On the Number of Modes of a Gaussian Mixture

Miguel Á. Carreira-Perpiñán[1] and Christopher K. I. Williams[2]

[1] Dept. of Computer Science, University of Toronto
miguel@cs.toronto.edu
[2] School of Informatics, University of Edinburgh
c.k.i.williams@ed.ac.uk

Abstract. We consider a problem intimately related to the creation of maxima under Gaussian blurring: the number of modes of a Gaussian mixture in D dimensions. To our knowledge, a general answer to this question is not known. We conjecture that if the components of the mixture have the same covariance matrix (or the same covariance matrix up to a scaling factor), then the number of modes cannot exceed the number of components. We demonstrate that the number of modes can exceed the number of components when the components are allowed to have arbitrary and different covariance matrices.

We will review related results from scale-space theory, statistics and machine learning, including a proof of the conjecture in 1D. We present a convergent, EM-like algorithm for mode finding and compare results of searching for all modes starting from the centers of the mixture components with a brute-force search. We also discuss applications to data reconstruction and clustering.

1 Introduction

We propose a mathematical conjecture about Gaussian mixtures (GMs): that, under certain conditions, the number of modes cannot exceed the number of components. Although we originally came across this conjecture in a pattern recognition problem (sequential data reconstruction), it is intimately related to scale-space theory (since some GMs are the convolution of a delta mixture with a Gaussian kernel) and statistical smoothing (since Gaussian kernel density estimates are GMs). Bounding the number of modes and the region where they lie, and finding all these modes, is of interest in these areas. The widespread use of GMs makes the conjecture relevant not only theoretically but also in applications of these areas, such as data reconstruction, image segmentation or clustering.

We state formally the conjecture and prove part of it in Sect. 2, and review related proof approaches in Sect. 3. We show the convergence of an algorithm that tries to find all modes in Sect. 4 and discuss applications in Sect. 5. An extended version of this paper appears as [1].

2 The Conjecture

Consider a GM density of $M > 1$ components in $\mathbb{R}^D$ for $D \geq 1$, with mixture proportions $\{\pi_m\}_{m=1}^M \subset (0,1)$ satisfying $\sum_{m=1}^M \pi_m = 1$, component means $\{\boldsymbol{\mu}_m\}_{m=1}^M \subset \mathbb{R}^D$ and positive definite covariance matrices $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$:

$$p(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{m=1}^M p(m)p(\mathbf{x}|m) \stackrel{\text{def}}{=} \sum_{m=1}^M \pi_m p(\mathbf{x}|m) \qquad \forall \mathbf{x} \in \mathbb{R}^D \qquad \mathbf{x}|m \sim \mathcal{N}_D(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m).$$

In general, there is no explicit expression for the modes of p, i.e., no analytic solution for the stationary points in eq. (1); we do not even know how many modes p has. Intuitively, it seems reasonable that the number of modes of p will not exceed the number M of components in the GM: the more the different components interact (depending on their mutual separation and on their covariance matrices), the more they will coalesce and the fewer modes will exist. Besides, modes should always appear inside the region enclosed by the component centroids $\{\boldsymbol{\mu}_m\}_{m=1}^M$—more precisely, in their convex hull[3]. Based on this reasoning, Carreira-Perpiñán [2] (see also [3]) proposed the following conjecture.

Conjecture. Let $p(\mathbf{x}) = \sum_{m=1}^M p(m)p(\mathbf{x}|m)$, where $\mathbf{x}|m \sim \mathcal{N}_D(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, be a mixture of M D-variate normal distributions. Then $p(\mathbf{x})$ has M modes at most, all of which are in the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^M$, if one of the following conditions holds:

1. $D = 1$ *(one-dimensional mixture).*
2. $D \geq 1$ and the covariance matrices are arbitrary but equal: $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ $\forall m = 1, \ldots, M$ *(homoscedastic mixture).*
3. $D \geq 1$ and the covariance matrices are isotropic: $\boldsymbol{\Sigma}_m = \sigma_m^2 \mathbf{I}_D$ *(isotropic mixture).*

Several parts of this conjecture hold, namely the modes (and all other stationary points) lie in the convex hull, and for $D = 1$ the number of modes does not exceed M. We will prove this below. Besides, the conditions of the conjecture are necessary. Figure 1 gives examples of a GM with nonisotropic, different component covariance matrices that has more modes than components and the modes lie outside the convex hull of the centroids. Also, if the kernel $p(\mathbf{x}|m)$ is not Gaussian one can construct examples where the conjecture does not hold. This may seem counterintuitive, since one may expect that localised, tapering kernels would behave like the Gaussian. However, small modes can typically arise where kernels interact—although it may occur only rarely. The necessity that the kernel be Gaussian has been established in scale-space theory (see Sect. 3.2).

The Modes Lie in the Convex Hull for Any Dimension D. All modes lie in the convex hull of the centroids for the case of isotropic GMs. One proof is given by the stationary-point eq. (3), which also shows that in generic cases the modes must lie strictly in the interior of the convex hull and not on its boundary. An alternative proof is given in [2, p. 218].

[3] Defined as the set $\left\{\mathbf{x} : \mathbf{x} = \sum_{m=1}^M \lambda_m \boldsymbol{\mu}_m \text{ with } \{\lambda_m\}_{m=1}^M \subset [0,1] \text{ and } \sum_{m=1}^M \lambda_m = 1\right\}$.

