ORIGINAL PAPER

Gaussian mixture model decomposition of multivariate signals

Gustav Zickert[1] · **Can Evren Yarman[2]**

Abstract

We propose a greedy variational method for decomposing a non-negative multivariate signal as a weighted sum of Gaussians, which, borrowing the terminology from statistics, we refer to as a Gaussian mixture model. Notably, our method has the following features: (1) It accepts multivariate signals, i.e., sampled multivariate functions, histograms, time series, images, etc., as input. (2) The method can handle general (i.e., ellipsoidal) Gaussians. (3) No prior assumption on the number of mixture components is needed. To the best of our knowledge, no previous method for Gaussian mixture model decomposition simultaneously enjoys all these features. We also prove an upper bound, which cannot be improved by a global constant, for the distance from any mode of a Gaussian mixture model to the set of corresponding means. For mixtures of spherical Gaussians with common variance σ^2, the bound takes the simple form $\sqrt{n}\sigma$. We evaluate our method on one- and two-dimensional signals. Finally, we discuss the relation between clustering and signal decomposition, and compare our method to the baseline expectation maximization algorithm.

Keywords Gaussian mixture model (GMM) · Matching pursuit (MP) · Variational greedy approximation · Electron microscopy

1 Introduction

Mixtures of Gaussians are often used in clustering to fit a probability distribution to some given sample points. In this work we are concerned with the related problem of approximating a non-negative but otherwise arbitrary signal by a sparse linear combination of potentially anisotropic Gaussians. Our interest in this problem stems mainly from its applications in transmission electron microscopy (TEM), where it is common to express the reconstructed 3D image as a linear combination of Gaussians [5–8]. Consequently, because the projection of a Gaussian is a Gaussian, TEM projection data can be treated as 2D images made up of linear combinations of Gaussians [14].

Methods for sparse decomposition of multivariate signals as Gaussian mixture models (GMMs) may be considered in two classes. The first class contains methods based on the expectation maximization algorithm that is commonly used for fitting a GMM to a point-cloud and adapt it to input data

in the form of multivariate signals [7,8]. The second class is the class of greedy variational methods [4,5,9]. The proposed method, which belongs to the latter class, is similar to—and is inspired by—both [9] and [5]. It is a continuously parameterized analogue of orthogonal matching pursuit where at each iteration the L^2-norm of the error is non-increasing. The significance of the proposed method, and what distinguishes it from earlier work on GMM decomposition of signals, is that

1. The resulting GMM may contain ellipsoidal Gaussians. This allows for a sparser representation than what could be achieved with a GMM only consisting of spherical Gaussians.
2. The number of Gaussians does not need to be set beforehand.

We are not aware of previous methods for GMM decomposition of signals that enjoys both of these properties at the same time.

We complement our algorithm with a theorem (Theorem 1) that upper bounds the distance from a local maximum of a GMM to the set of mean vectors. This provides theoretical support for our initialization of each new mean vector

✉ Gustav Zickert
gzickert@gmail.com

[1] Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

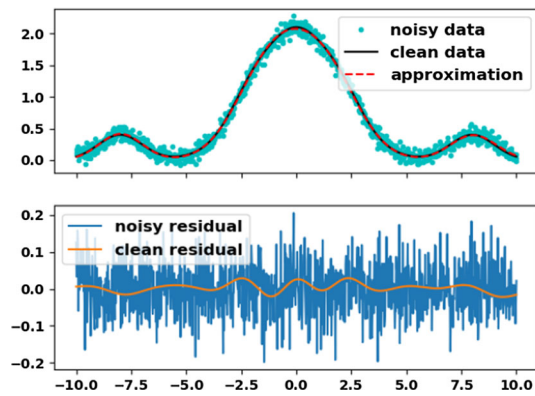[2] Schlumberger Cambridge Research, Cambridge, UK

Fig. 1 Results from Experiment 1

at a maximum of the residual. We remark that Theorem 1 could also be of interest in its own right. Indeed the number of modes of Gaussian mixtures has been investigated previously [1,3], but the authors of this paper are not aware of any existing quantitative bounds on the distance from a mode of a GMM to its mean vectors in the multivariate setting.

The rest of this paper is organized as follows. In Sect. 2 we define the GMM decomposition problem. Section 3 contains a description of the proposed algorithm together with numerical examples. Section 4 is devoted to theoretical questions; in particular, we state and prove the afore-mentioned upper bound. Finally, in Sect. 5 we provide a conclusion.

2 Problem statement

For $x_0 \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ symmetric non-negative definite we define $g(x_0, \Sigma)$ as the Gaussian density in n dimensions with mean vector x_0 and covariance matrix Σ^2, i.e.,

$$g(x_0, \Sigma)(x) := C_\Sigma \exp\left\{-\frac{1}{2}(x - x_0)^T \Sigma^{-2} (x - x_0)\right\},$$

where C_Σ is a normalizing factor, ensuring that $|g(x_0, \Sigma)|_1 = 1$. Further, by a GMM we mean a linear combination of the form[1]

$$\sum_{m=1}^{M} a_m g(x_m, \Sigma_m), \quad a_m > 0, M > 0.$$

The problem we consider is to construct an algorithm with the following properties. Given input in the form of a non-negative signal $d \in \mathbb{R}^{k_1 \times \cdots \times k_n}$, where k_i is the number of grid points along the i:th variable, the output should be a list of GMM parameters, i.e., it is a list $(a_m^*, x_m^*, \Sigma_m^*)_{m=1}^{M}$ of weights, mean vectors and square roots of covariance matrices. The output should be such that

1. the residual

$$r := d - \sum_{m=1}^{M} a_m^* g(x_m^*, \Sigma_m^*)$$

has a small L^2-norm. (We remark that any sufficiently regular non-negative function can be uniformly approximated arbitrarily well using GMMs; see Sect. 4.)

