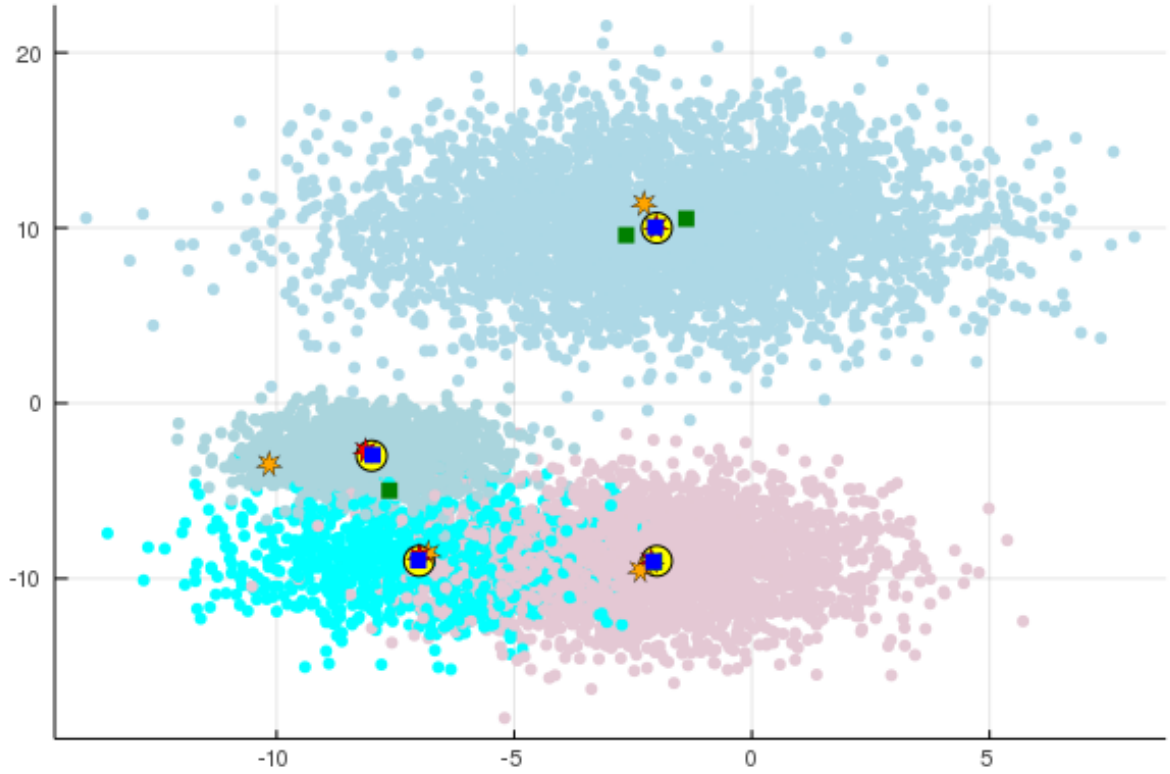


Figure 1: Yellow circle: the original means, Orange star: the method of moments, red star: symmetric tensor approximation method (RNE) [KKM22], blue square: EM with initial point the red star point, green square: EM with k-means initialization.

Simulation

Example 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: $\sigma_1^2 = 1.5$, $\sigma_2^2 = 2.5$, $\sigma_3^2 = 5.0$, $\sigma_4^2 = 15.0$.

Example 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: $\sigma_1^2 = 5.0$, $\sigma_2^2 = 10.0$, $\sigma_3^2 = 15.0$.