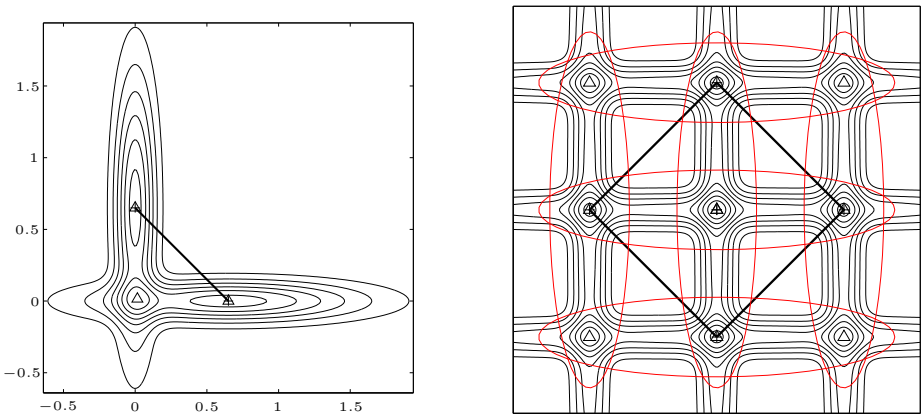

Fig. 1. GMs in dimension $D \geq 2$ that have different, non-isotropic covariances do not generally verify conjecture 2. The left graph shows a contour plot for a bicomponent GM with $\pi_1 = \pi_2 = \frac{1}{2}$, $\boldsymbol{\mu}_1 = \left(\begin{smallmatrix} 0.6 \\ 0 \end{smallmatrix}\right)$, $\boldsymbol{\mu}_2 = \left(\begin{smallmatrix} 0 \\ 0.6 \end{smallmatrix}\right)$, $\boldsymbol{\Sigma}_1 = \left(\begin{smallmatrix} 0.65 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$ and $\boldsymbol{\Sigma}_2 = \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.65 \end{smallmatrix}\right)$. This GM has three modes (marked "△"): two nearly coincident with the centroids $\boldsymbol{\mu}_m$ (marked "+") and a third one near the meeting point of the components' principal axes. All the modes are outside the convex hull of the centroids (marked by the thick line). More complicated arrangements can result in a multiplicity of modes, as shown in the right graph (inspired by Fig. 2 of [4]).

The Homoscedastic Case is Equivalent to the Homoscedastic Isotropic One. The following theorem shows that the modes problem for a homoscedastic GM with a given arbitrary covariance $\boldsymbol{\Sigma}$ is equivalent to that of another homoscedastic GM with isotropic covariance $\sigma^2\mathbf{I}$ (for a certain σ). Thus, one can try to prove a result for the simple case of isotropic covariances and then the result will also hold for $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ arbitrary. The reason is that, by rotating and rescaling the coordinate axes, we can spherise each component.

Theorem 1. *The mixtures* $p(\mathbf{x}) = \sum_{m=1}^{M} \pi_m \left|2\pi\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)}$ *(arbitrary but equal covariances) and* $p(\mathbf{u}) = \sum_{m=1}^{M} \pi_m (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{u}-\boldsymbol{\nu}_m\|^2}$ *(unit covariances), related by a rotation and scaling, have the same number of modes, which lie in the respective centroid convex hulls.*

Proof. Let $\boldsymbol{\Sigma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ be the spectral decomposition of $\boldsymbol{\Sigma}^{-1}$, with $\mathbf{U}$ orthogonal and $\boldsymbol{\Lambda}$ diagonal and positive definite. Consider the coordinate transformation $\mathbf{u} \stackrel{\text{def}}{=} \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\mathbf{x}$ (orthogonal rotation followed by scaling), so that $p(\mathbf{u}) = \sum_{m=1}^{M} \pi_m (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{u}-\boldsymbol{\nu}_m\|^2}$ and $\nabla_{\mathbf{x}} p = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}} \nabla_{\mathbf{u}} p$, and define $\boldsymbol{\nu}_m \stackrel{\text{def}}{=} \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\boldsymbol{\mu}_m$. Since $\mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$ is nonsingular, $\nabla_{\mathbf{x}} p = \mathbf{0} \Leftrightarrow \nabla_{\mathbf{u}} p = \mathbf{0}$ and so the stationary points are preserved by the transformation.

Now, if $\mathbf{x}$ is a point in the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^{M}$ then $\mathbf{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{\mu}_m$ where $\{\lambda_m\}_{m=1}^{M} \subset [0,1]$ and $\sum_{m=1}^{M} \lambda_m = 1$. So $\mathbf{u} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\mathbf{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\boldsymbol{\mu}_m = \sum_{m=1}^{M} \lambda_m \boldsymbol{\nu}_m$ which is in the convex hull of $\{\boldsymbol{\nu}_m\}_{m=1}^{M}$. □

Theorem 1 shows that case 2 of conjecture 2 is a particular case of case 3 (case 1 is also a particular case of case 3, obviously).

The Conjecture Holds for $D = 1$. We can prove this using the scale-space theory proofs of non-creation of maxima with Gaussian blurring (Sect. 3.2). The intuitive idea is that, by alternating the operations of "planting" a delta function (of value π_m) at a centroid location $\boldsymbol{\mu}_m$ and applying Gaussian blurring (to fatten the delta) we can create any isotropic GM. If planting a delta adds a single mode and Gaussian blurring never creates modes (this latter result given by the mentioned proofs), then the number of modes will never exceed the number of components M. Our proof is by induction. Note that the only step that requires $D = 1$ is the application of the scale-space theorem.

Theorem 2. *In 1D, any Gaussian mixture with M components has at most M modes.*

Proof. By induction on M. The statement holds trivially for $M = 1$. Assume it holds for $M - 1$ components and consider an arbitrary GM p with $M > 1$ components. Consider the component with narrowest variance and call this σ_M^2, perhaps by reordering the components, so that $\sigma_M < \sigma_m \ \forall m < M$ (in the nongeneric case of ties, simply choose any of the narrowest ones and the argument holds likewise). Now apply Gaussian deblurring of variance σ_M^2, recalling that the convolution of two isotropic Gaussians of variances σ_a^2 and σ_b^2 is a Gaussian of variance $\sigma_a^2+\sigma_b^2$ (the semigroup structure). We obtain a mixture density p' where each component for $m = 1, \ldots, M-1$ is a Gaussian of mixing proportion π_m and variance $\sigma_m^2 - \sigma_M^2$, and component M is a delta function of mixing proportion π_M. Thus, p' is a mixture of a delta and a GM with $M - 1$ components. By the induction hypothesis the latter has $M - 1$ modes at most, so p' has M modes at most. Now apply Gaussian blurring to p'. By the scale-space theorems of Sect. 3.2, no new maxima can appear, and so the original GM p has M modes at most. □

The following corollary results from the fact that all marginal and conditional distributions of a GM (of arbitrary covariances) are also GMs.

Corollary 1. *Any 1D projection (marginal or conditional distribution) of any Gaussian mixture in D dimensions with M components has at most M modes.*

3 Approaches to Proving the Conjecture

We review results from different fields that concern the conjecture. Additional results applicable only in particular cases are given in [2].