2. the approximation is sparse, i.e., the number of Gaussians M in the sum should ideally be as small as possible given the L^2-norm of the residual.

3 Proposed method

In each iteration of our algorithm, a new Gaussian is added to the GMM by a procedure that corresponds to one iteration of a continuously parameterized version of matching pursuit (MP), c.f. [10]: The starting guess x_0 for the mean vector is defined to be a global maximum of r', which is a smoothed version of the current residual r. Likewise Σ_0 is set to be a square root of the matrix of second-order moments of $\frac{r''}{|r''|_1}$, where r'' is r restricted to a neighborhood of x_0. The initial weight a_0 is given by projecting the Gaussian atom defined by (x_0, Σ_0) onto the residual, i.e., it is given as the global minimum of the convex objective $a \to |r - a g(x_0, \Sigma_0)|_2^2, a \geq 0$.

We found that this vanilla MP type of approach was by itself not capable of producing a good approximation. Because of this, we update the already obtained Gaussians in a way that bears some resemblance to the projection step of orthogonal MP (OMP) [13]: Starting from (a_0, x_0, Σ_0), the parameters defining the most recently added Gaussian are updated by minimizing the L^2-norm of the residual.[2] The final step of an outer iteration is to simultaneously adjust all Gaussians in the current GMM, again by minimizing the L^2-norm of the residual. We propose to use the L-BFGS-B [2] method with non-negativity constraints on the weights for all of the three afore-mentioned minimization problems. Our algorithm runs until some user-specified stopping criterion is met and is summarized in pseudo-code in Algorithm 1. Our implementation of Algorithm 1 is in Python. The source code and consequent examples below are available upon request.

We now assess the space and time complexity of the proposed method in terms of the final number M of Gaussians used and the dimension n of the input signal. Let N_i denote the number of parameters of i Gaussians, so $N_i = \mathcal{O}(in^2)$.

---

[1] We do not require the GMM to be normalized, i.e., we do not require that $\sum_{m=1}^{M} a_m = 1$.

[2] This step is not crucial for the performance of the algorithm, but was empirically found to improve the final data fit.
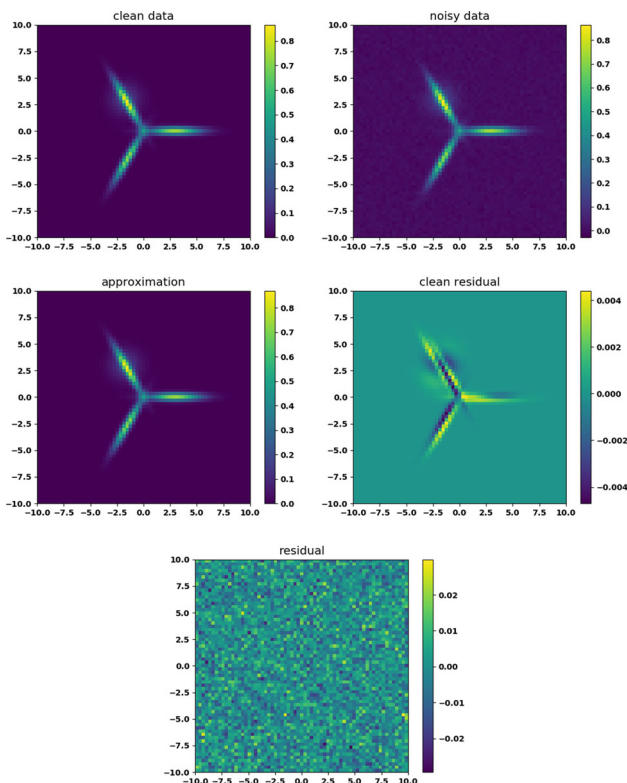
Fig. 2 Results from Experiment 2. Note the faint spherical Gaussian on the upper left "leg."
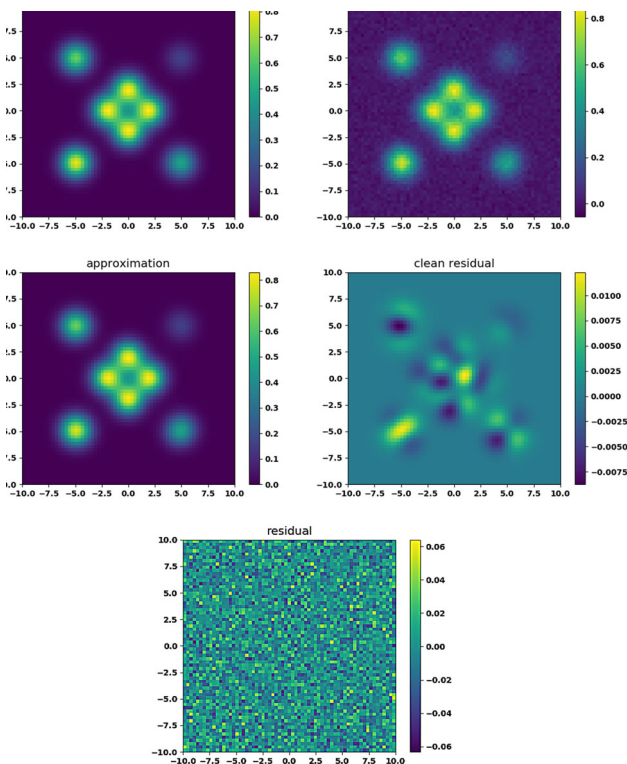
Fig. 3 Results from Experiment 3