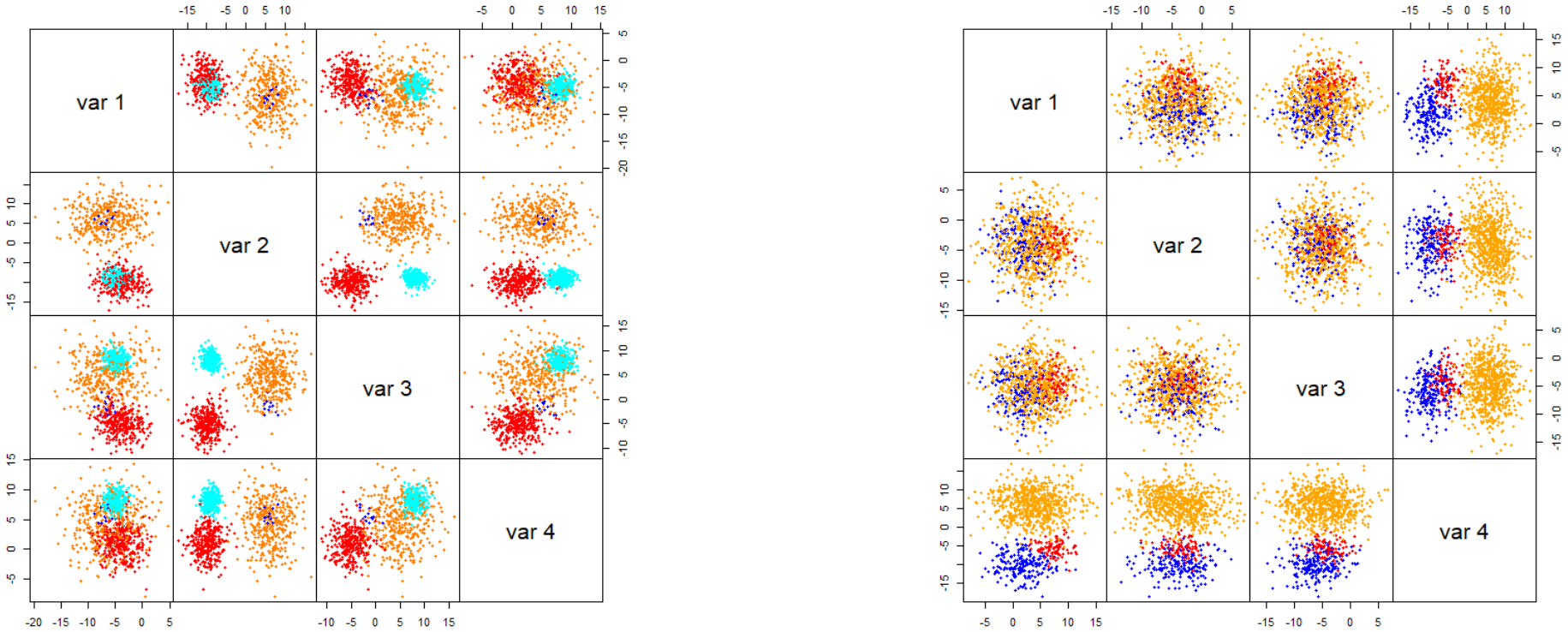


Figure 2: Scatterplot matrix for the sampled dataset of **Example 1** (resp. **2**) projected onto the first four features: upper (resp. lower) panels show scatterplots for pairs of variables in the original clustering (resp. the clustering obtained by applying the EM algorithm initialised by the method of moments).

- The comparison 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.
- 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 error rate measure can be viewed as an alternative of the ARI, it measures the minimum error between the predicted clustering and the true clustering.

Table 1: Estimation of the stability of **Example 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)

Table 2: Estimation of the stability of **Example 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)

Real data

Example1: 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 $n = 4$ and the number of clusters is $r = 3$.

Setosa

Virginica

Versicolor

Method	BIC	ARI	errorRate	time(s)
em_km	-1227.67	0.6199	0.167	0.007
em_mom	-1227.67	0.6410	0.153	0.203
em_mbhc	-1227.67	0.6199	0.167	0.007
em_emEM	-1227.67	0.6302	0.160	0.045

Example MNIST dataset

❑ The MNIST digit image database 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 **28×28=784**.

❑ 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, for instance the **Principal Component Analysis**.