3.1 System of Equations for the Stationary Points of the Density

In [2] the problem was approached by trying to determine the stationary (or critical) points of the GM density p as follows. Consider the case with $\boldsymbol{\Sigma}_m =$

$\boldsymbol{\Sigma},\ m = 1, \ldots, M$ (homoscedastic GM) and assume $\mathbf{x}$ is a stationary point of p. Then

$$\nabla p(\mathbf{x}) = \sum_{m=1}^{M} p(\mathbf{x}, m)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_m - \mathbf{x}) = \mathbf{0} \Longrightarrow \mathbf{x} = \sum_{m=1}^{M} p(m|\mathbf{x})\boldsymbol{\mu}_m. \tag{1}$$

This is a nonlinear system of D equations and D unknowns $x_1, \ldots, x_D \subset \mathbb{R}$. Since $p(m|\mathbf{x}) \in (0,1)$ for all m and $\sum_{m=1}^{M} p(m|\mathbf{x}) = 1$, $\mathbf{x}$ is a convex linear combination of the centroids and so all stationary points lie in the convex hull of the centroids. Instead, write $\mathbf{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{\mu}_m$ with $\lambda_m \in (0,1)$ and $\sum_{m=1}^{M} \lambda_m = 1$. Then we can consider $(m = 1, \ldots, M)$:

$$\lambda_m = p(m|\mathbf{x}) = \frac{\pi_m e^{-\frac{1}{2}\mathbf{u}_m^T \boldsymbol{\Sigma}^{-1}\mathbf{u}_m}}{\sum_{m'=1}^{M} \pi_{m'} e^{-\frac{1}{2}\mathbf{u}_{m'}^T \boldsymbol{\Sigma}^{-1}\mathbf{u}_{m'}}} \qquad \mathbf{u}_m \stackrel{\text{def}}{=} \mathbf{x} - \boldsymbol{\mu}_m = \sum_{m'=1}^{M} \lambda_{m'}\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m \tag{2}$$

as a nonlinear system of M equations and M unknowns $\lambda_1, \ldots, \lambda_M \in (0,1)$ subject to $\sum_{m=1}^{M} \lambda_m = 1$.

For $M = 2$ with $\lambda \stackrel{\text{def}}{=} \lambda_1$, $\lambda_2 = 1 - \lambda$, and $\pi \stackrel{\text{def}}{=} p(1)$, $p(2) = 1 - \pi$, eq. (2) reduces to the transcendental equation $\lambda = \frac{1}{1+e^{-\alpha(\lambda - \lambda_0)}}$ with $\alpha = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in (0, \infty)$ and $\lambda_0 = \frac{1}{2} + \frac{1}{\alpha}\log\frac{1-\pi}{\pi} \in (-\infty, \infty)$. This can have at most 3 roots in $(0,1)$, as can be easily seen geometrically in Fig. 2, and so at most 2 can be maxima.

Unfortunately, for higher M the system becomes very difficult to study. Besides, if a counterexample to the conjecture does exist, it is likely to require a nontrivial number of components M in $D \geq 2$, which makes very difficult to look for such a counterexample in terms of the λ_m's.

In case $\boldsymbol{\Sigma}_m = \sigma_m^2 \mathbf{I},\ m = 1, \ldots, M$ (isotropic components), we get the system

$$\lambda_m = q(m|\mathbf{x}) \stackrel{\text{def}}{=} \frac{p(m|\mathbf{x})\sigma_m^{-2}}{\sum_{m'=1}^{M} p(m'|\mathbf{x})\sigma_{m'}^{-2}} = \frac{\pi_m \sigma_m^{-(D+2)} e^{-\frac{1}{2}\left\|\frac{\mathbf{u}_m}{\sigma_m}\right\|^2}}{\sum_{m'=1}^{M} \pi_{m'} \sigma_{m'}^{-(D+2)} e^{-\frac{1}{2}\left\|\frac{\mathbf{u}_{m'}}{\sigma_{m'}}\right\|^2}} \tag{3}$$

again with $\mathbf{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{\mu}_m$, where $\lambda_m \in (0,1)$ and $\sum_{m=1}^{M} \lambda_m = 1$, and $\mathbf{u}_m$ as in eq. (2). In effect, λ_m equals the responsibility $p(m|\mathbf{x})$ but reweighted by the inverse variance and renormalised. An analogous analysis shows that there are 3 stationary points at most for the case $M = 2$ but becomes difficult in general.

A further problem with this approach is that the equations apply to all stationary points (maxima, minima and saddles) rather than to maxima only.

Note that the modes must lie strictly in the interior of the convex hull and not on its boundary, since $p(m|\mathbf{x})$ and $q(m|\mathbf{x}) < 1$ (except in non-generic cases such as when all the centroids are equal or some σ_m is zero).

3.2 Scale-Space Theory

We give here a short summary of the creation of maxima with Gaussian blurring in the scale space framework. The central issue of linear Gaussian scale space [5]

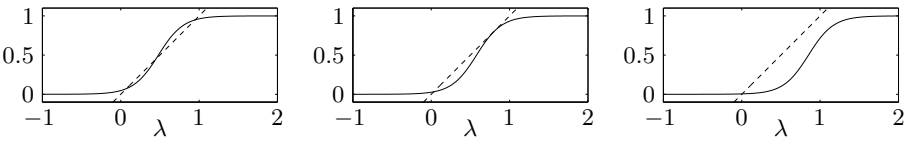

Fig. 2. Three possible cases for the solutions of the equation $\lambda = \frac{1}{1+e^{-\alpha(\lambda-\lambda_0)}}$.

is the generation of a family of functions $I(\mathbf{x}; s)$ by convolution or blurring of the original D-dimensional function $I(\mathbf{x})$ (the "greyscale image") with a Gaussian kernel of scale $s = \sigma^2$:

$$I(\mathbf{x}; s) \stackrel{\text{def}}{=} (G_s * I)(\mathbf{x}) = \int (2\pi s)^{-\frac{D}{2}} e^{-\frac{1}{2s}\|\mathbf{y}\|^2} I(\mathbf{x} - \mathbf{y})\, d\mathbf{y} \qquad \mathbf{x} \in \mathbb{R}^D$$