The most demanding step of a single outer iteration of the proposed algorithm is the simultaneous adjustment of all Gaussians. If this step is done using L-BFGS-B, this amounts to a cost of $\mathcal{O}(N_i)$, both in terms of space and time [2]. Summing over the M iterations we hence find the space and time requirements of the proposed method to be $\mathcal{O}(M^2 n^2)$.

Algorithm 1

1: **procedure** GMM DECOMPOSITION
    Input: Signal d, hyper-parameters $\tau_1, \tau_2 \in \mathbb{Z}_+$
    Output: GMM parameters $\left(a_m^*, x_m^*, \Sigma_m^*\right)_{m=1}^M$
2:    $M \leftarrow 0, r \leftarrow d$
3:    **while** stopping criterion is not met **do**
4:        $M \leftarrow M + 1$
5:        $r'(y) \leftarrow \frac{1}{\tau_1} \sum_{z \in A(y)} r(z),$
          $A(y) := \{\tau_1 \text{ grid points closest to } y\}$
6:        $x_0 \leftarrow \arg\max_y r'(y)$
7:        $r'' \leftarrow$ restriction of r to the τ_2 grid points
          closest to x_0
8:        $\Sigma_0 \leftarrow$ square root of matrix of second order
          moments of $\frac{r''}{|r''|_1}$
9:        $a_0 \leftarrow \arg\min_{a \geq 0} |r - a g(x_0, \Sigma_0)|_2^2$
10:      $(a^*, x^*, \Sigma^*) \leftarrow \underbrace{\text{loc. min.} |r - a g(x, \Sigma)|_2^2, a \geq 0}_{\text{initialized at } a_0, x_0, \Sigma_0}$
11:      $\left(a_M^*, x_M^*, \Sigma_M^*\right) \leftarrow (a^*, x^*, \Sigma^*)$
12:      $\left(a_m^{**}, x_m^{**}, \Sigma_m^{**}\right)_{m=1}^M \leftarrow \left(a_m^*, x_m^*, \Sigma_m^*\right)_{m=1}^M$
13:      $\left(a_m^*, x_m^*, \Sigma_m^*\right)_{m=1}^M \leftarrow$
        $\underbrace{\text{loc. min.} \left|d - \sum_{m=1}^M a_m g(x_m, \Sigma_m)\right|_2^2, a_m \geq 0}_{\text{initialized at } (a_m^{**}, x_m^{**}, \Sigma_m^{**})_{m=1}^M}$
14:      $r \leftarrow d - \sum_{m=1}^M a_m^* g(x_m^*, \Sigma_m^*)$

3.1 Main numerical results

The theoretical results that our algorithm relies on are valid for any dimension. However, motivated by TEM data acquisition, we limit our examples up to 2D, which is sufficient for recovering the 3D structure [14]. As proof of concept for the proposed method, we ran our algorithm on small toy examples in 1D and 2D. We refer to the examples as Experiment 1–Experiment 3, see Figs. 1, 2, 3. A clean signal d_{clean} was generated by discretizing GMMs with parameters as in Table 1, 2, 3 to the grids $\left\{y_k = -10 + \frac{20k}{1000}\right\}_{k=0}^{1000}$ and $\left\{y_{k,\ell} = \left(-10 + \frac{20k}{65}, -10 + \frac{20\ell}{65}\right)\right\}_{k,\ell=0}^{64}$ in the 1D and 2D experiments, respectively. From this signal, noisy data were then generated as $d = d_{\text{clean}} + \epsilon$ where ϵ denotes white Gaussian noise with standard deviation σ_{noise}. The latter was chosen so that SNR = 20, where SNR := $10 \log_{10} \frac{\text{Var}(d_{\text{clean}})}{\sigma_{\text{noise}}^2}$. As stopping criterion we used $\text{SNR}_{\text{stop}} \geq 20$, where $\widehat{\text{SNR}}_{\text{stop}} := 10 \log_{10} \frac{\text{Var}(d_{\text{est}})}{\text{Var}(d - d_{\text{est}})}$, and where $d_{est} := \sum_m a_m^* g(x_m^*, \Sigma_m^*)$. In all minimization sub-procedures we used the L-BFGS-B

Table 1 Input and output GMM parameters for Exp. 1

Input			Output		
a_m	x_m	Σ_m^2	a_m	x_m	Σ_m^2
1	0	1	5.0272	1.3052	2.2632
8	0	4	1.0116	−7.9756	0.9512
1	−2	1	1.0351	8.0478	1.1041
1	2	1	5.8708	−1.1215	2.4119
1	−8	1			
1	8	1			