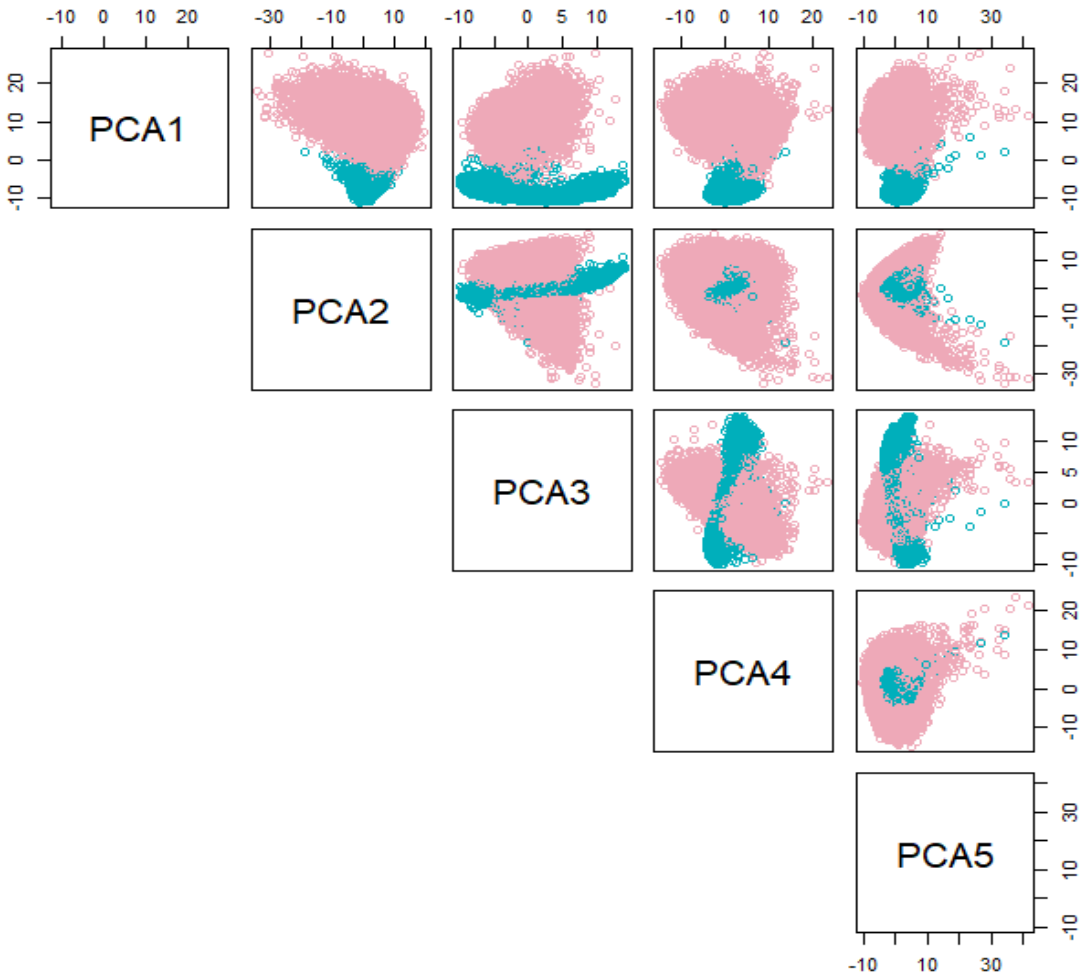


Figure 3: Scatterplot for pairs of variables: upper panels show the first five features obtained by applying the PCA transformation on the dataset. The graphs points marked according to the true two classes 0 and 1.

Table 3: Numerical results of the Example.

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

Conclusion

- We considered the method of moments to recover spherical Gaussian mixtures.
- We used the estimation given by the method of moments as an initial point to the EM algorithm.
- We demonstrated in the experimentations that tensor decomposition techniques ❶ can provide a good initial point for the EM algorithm, ❷ outperforming the other state-of-the-art strategies in term of accuracy, when datasets are well represented by spherical Gaussian mixture models.
- **Some theoretical results:**
  - We prove that symmetric tensors with interpolation degree strictly less than half their order are identifiable ☞ present an algorithm, based on simple linear algebra operations, to compute their decomposition.

Perspectives ⟹ ...

- Since the method of moments works well in term of accuracy, a straightforward question is to try to reduce its complexity → dimensionality reduction strategies and further approaches can be investigated.
- Extend the method of moments to the general covariance matrices case.

References I


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


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.


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

References II


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.


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.


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

For more details concerning this work:

R. Khouja, P-A. Mattei, B.Mourrain, *Tensor decomposition for learning Gaussian mixtures from moments*. Journal of Symbolic Computation, 2022.

Thank you for your attention!