with $I(\mathbf{x}) \equiv I(\mathbf{x}; 0)$. As the scale increases, $I(\mathbf{x}; s)$ represents coarser structure. Several researchers (among others [6, 7, 8, 9]) proved that the Gaussian kernel never creates new maxima in 1D and, further, is the only kernel to do so. Their proofs are typically based on the following points. (1) Causality principle: since the Gaussian kernel is the Green's function of the diffusion equation, the family $I(\mathbf{x}; s)$ is the solution of the diffusion equation (where the time is given by the scale s) with initial condition $I(\mathbf{x}; 0) = I(\mathbf{x})$: $\frac{\partial I}{\partial s} = \frac{1}{2}\nabla^2_{\mathbf{x}} I$. (2) Particular properties of the blurring process and the Gaussian kernel, such as semigroup structure, homogeneity or isotropy. (3) The implicit function theorem applied to the variables $\mathbf{x}$ and s guarantees that the maxima trajectories $\mathbf{x} = \mathbf{x}(s)$ (along which the gradient $\nabla_{\mathbf{x}} I$ is zero) are continuously differentiable except at bifurcation points[4] where the Hessian of I with respect to $\mathbf{x}$ becomes singular and the topology changes. A concave level surface corresponds to the annihilation of a pair (a maximum with a minimum or saddle-point), while a convex one corresponds to the creation of a pair. The fact that the family satisfies the diffusion equation forbids the latter.

However, this does not hold in 2D (counterintuitive as it may seem, and though some of the mentioned proofs claimed it did) as originally evidenced by a counterexample proposed by Lifshitz and Pizer [10]: the original image is unimodal, made up by a low hill from whose summit a narrow ramp ascends over a deep valley towards a high hill (which contains the maximum). Gaussian blurring produces a dip in the ramp, creating a new maximum on the low hill, that later annihilates with the dip. This and further examples are analysed by Kuijper and Florack [11, 12], who also suggest that such created maxima are rare (being associated with elongated structures) and short-lived in scale space.

The definitive explanation of the creation of maxima was given by Damon [13] using Morse theory and catastrophe theory. Thom's theorem classifies the behaviour at bifurcation points of a family of functions dependent on parameters (such as the scale). This cannot be applied directly because the family is not unconstrained, but must obey the diffusion equation. Damon showed that maxima

[4] Also called degenerate critical points, top-points or catastrophes.

creations are associated with an umbilic catastrophe that occurs generically, i.e., does not disappear by perturbing the function.

In summary, in 2D or higher, there exist functions upon which Gaussian blurring results in occasional, but generic, creations of maxima as the scale increases. In 1D no such functions exist: Gaussian blurring never creates maxima, and is the only kernel to do so—for any other kernel, there exist functions on which it creates maxima.

How does this apply to the GM case? Our original "image" is a delta mixture $I(\mathbf{x}) = \sum_{m=1}^{M} \pi_m \delta(\mathbf{x} - \boldsymbol{\mu}_m)$, which by convolution with a Gaussian of variance $s = \sigma^2$ results in a homoscedastic isotropic GM with component covariances $\boldsymbol{\Sigma}_m = \sigma^2 \mathbf{I}_D$. At zero scale the mixture has M modes, one on each centroid $\boldsymbol{\mu}_m$. Therefore, in 1D the scale-space theorems state that no new modes appear as σ increases, which proves the conjecture for the homoscedastic case; and our Theorem 2 extends the proof to the *isotropic* case. In 2D or higher, the possibility that the Gaussian blurring may create modes does not necessarily disprove the conjecture. Firstly, it could be that for mixtures of Gaussians or deltas new modes can never appear; we have never succeeded to replicate a sequence of events such as that of [10]. Perhaps an approach based on catastrophe theory but restricted to initial images which are delta mixtures would resolve this question. Secondly, even if new modes can appear when blurring a GM, the total number of modes may still never exceed M. In other words, a situation of mode creation may require a large number of Gaussian components that interact to result in a GM with only a few modes before and after the creation. A brute-force search has failed to find counterexamples of the conjecture (see Sect. 4.3).

Note also that all catastrophes, being stationary points, must lie in the interior of the convex hull of the centroids (or, nongenerically, on its boundary), as mentioned in Sect. 2.

3.3 Kernel Density Estimation in 1D

Given a data sample $\{x_n\}_{n=1}^{N} \subset \mathbb{R}$, Silverman [14] considers the 1D Gaussian kernel density estimate $p(x; h)$ (Sect. 5.2). This is, of course, a homoscedastic isotropic GM of centroids $\{x_n\}_{n=1}^{N}$, variance h^2 and and equal mixing proportions $\pi_n = \frac{1}{N}$. In his proof, which we believe is not known to the scale-space community, Silverman shows that the number of maxima of $p(x; h)$ (or generally of $\partial^m p / \partial x^m$ for integer $m \geq 0$) is a right continuous decreasing function of h. His proof is based on the total positivity and the semigroup structure of the Gaussian kernel and the variation diminishing property of functions generated by convolutions with totally positive kernels. However, the proof uses the counts of sign changes of the mixture derivative and so it seems difficult to extend it to dimensions higher than 1.

4 Algorithms for Finding All the Modes

We now turn to the practically important question of finding all the modes of a GM. No direct solution exists, so we need to use numerical iterative methods.

Carreira-Perpiñán [15] suggested starting a mode-seeking algorithm from every centroid to locate all the modes. He gave two hill-climbing algorithms applicable to GMs with components of arbitrary covariance: a gradient-quadratic one and a fixed-point iteration one. Here we deal only with the latter because it allows to define in a unique way a basin of attraction for each mode, which is relevant both for the conjecture and for mean-shift algorithms. We also prove its convergence by deriving it as an EM algorithm.

4.1 The Fixed-Point Iteration Algorithm as an EM Algorithm

By equating the gradient of the GM density to zero, using Bayes' theorem and rearranging we obtain a fixed-point iterative scheme [15]:

$$\mathbf{x}^{(\tau+1)} = \mathbf{f}(\mathbf{x}^{(\tau)}) \text{ with } \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} \left(\sum_{m=1}^{M} p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} \right)^{-1} \sum_{m=1}^{M} p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m. \quad (4)$$

Following a suggestion from the second author, Carreira-Perpiñán [2] showed that this algorithm can also be derived as an *expectation-maximisation (EM) algorithm* [16, 17] as follows[5]. Consider the following density model with parameters $\mathbf{v} = (v_1, \dots, v_D)^T$ and fixed $\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$:

$$p(\mathbf{x}|\mathbf{v}) = \sum_{m=1}^{M} \pi_m \, |2\pi \boldsymbol{\Sigma}_m|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-(\boldsymbol{\mu}_m-\mathbf{v}))^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{x}-(\boldsymbol{\mu}_m-\mathbf{v}))}.$$