Table 2 Input and output GMM parameters for Exp. 2

Input		
a_m	x_m	Σ_m^2
2	$(-1.5, -2.5981)$	$\begin{pmatrix} 0.7969 & 1.272 \\ 1.272 & 2.2656 \end{pmatrix}$
2	$(-1.5, 2.5981)$	$\begin{pmatrix} 0.7969 & -1.272 \\ -1.272 & 2.2656 \end{pmatrix}$
2	$(3, 0)$	$\begin{pmatrix} 3 & 0 \\ 0 & 0.0625 \end{pmatrix}$
1	$(-1.75, -3.0311)$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Output		
a_m	x_m	Σ_m^2
1.9833	$(-1.4937, -2.5862)$	$\begin{pmatrix} 0.5758 & 1.1181 \\ 1.1181 & 2.4908 \end{pmatrix}$
1.9901	$(3.0075, -9.7258 \times 10^{-4})$	$\begin{pmatrix} 2.982 & 0.0082 \\ 0.0082 & 0.0623 \end{pmatrix}$
2.001	$(-1.4994, 2.5924)$	$\begin{pmatrix} 0.7858 & -1.2663 \\ -1.2663 & 2.2805 \end{pmatrix}$
1.0101	$(-1.7466, -3.038)$	$\begin{pmatrix} 0.9878 & 0.0126 \\ 0.0126 & 0.9731 \end{pmatrix}$

method with non-negativity constraints on the weights and ran it until convergence. Hyper-parameters τ_1 and τ_2 (introduced in line 5 and 7 in Algorithm 1) should ideally be chosen based on the amount of noise in the data; however, we found that the exact values were not so important in our toy examples, and we used the ad-hoc chosen values $\tau_1 = 10$ and $\tau_2 = 20$. We leave a more careful sensitivity analysis with respect to these parameters to future studies. The total runtime was a few minutes on a computer with an Intel Pentium CPU running at 2.90GHz, using ~ 8 GB of RAM. The final results of our experiments are tabulated in Table 1, 2, 3 and are plotted in Figure 1, 2, 3. Intermediate results for the 2D experiments are plotted in Figure 4, 5.

Remark 1 The number of modes of a GMM is not necessarily equal to the number of means. Indeed, there might be more modes than means [1]. Alternatively, there could be fewer modes than means, for example, in Experiment 2, a low-amplitude Gaussian and high-amplitude Gaussian are close to each other, and there is only one mode within the vicinity of both means. Our method successfully recovered the parameters of both Gaussians. There is also a mode at the center dominantly formed by the superposition of the tails of three anisotropic Gaussians, where our algorithm did not introduce a spurious Gaussian. Similar performance is observed in Experiment 3. We attribute this to the greedy nature of the algorithm where a predefined number of Gaussians is not enforced in the decomposition.

3.2 Clustering and comparison to expectation maximization

As one of the main techniques in unsupervised learning, clustering seeks to subdivide data into groups based on some similarity measure. A common approach in clustering is based on Gaussian mixtures. In this approach, one is given a point-cloud $P \subset \mathbb{R}^n$ and the goal is to find a GMM f such that the probability that P is a sample from f is maximal among all GMMs with a prescribed number of components.

There are many ways of transforming a point-cloud into a signal and vice-versa. Thus one may use the method proposed

in this paper for GMM-based clustering, and conversely, one may use methods from clustering in order to decompose a signal into a GMM. Below we illustrate both directions of this "problem transformation" with some numerical examples.

3.2.1 Clustering via signal decomposition

A sample point-cloud P of size 10^5 was drawn from a normalized version of the GMM in Experiment 3 above. Then a signal d was generated from P by computing a histogram based on the 2D grid used in the previous experiments. Results from the expectation maximization (EM) algorithm and the proposed method applied to P and d, respectively, are shown in Fig. 6. The two methods resulted in approximately the same likelihood of P given the computed GMMs.
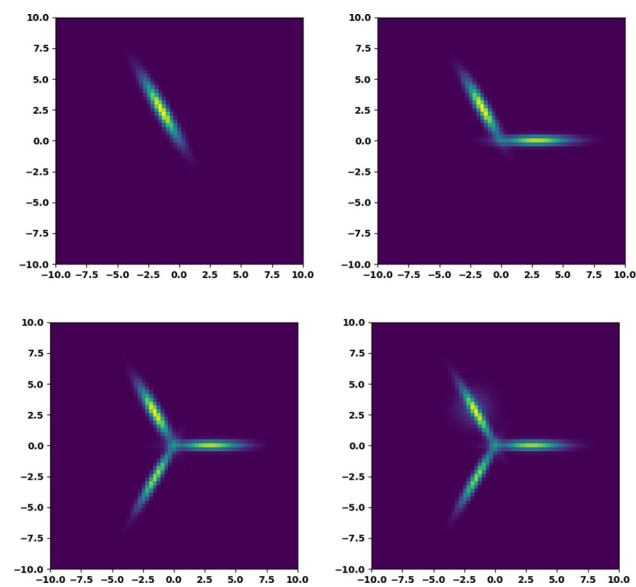
3.2.2 Signal decomposition via clustering