That is, $\mathbf{x}|\mathbf{v}$ is a D-dimensional GM where component m has mixing proportion π_m (fixed), mean vector $\boldsymbol{\mu}_m - \mathbf{v}$ ($\boldsymbol{\mu}_m$ fixed) and covariance matrix $\boldsymbol{\Sigma}_m$ (fixed). Varying $\mathbf{v}$ results in a rigid translation of the whole GM as a block rather than the individual components varying separately. Now consider fitting this model by maximum likelihood to a data set $\{\mathbf{x}_n\}_{n=1}^N$ and let us derive an EM algorithm to estimate the parameters $\mathbf{v}$. Call $z_n \in \{1, \dots, M\}$ the (unknown) index of the mixture component that generated data point $\mathbf{x}_n$. Then:

E step The complete-data log-likelihood, as if all $\{z_n\}_{n=1}^N$ were known, and assuming iid data, is $\sum_{n=1}^N \mathcal{L}_{n,\text{complete}}(\mathbf{v}) = \sum_{n=1}^N \log p(\mathbf{x}_n, z_n|\mathbf{v})$ and so its expectation with respect to the current posterior distribution is

$$\begin{aligned} Q(\mathbf{v}|\mathbf{v}^{(\tau)}) &\stackrel{\text{def}}{=} \sum_{n=1}^{N} \mathrm{E}_{p(z_n|\mathbf{x}_n,\mathbf{v}^{(\tau)})} \{\mathcal{L}_{n,\text{complete}}(\mathbf{v})\} \\ &= \sum_{n=1}^{N} \sum_{z_n=1}^{M} p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \log\{p(z_n|\mathbf{v}) p(\mathbf{x}_n|z_n, \mathbf{v})\} \\ &= \sum_{n=1}^{N} \sum_{z_n=1}^{M} p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \log p(\mathbf{x}_n|z_n, \mathbf{v}) + K \end{aligned}$$

where $K \stackrel{\text{def}}{=} \sum_{n=1}^N \sum_{z_n=1}^M p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \log \pi_{z_n}$ is independent of $\mathbf{v}$.

[5] We recently learned of an independent derivation by Y. Weiss (unpubl. manuscript).

M step The new parameter estimates $\mathbf{v}^{(\tau+1)}$ are obtained from the old ones $\mathbf{v}^{(\tau)}$ as $\mathbf{v}^{(\tau+1)} = \arg\max_{\mathbf{v}} Q(\mathbf{v}|\mathbf{v}^{(\tau)})$. To perform this maximisation, we equate the gradient of Q with respect to $\mathbf{v}$ to zero:

$$\frac{\partial Q}{\partial \mathbf{v}} = \sum_{n=1}^{N} \sum_{z_n=1}^{M} p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \frac{1}{p(\mathbf{x}_n|z_n, \mathbf{v})} \frac{\partial p(\mathbf{x}_n|z_n, \mathbf{v})}{\partial \mathbf{v}} = \mathbf{0}. \tag{5}$$

Solving for $\mathbf{v}$ in eq. (5) results in

$$\mathbf{v}^{(\tau+1)} = \left(\sum_{n=1}^{N} \sum_{z_n=1}^{M} p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \boldsymbol{\Sigma}_{z_n}^{-1} \right)^{-1} \sum_{n=1}^{N} \sum_{z_n=1}^{M} p(z_n|\mathbf{x}_n, \mathbf{v}^{(\tau)}) \boldsymbol{\Sigma}_{z_n}^{-1} (\boldsymbol{\mu}_{z_n} - \mathbf{x}_n).$$

If now we choose the data set as simply containing the origin, $\{\mathbf{x}_n\}_{n=1}^{N} = \{\mathbf{0}\}$, rename $z_1 = m$ and omit $\mathbf{x}_1 = \mathbf{0}$ for clarity, we obtain the M step as:

$$\mathbf{v}^{(\tau+1)} = \left(\sum_{m=1}^{M} p(m|\mathbf{v}^{(\tau)}) \boldsymbol{\Sigma}_{m}^{-1} \right)^{-1} \sum_{m=1}^{M} p(m|\mathbf{v}^{(\tau)}) \boldsymbol{\Sigma}_{m}^{-1} \boldsymbol{\mu}_{m} \tag{6}$$

which is formally identical to the iterative scheme of eq. (4).

General properties of the EM algorithm for GMs [16, 18, 17] show that the convergence of (6) is global and linear. Firstly, at every iteration τ, the iterative scheme (6) will either increase or leave unchanged the log-likelihood $\sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{v}) = \log p(\mathbf{0}|\mathbf{v})$ so, correspondingly, the iterative scheme (4) will monotonically increase the density value $p(\mathbf{x})$ or leave it unchanged. Thus, (4) converges from any initial value of $\mathbf{x}$ to a local stationary point of $p(\mathbf{x})$ [17, Th. 3.2]. Although convergence can occur to a saddle point or to a minimum as well as to a mode. Since both saddle points and minima are unstable for maximisation, a small random perturbation will cause the EM algorithm to diverge from them. Thus, practical convergence will almost always be to a mode. Secondly, its convergence rate is linear (first-order) and so is very slow except when the mixture components are very separated, in which case the convergence becomes superlinear. Note in Fig. 3 the slow crawl along ridges of the density and how the iterates may be attracted to saddle points, to then deviate towards a mode.

The EM view of the fixed-point algorithm should also be applicable to mixtures of other kernels.

4.2 Particular Cases

In the case of isotropic GMs the fixed-point scheme reduces to:

$$\mathbf{x}^{(\tau+1)} = \sum_{m=1}^{M} q(m|\mathbf{x}^{(\tau)}) \boldsymbol{\mu}_m \qquad q(m|\mathbf{x}) = \frac{p(m|\mathbf{x})\sigma_m^{-2}}{\sum_{m'=1}^{M} p(m'|\mathbf{x})\sigma_{m'}^{-2}} \tag{7}$$

where $p(m|\mathbf{x})$ is the posterior probability or responsibility of component m given point $\mathbf{x}$ and the $q(m|\mathbf{x})$ values are the responsibilities $p(m|\mathbf{x})$ reweighted by the

inverse variance and renormalised. For homoscedastic GMs, this simplifies even more with $q(m|\mathbf{x}) = p(m|\mathbf{x})$ so that the new point $\mathbf{x}^{(\tau+1)}$ is the conditional mean of the GM under the current point $\mathbf{x}^{(\tau)}$. This is formally akin to clustering by deterministic annealing [19], to algorithms for finding pre-images in kernel-based methods [20] and to mean-shift algorithms (Sect. 5.2).

In both cases, each iterate is a convex linear combination of the centroids, as are the stationary points, and so the sequence lies in the interior of the convex hull of the centroids. In general for finite mixtures of densities from the exponential family, the EM algorithm always stays in the convex hull of a certain set of parameters [18, eq. (5.3)].

4.3 Brute-Force Search for Counterexamples

Whether starting the algorithm from each centroid can indeed find all modes depends on the conjecture. It certainly does not hold in the general case where the covariance matrices are not isotropic and different, since then we can have more modes than centroids (although we may expect the algorithm to find many of the modes). Deriving an efficient algorithm to find all modes for this case is difficult, because we do not even know where to look for the modes: they need not lie inside the convex hull of the centroids, and may lie far away from them.

What happens in the cases where the conjecture may hold? Even if the number of modes is fewer than or equal to the number of components, some modes might conceivably not be reachable from any centroid. Since we can associate almost every point $\mathbf{x} \in \mathbb{R}^D$ with a unique mode (except for saddles, minima and points converging to them), we can define the *basin of attraction* of each mode as the region of $\mathbb{R}^D$ of all points that converge to that mode. The claim that the algorithm finds all modes if started from every centroid is equivalent to the claim that the basin of attraction of every mode contains at least one centroid.

Theoretically, this question seems as difficult as the modes conjecture, so we decided to run a brute-force search to look for counterexamples (we thank Geoff Hinton for suggesting us this idea). We uniformly randomly generated $M = 30$ centroids in the rectangle $[0,1] \times [0,0.7]$, mixing proportions $\pi_m \in (0,1)$ and isotropic covariance matrices with $\sigma_m \in [0.05, 0.15]$. Then we run the algorithm starting (a) from every centroid and (b) from every point in a grid of 100×70 of the rectangle. Call π_a and π_b the number of modes found in each case, respectively. We repeated the process 1500 times and considered only those cases where $\pi_a \neq \pi_b$; cases where $\pi_a = \pi_b$ cannot disprove the modes conjecture since by construction $\pi_a \leq M$. A difference $\pi_a \neq \pi_b$ was considered a false alarm if due to a single mode appearing as two or more with a small numerical difference[6]. We found 3 differences in the homoscedastic GM case (all false alarms) and 10 in the isotropic case (7 genuine, 3 false alarms). One of the genuine differences is shown in Fig. 3(right column): note how the green basin of attraction at the top

[6] The implementation of the algorithm considers that two modes are the same if their distance is less than a user parameter `min_diff` that has a very small value [15]. This helps to remove duplicated modes, but can occasionally fail.

right contains no centroids. The associated mode lies in a very flat area of the density, as indicated by the lack of contours[7]; the same happened in all other cases. We conclude that, in the isotropic case, it is possible (but rare) that a mode may not be reachable from any centroid.

In all our experiments the number of modes found by brute-force search π_b was $\leq M$. This reinforces our belief that the modes conjecture holds, or that if it does not, then it may fail only rarely. The results also show that the algorithm almost always finds all the modes when the component covariances are isotropic, perhaps always when they are equal.

Figure 4 shows that, unlike a Voronoi tessellation, the basins of attraction need be neither convex (left plot) nor connected sets (right plot). The points on the basin boundaries are either saddle points or minima, or converge to a saddle point. Note the following: (1) the basins often have very thin streaks extending for long distances, sandwiched between other basins; (2) one basin can be completely included in another; and (3) some points (typically minima) lie in the boundary of several basins simultaneously. Also, the sharper a mode is (e.g. for high π_m and low σ_m), the smaller its basin is. However, such small-basin modes will not be missed since they will lie near a centroid.

The brute-force search extension to 3D is computationally prohibitive.

5 Applications

The conjecture and mode-finding algorithms are relevant in statistical and machine learning applications such as function approximation, data visualisation, data reconstruction, clustering or image processing. The basic idea is that modes can be associated with important structure in an empirical distribution. We discuss the problems of regression and clustering.

5.1 Multivalued Regression and Data Reconstruction

In traditional nonlinear regression, one wants to derive a (parametric or non-parametric) mapping $\mathbf{y} = \mathbf{f}(\mathbf{x})$ given data pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^D \times \mathbb{R}^E$. The mapping $\mathbf{f}$ assigns a unique value $\mathbf{y}$ to every input $\mathbf{x}$; often some unimodal noise model is also assumed (e.g. Gaussian noise for a sum-squared error function). However, this is an unreasonable model if $p(\mathbf{y}|\mathbf{x})$ can be multimodal; typically this occurs when inverting a non-injective forward mapping such as $g(x) = x^2$. Representing $p(\mathbf{y}|\mathbf{x})$ as a mixture model has been proposed in a number of contexts. For example it arises with the mixture of experts model [21] (see also the mixture density networks [22, §6.4]), where the number of mixtures is chosen in some fashion. Also, the Nadaraya-Watson estimator [23] gives rise to an N-component mixture for $p(\mathbf{y}|\mathbf{x})$. Carreira-Perpiñán [24] proposed a flexible way to represent multivalued mappings by first estimating a probability density function $p(\mathbf{x}, \mathbf{y})$ for the joint variables from the training data, and then

[7] It might be argued that perhaps such modes are not really modes, but lie in the limit of numerical accuracy.