A noisy signal d was generated as in Experiment 3 described in Sect. 3.1, except that the GMM was normalized before discretization. We then associated a point-cloud P to the noisy signal, following the methodology in [7]. More precisely, P had $\lfloor C^{-1}d(y) \rfloor$ points at grid-point y, where the constant $C := 10^{-5} \sum_y d(y)$ was chosen so that P had roughly 10^5 points in total. Using P we estimated a GMM with the EM algorithm. The performance of EM depended on the random initialization. A typical result using EM applied to P, along with the result from the proposed method applied to d, is shown in Fig. 7. We remark that it is possible that a better
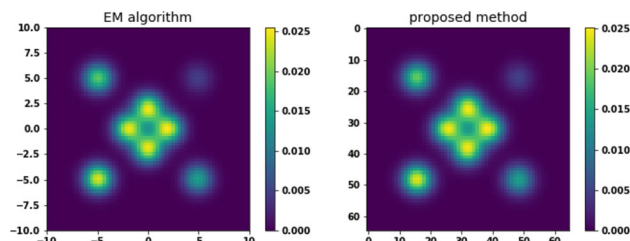
Table 3 Input and output GMM parameters for Exp. 3

Input		
a_m	x_m	Σ_m^2
5	$(-5, 5)$	Id_2
1	$(5, -5)$	Id_2
3	$(5, 5)$	Id_2
4	$(-5, -5)$	Id_2
5	$(-2, 0)$	Id_2
5	$(0, -2)$	Id_2
5	$(2, 0)$	Id_2
5	$(0, 2)$	Id_2
Output		
a_m	x_m	Σ_m^2
4.9685	$(9.7331 \times 10^{-4}, 2.0079)$	$\begin{pmatrix} 0.9781 & 0.0125 \\ 0.0125 & 1.012 \end{pmatrix}$
4.9517	$(-4.9988, 5.0024)$	$\begin{pmatrix} 1.0006 & 0.0127 \\ 0.0127 & 1.0103 \end{pmatrix}$
3.9821	$(-5.005, -4.9978)$	$\begin{pmatrix} 0.9919 & 0.0017 \\ 0.0017 & 0.9731 \end{pmatrix}$
2.9853	$(4.9935, 5.0071)$	$\begin{pmatrix} 0.9955 & -0.0308 \\ -0.0308 & 0.9976 \end{pmatrix}$
5.0608	$(-1.9916, 0.0076)$	$\begin{pmatrix} 1.0128 & 0.0154 \\ 0.0154 & 0.9977 \end{pmatrix}$
5.0322	$(0.0034, -1.9952)$	$\begin{pmatrix} 0.9968 & 0.0145 \\ 0.0145 & 1.0038 \end{pmatrix}$
1.0048	$(5.0158, -5.0174)$	$\begin{pmatrix} 1.007 & 0.0083 \\ 0.0083 & 1.0034 \end{pmatrix}$
4.9247	$(2.0213, 0.0077)$	$\begin{pmatrix} 0.9743 & -0.0097 \\ -0.0097 & 0.9997 \end{pmatrix}$
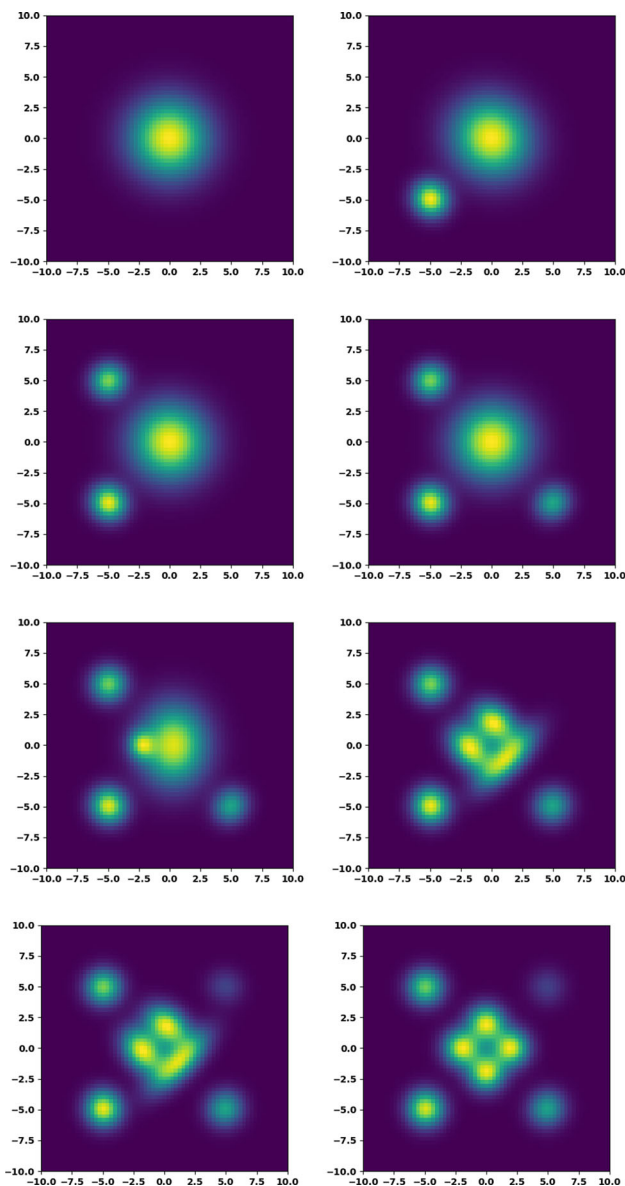
Fig. 5 Iterations from Experiment 3