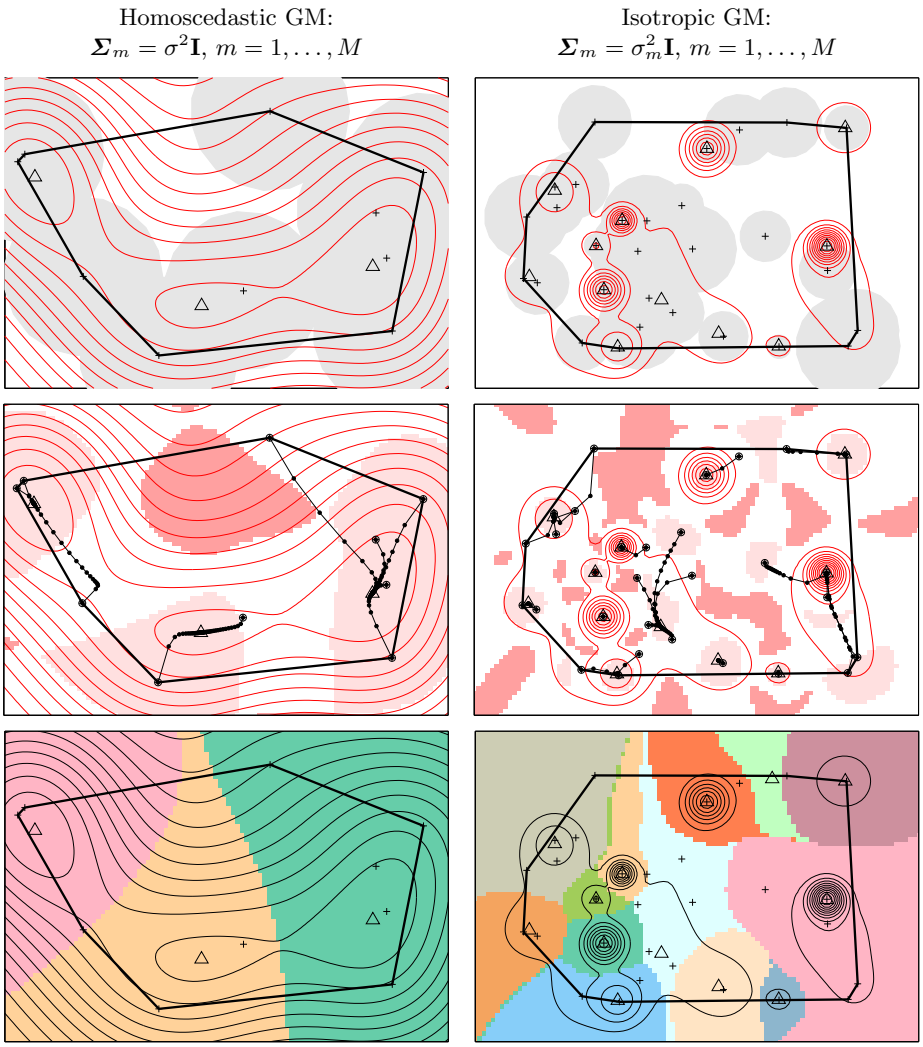

Fig. 3. The fixed-point iterative algorithm for exhaustive mode finding in 2D. The left column shows an example of homoscedastic GM ($\boldsymbol{\Sigma}_m = \sigma^2\mathbf{I}$) and the right one an example of isotropic GM ($\boldsymbol{\Sigma}_m = \sigma_m^2\mathbf{I}$). The latter is a very rare case where the algorithm did not find all modes (compare the top and bottom rows: in the top-row plot, a mode is missing at the top right). All parameters $\boldsymbol{\mu}_m$, π_m, σ_m and σ were drawn randomly. The GM modes are marked "△" and the GM centroids "+". The thick-line polygon is the convex hull of the centroids. *Top row*: contour plot of the GM density $p(\mathbf{x})$. Each original component is indicated by a grey disk of radius σ or σ_m centred on the corresponding mean vector $\boldsymbol{\mu}_m$ (marked "+"). *Middle row*: plot of the Hessian character (dark colour: positive definite; white: indefinite; light colour: negative definite). The search paths from the centroids are given. *Bottom row*: plot of the basins of attraction of each mode (i.e., the geometric locus of points that converge to each mode). Figures 3 and 4 may require to be viewed in colour to appreciate the different basins.

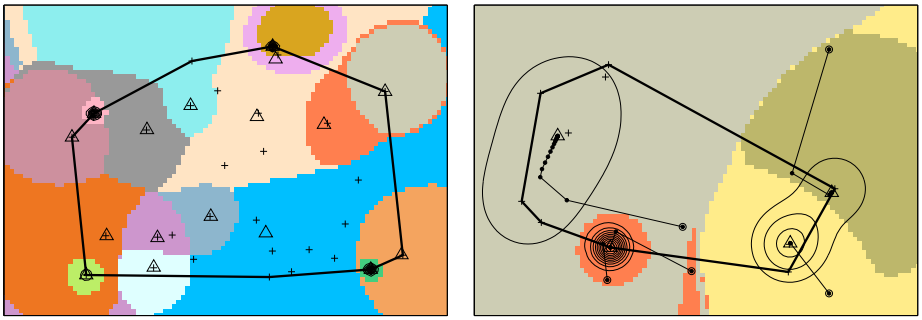

Fig. 4. Basins of attraction of each mode for examples that are heteroscedastic GMs with isotropic components. *Left*: basins may not be convex sets. *Right*: basins may not be connected sets (note the sample search paths).

defining a multivalued mapping $\mathbf{y} = \mathbf{f}(\mathbf{x})$ as the collection of modes of the conditional distribution $p(\mathbf{y}|\mathbf{x})$. It is computationally convenient to model $p(\mathbf{x}, \mathbf{y})$ as a homoscedastic GM[8], since then computing $p(\mathbf{y}|\mathbf{x})$ or any other conditional distribution is trivial, and we can use the algorithms of [15] (such as that of Sect. 4.1) to find the modes. Since (ideally at least) *every mode corresponds to a branch of the multivalued mapping and vice versa* it is of interest to locate *all* the modes of the conditional distribution, which leads us to the conjecture.

These ideas can be used to reconstruct missing data in a sequence of vectors $\mathbf{t}_1, \ldots, \mathbf{t}_N$ in a two-step procedure. First, at each vector in the sequence, one finds all the modes of the conditional distribution $p(\mathbf{t}_{n,\mathcal{M}}|\mathbf{t}_{n,\mathcal{P}})$, where $\mathbf{t}_{n,\mathcal{M}}$ (resp. $\mathbf{t}_{n,\mathcal{P}}$) means the missing variables (resp. present) at vector $\mathbf{t}_n$. This gives several candidate reconstructions for each vector. Second, a unique candidate at each n is selected by minimising a continuity constraint (such as the trajectory length) over the whole sequence. This results in a unique reconstruction of the whole sequence. This method was applied [2] to inverse mappings in speech (the acoustic-to-articulatory mapping) and robotics (the inverse kinematics).

5.2 Clustering

Given an unlabelled training set $\{\mathbf{x}_n\}_{n=1}^N \subset \mathbb{R}^D$, we want to obtain a clustering of these points and classify a new data point $\mathbf{x}$. One possible clustering approach is as follows. First, compute a kernel density estimate from the data of kernel K and window width $h > 0$ (which controls the amount of smoothing [23]):

$$p(\mathbf{x}; h) = \frac{1}{Nh^D} \sum_{n=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right). \tag{8}$$

[8] Or any other model that results in it, such as the generative topographic mapping [25] or kernel density estimation.