Fig. 6 Results from clustering via signal decomposition

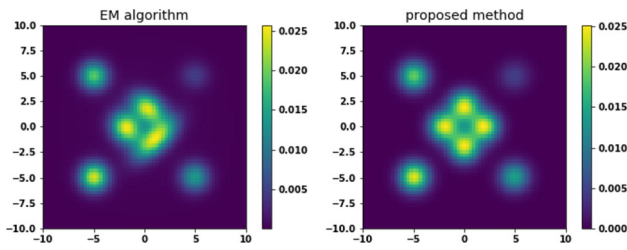**Fig. 4** Iterations from Experiment 2

Fig. 7 Results from signal decomposition via clustering

result with EM may be obtained using some other procedure for generating the point-cloud from the given signal.

4 Theoretical considerations

The main theoretical contribution of our paper is Theorem 3.4, which says that a mode of a GMM cannot lie too far away from the set of means. This theorem generalizes the 1D result that any mode will lie within one standard deviation away from some mean, and provides theoretical support for our way of initializing each iteration of our method. Before turning to this theorem, we need three lemmas. In the first two lemmas we compute some integrals over the n-sphere of certain polynomials, and in the third lemma we provide expressions for the gradient and Hessian of a GMM. We also prove, in Proposition 6, that any sufficiently nice non-negative function may be approximated in L^∞ to any desired accuracy by a GMM.

4.1 Location of modes of Gaussian mixtures

Lemma 1

$$\int_{S^{n-1}} (x, y)^k dy = |x|^k C_k, \quad x \in \mathbb{R}^n, k \in \mathbb{N},$$

where $C_k := \int_{S^{n-1}} (e_1, y)^k \, dy$.

Proof Omitted due to space constraints. □

Lemma 2 *Let A be a symmetric $n \times n$ matrix. Then*

$$\int_{S^{n-1}} (y, Ay)dy = Tr(A)C_2, \ \text{where } C_2 \text{ is as in Lemma 1.}$$

Proof Let $\lambda_i, i = 1, \ldots, n$ be the eigenvalues of A and take $U \in O(n)$ s.t. $A = U^T D U$, $D := \text{diag}(\lambda_1, \ldots, \lambda_n)$.

$$\int_{S^{n-1}} (y, Ay)dy = \int_{S^{n-1}} (y, U^T DU y)dy$$
$$= \int_{S^{n-1}} (Uy, DUy)dy = \int_{S^{n-1}} (y, Dy)dy$$

$$= \sum_{m=1}^{n} \lambda_m \int_{S^{n-1}} (y, e_m e_m^T y)dy$$
$$= \sum_{m=1}^{n} \lambda_m \int_{S^{n-1}} (e_m, y)^2 dy = C_2 \sum_{m=1}^{n} \lambda_m = C_2 \text{Tr}(A).$$

□

Lemma 3 *The gradient and Hessian of a Gaussian mixture model* [3]

$$f(x) = \sum_m a_m \exp\left\{-\frac{1}{2}(x - x_m)^T \Sigma_m^{-2} (x - x_m)\right\},$$
$$a_m > 0, \Sigma_m > 0, \text{ are given by}$$

$$\nabla f = \sum_m a_m \left(-\Sigma_m^{-1}\right) \left(\Sigma_m^{-1} x - \Sigma_m^{-1} x_m\right)$$
$$\times \exp\left\{-\frac{1}{2}(x - x_m)^T \Sigma_m^{-2} (x - x_m)\right\}$$
$$Hf = \sum_m a_m \left[\left(\Sigma_m^{-2}(x - x_m)\right)\left(\Sigma_m^{-2}(x - x_m)\right)^T - \Sigma_m^{-2}\right]$$
$$\times \exp\left\{-\frac{1}{2}(x - x_m)^T \Sigma_m^{-2} (x - x_m)\right\}$$

Proof Since this is a standard result, we omit the straightforward proof. □

Theorem 1 *Consider a Gaussian mixture model in n dimensions*

$$f(x) = \sum_m a_m \exp\left\{-\frac{1}{2}(x - x_m)^T \Sigma_m^{-2} (x - x_m)\right\},$$

$a_m > 0, \Sigma_m > 0$. *Let $\sigma_{m,\max}$ and $\sigma_{m,\min}$ be the maximal and minimal eigenvalues of Σ_m. Let x' be a local maximum of f. Then there exists an index m such that*

$$|x' - x_m| \leq \sqrt{n}\sigma_{m,\max}^2 \sigma_{m,\min}^{-1}.$$

Proof Since x' is a local maximum we have $Hf(x') \leq 0$, i.e., $(y, Hf(x')y) \leq 0, y \in \mathbb{R}^n$. We integrate the last inequality over the unit-sphere and make use of Lemma 3 to conclude:

$$\sum_m a_m \int_{S^{n-1}} y^T A_m(x', y)y dy \exp\{\cdots\} \leq 0, \ \text{where}$$

$$A_m(x', y) := \left[\left(\Sigma_m^{-2}(x' - x_m)\right)\left(\Sigma_m^{-2}(x' - x_m)\right)^T - \Sigma_m^{-2}\right].$$

Hence there exist some index m such that

$$\int_{S^{n-1}} y^T A_m(x', y)y dy \leq 0,$$

---

[3] In the interest of readability we have abused notation and absorbed the normalizing constants C_{Σ_m} into the weights a_m.

which leads to

$$\int_{S^{n-1}} \left(y, \Sigma_m^{-2} \left(x' - x_m \right) \right)^2 dy - \int_{S^{n-1}} \left(y, \Sigma_m^{-2} y \right) dy \le 0.$$

We apply Lemma 1 and Lemma 2 and obtain:

$$C_2 \left| \Sigma_m^{-2} \left(x' - x_m \right) \right|^2 - C_2 \mathrm{Tr} \left(\Sigma_m^{-2} \right) \le 0.$$

Now $C_2 > 0$ since C_2 is an integral of a continuous non-negative function that is not everywhere zero. Hence $\left| \Sigma_m^{-2} \left(x' - x_m \right) \right|^2 \le \mathrm{Tr} \left(\Sigma_m^{-2} \right)$. Note that $\sigma_{m,\max}^{-4} \left| x' - x_m \right|^2 \le \left| \Sigma_m^{-2} \left(x' - x_m \right) \right|^2$ and that $\mathrm{Tr} \left(\Sigma_m^{-2} \right) \le n \sigma_{m,\min}^{-2}$. So $\sigma_{m,\max}^{-4} \left| x' - x_m \right|^2 \le n \sigma_{m,\min}^{-2}$, and the claim of the theorem follows. $\square$

Corollary 4 *Let f be a mixture of spherical Gaussians with common variance σ^2, i.e.,*

$$f(x) = \sum_m a_m \exp \left\{ -\frac{1}{2\sigma^2} |x - x_m|^2 \right\}, \sigma > 0, a_m > 0.$$

If x' is local maximum of f, then there is an index m such that $|x' - x_m| \le \sqrt{n}\sigma$.

Proof This is an immediate consequence of Theorem 1. $\square$

Proposition 5 *The bound in Theorem 1 cannot be improved by a constant, i.e., for any $\delta > 0$ there exist a GMM f such that some mode x' of f satisfies*

$$|x' - x_m| > \sqrt{n}\sigma_{m,\max}^2 \sigma_{m,\min}^{-1} - \delta,$$

for all mean vectors x_m of f.

Proof We explicitly construct a family $\{f_\epsilon\}_\epsilon$ of functions that satisfy the statement of this proposition. For ϵ such that $\min \left(\delta, \sqrt{n}\sigma \right) > \epsilon > 0$ let f_ϵ be the $2n$ component n-dimensional spherical GMM with common variance σ^2, common amplitude a and with means at $\pm(\sqrt{n}\sigma - \epsilon)e_i$, for $i = 1, 2, \ldots, n$. We shall prove that f_ϵ has a mode in the origin, and thus it has a mode at distance $\sqrt{n}\sigma - \epsilon$ from the set of means of f_ϵ.[4] Lemma 3 implies

$$\nabla f_\epsilon(0) = a \sum_m \sigma^{-2} x_m \exp \left\{ -\frac{1}{2\sigma^2} |x_m|^2 \right\}$$

$$= a\sigma^{-2} \exp \left\{ -\frac{(\sqrt{n}\sigma - \epsilon)^2}{2\sigma^2} \right\} \sum_m x_m.$$

---

[4] Numerical experiments suggest that f_ϵ has a mode in the origin also for $\epsilon = 0$. A proof of this (if it is true) would however need an argument different from the one given here, since $Hf_0(0) = 0$.

By symmetry $\sum_m x_m = 0$, so f_ϵ has a critical point in the origin. Hence we are done if we prove that $Hf_\epsilon(0) < 0$. Again by Lemma 3:

$$Hf_\epsilon(0) = a \sum_m \left(\sigma^{-4} x_m x_m^T - \sigma^{-2} I_n \right) e^{-\frac{1}{2\sigma^2} |x_m|^2}$$

$$= 2a e^{-\frac{(\sqrt{n}\sigma - \epsilon)^2}{2\sigma^2}} \sum_{i=1}^n \left(\sigma^{-4} (\sqrt{n}\sigma - \epsilon)^2 e_i e_i^T - \sigma^{-2} I_n \right)$$

$$= 2a e^{-\frac{(\sqrt{n}\sigma - \epsilon)^2}{2\sigma^2}} \sigma^{-2} \underbrace{\left[\sigma^{-2} (\sqrt{n}\sigma - \epsilon)^2 - n \right]}_{<0} I_n.$$

Hence $Hf_\epsilon(0) < 0$. $\square$

4.2 Function approximation by Gaussian mixtures

The ability of GMMs to approximate functions in L^p spaces has been investigated previously, see, e.g., [12] where it is noted that any probability density function may be approximated in the sense of L^1 by GMMs. For completeness, we here give a proof of the density of GMMs in the L^∞-norm. The proof relies on the machinery of *quasi-interpolants* [11].

Proposition 6 *Let u be a twice-differentiable and compactly supported non-negative function on $\mathbb{R}^n$ such that u and all its partial derivatives up to order two are bounded. Then u may be approximated in L^∞ arbitrarily well by Gaussian mixtures.*

Proof Let $g = g(0, \Sigma)$ be an n-dimensional centered Gaussian density function with non-degenerate covariance matrix Σ^2. As is noted on page 50 in [11], g generates an approximate quasi-interpolant of order 2. This follows from the observation that f satisfies both Condition 2.15 in [11, p. 33], with moment order $N = 2$, and Condition 2.12 in [11, p. 32], with decay order K, for all $\mathbb{Z} \ni K > n$. Hence Theorem 2.17 in [11, p. 35] is applicable and implies that for any twice-differentiable real-valued function u on $\mathbb{R}^n$ such that u and all its partial derivatives up to order two are bounded, one has that for any $\epsilon > 0$ that there exists a $\mathcal{D} = \mathcal{D}(\epsilon) > 0$ such that for all h, $\left| u - \mathcal{M}_{h,\mathcal{D}} u \right|_\infty$ is bounded from the above by $c\mathcal{D}h^2 \max_{|\alpha|=2} |\partial^\alpha u|_\infty + \epsilon \left(|u|_\infty + \sqrt{\mathcal{D}}h|\nabla u|_\infty \right) \le A \left(c\mathcal{D}h^2 + \epsilon + \sqrt{\mathcal{D}}h \right),$

where $A := \sqrt{n} \max_{0 \le |\alpha| \le 2} |\partial^\alpha u|_\infty$, c is a constant independent of u, h and $\mathcal{D}$ and the quasi-interpolant $\mathcal{M}_{h,\mathcal{D}} u(x)$ is defined by

$$\mathcal{M}_{h,\mathcal{D}} u(x) := \mathcal{D}^{-n/2} \sum_{m \in \mathbb{Z}^n} u(hm) g \left(\frac{x - hm}{\sqrt{\mathcal{D}}h} \right). \tag{1}$$

Note that the weights $u(hm)$ are non-negative since u is non-negative, and that the compact support of u allows us

to restrict the domain of summation in (1) to a finite subset of $\mathbb{Z}^n$. Hence $\mathcal{M}_{h,\mathcal{D}}u(x)$ is a GMM. Now for a given $\epsilon > 0$ we let $\epsilon' := \frac{\epsilon}{2A}$ and pick $\mathcal{D} > 0$ such that:

$$\left| u - \mathcal{M}_{h,\mathcal{D}}u \right|_\infty \le A \left(c\mathcal{D}h^2 + \epsilon' + \sqrt{\mathcal{D}}h \right)$$
$$= \epsilon/2 + A \left(c\mathcal{D}h^2 + \sqrt{\mathcal{D}}h \right).$$

Next we take h small enough, so that $A \left(c\mathcal{D}h^2 + \sqrt{\mathcal{D}}h \right) \le \epsilon/2$, and conclude that $\left| u - \mathcal{M}_{h,\mathcal{D}}u \right|_\infty \le \epsilon$. $\qquad\square$

5 Conclusion

Motivated primarily by applications in TEM, we have developed a new algorithm for decomposing a non-negative multivariate signal as a sum of Gaussians with full covariances. We have tested it on 1D and 2D data. Moreover, we have also proved an upper bound for the distance from a local maximum of a GMM to the set of its mean vectors. This upper bound provides motivation for a key step in our method, namely the initialization of each new Gaussian at the maximum of the residual. Finally we remark that while we have only tested the proposed method on functions sampled on uniform grids, it is straightforward to extend the method to handle input data in the form of multivariate functions sampled on *non-uniform* grids.

Availability of data and material Upon request

Declaration

Conflict of interest The authors declare that they have no conflict of interest

Code availability Upon request.

References

1. Amendola, C., Engstrom, A. and Haase, C.: Maximum number of modes of gaussian mixtures. arXiv preprint arXiv:1702.05066, 2017
2. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
3. Carreira-Perpinan, M.A.: Mode-finding for mixtures of Gaussian distributions. Technical report, Dept. of Computer Science, University of Sheffield, UK, 2000
4. Fatma, K.B. and Cetin, A.E.: Design of gaussian mixture models using matching pursuit. *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 1999
5. Jonic, S., Sorzano, C.O.S.: Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: Algorithm to automatically control accuracy of approximation. IEEE J. Sel. Topics Signal Process. **10**(1), 161–173 (2016)
6. Jonic, S. and Sorzano, C.O.S.: Versatility of approximating single-particle electron density maps using pseudoatoms and approximation-accuracy control. BioMed Research International (2016)
7. Joubert, P.: *A Bayesian approach to initial model inference in cryo-electron microscopy*. PhD thesis, Georg-August University School of Science, 2016
8. Kawabata, T.: Gaussian-input Gaussian mixture model for representing density maps and atomic models. J. Struct. Biol. **203**, 1–16 (2018)
9. Keriven, N., Bourrier, A., Gribonval, R., Perez, P.: Sketching for large-scale learning of mixture models. Inf. Inference **7**(3), 447–508 (2018)
10. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process **41**(12), 3397–3415 (1993)
11. Mazya V. G., and Schmidt G.: *Approximate approximations*. Number 141. American Mathematical Soc., 2007
12. Nestoridis, V., Schmutzhard, S., Stefanopoulos, V.: Universal series induced by approximate identities and some relevant applications. J. Approx. Theory **163**(12), 1783–1797 (2011)
13. Pati, Y.C., Rezaiifar, R. and Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993
14. Zickert, G.: *Analytic and data-driven methods for 3D electron microscopy*. PhD thesis, KTH Royal Institute of Technology, 2020