Then associate each mode of it with a cluster. If we define an iterative mode seeking algorithm such as gradient ascent or our EM algorithm, then we can assign a new point $\mathbf{x}$ to the mode to which the algorithm converges if started from $\mathbf{x}$. It is of interest to know how many modes exist at a given width h, which brings us to the conjecture when Gaussian kernels are used.

Perhaps the earliest proposal of this approach was the *mean-shift algorithm* of Fukunaga and Hostetler [26], recently extended in [27] and [28]. The mean-shift algorithm was defined as

$$\mathbf{x} \leftarrow \mathbf{m}(\mathbf{x}) = \frac{\sum_{n=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \mathbf{x}_n}{\sum_{n=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)}$$

where $\mathbf{m}(\mathbf{x}) - \mathbf{x}$ is called the mean shift. The algorithm was derived for the Epanechnikov kernel for computational convenience (since it has finite support) as gradient ascent on $\log p(\mathbf{x})$ with a variable step size; no convergence proof was given. For the Gaussian kernel it coincides with our algorithm for homoscedastic GMs of eq. (7) with $q(m|\mathbf{x}) = p(m|\mathbf{x})$—thus, *the mean-shift algorithm with the Gaussian kernel (and probably other kernels) is an EM algorithm*, which proves it has first-order convergence from any starting point. Comaniciu and Meer [28], in an image segmentation application, gave a different convergence proof for the mean-shift algorithm for certain isotropic kernels (including the Gaussian and Epanechnikov) and noted empirically its slow convergence for the Gaussian kernel. Note that the fact that the clusters defined by mean-shift may not be connected sets (Fig. 4) could be undesirable for some applications. Related clustering methods have been proposed [29, 30, 9, 31]. The mode trajectories in the scale space of h have also been used as a tool for data visualisation [32].

In scale-space clustering the mode-finding algorithms of [15] can also be used in a fast incremental way, where the modes at scale s_1 are found from the modes at scale $s_0 < s_1$ (rather than starting from every centroid). If the number of modes decreases with the scale, this will not miss any mode.

6 Conclusion

We have presented theoretical and experimental evidence for the conjecture that, in any dimension, the number of modes of a Gaussian mixture where all components are isotropic or equal cannot exceed the number of components (and proven it in 1D). It may hold even if Gaussian blurring of a delta mixture can create modes. A possible approach to (dis)prove the conjecture is to particularise Morse theory to Gaussian blurring of delta mixtures. Practically, it seems that the conjecture will hold for almost all isotropic Gaussian mixtures and that hill-climbing algorithms started from each centroid of the mixture will usually find all modes. The conjecture may also typically hold for mixtures of certain non-gaussian kernels even though these are known to create modes upon blurring. Our derivation of the fixed-point iterative algorithm (which can also be seen as a mean-shift algorithm) as an EM algorithm guarantees it has first-order convergence from any starting point.

Bibliography

[1] Carreira-Perpiñán, M.Á., Williams, C.K.I.: On the number of modes of a Gaussian mixture. Technical Report EDI–INF–RR–0159, School of Informatics, University of Edinburgh, UK (2003). Available online at `http://www.informatics.ed.ac.uk/publications/report/0159.html`.

[2] Carreira-Perpiñán, M.Á.: Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction. PhD thesis, Dept. of Computer Science, University of Sheffield, UK (2001)

[3] Carreira-Perpiñán, M.Á.: Mode-finding for mixtures of Gaussian distributions. Technical Report CS–99–03, Dept. of Computer Science, University of Sheffield, UK (1999), revised August 4, 2000. Available online at `http://www.dcs.shef.ac.uk/~miguel/papers/cs-99-03.html`.

[4] Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14** (2002) 1771–1800

[5] Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers Group, Dordrecht, The Netherlands (1994)

[6] Koenderink, J.J.: The structure of images. Biol. Cybern. **50** (1984) 363–370

[7] Babaud, J., Witkin, A.P., Baudin, M., Duda, R.O.: Uniqueness of the Gaussian kernel for scale-space filtering. IEEE Trans. on Pattern Anal. and Machine Intel. **8** (1986) 26–33

[8] Yuille, A.L., Poggio, T.A.: Scaling theorems for zero crossings. IEEE Trans. on Pattern Anal. and Machine Intel. **8** (1986) 15–25

[9] Roberts, S.J.: Parametric and non-parametric unsupervised cluster analysis. Pattern Recognition **30** (1997) 261–272

[10] Lifshitz, L.M., Pizer, S.M.: A multiresolution hierarchical approach to image segmentation based on intensity extrema. IEEE Trans. on Pattern Anal. and Machine Intel. **12** (1990) 529–540

[11] Kuijper, A., Florack, L.M.J.: The application of catastrophe theory to image analysis. Technical Report UU–CS–2001–23, Dept. of Computer Science, Utrecht University (2001). Available online at `ftp://ftp.cs.uu.nl/pub/RUU/CS/techreps/CS-2001/2001-23.pdf`.

[12] Kuijper, A., Florack, L.M.J.: The relevance of non-generic events in scale space models. In Heyden, A., Sparr, G., Nielsen, M., Johansen, P., eds.: Proc. 7th European Conf. Computer Vision (ECCV'02), Copenhagen, Denmark (2002)

[13] Damon, J.: Local Morse theory for solutions to the heat equation and Gaussian blurring. J. Diff. Equations **115** (1995) 368–401

[14] Silverman, B.W.: Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, B **43** (1981) 97–99

[15] Carreira-Perpiñán, M.Á.: Mode-finding for mixtures of Gaussian distributions. IEEE Trans. on Pattern Anal. and Machine Intel. **22** (2000) 1318–1323

[16] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the *EM* algorithm. Journal of the Royal Statistical Society, B **39** (1977) 1–38
[17] McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons (1997)
[18] Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review **26** (1984) 195–239
[19] Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proc. IEEE **86** (1998) 2210–2239
[20] Schölkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Müller, K.R., Rätsch, G., Smola, A.: Input space vs. feature space in kernel-based methods. IEEE Trans. Neural Networks **10** (1999) 1000–1017
[21] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3** (1991) 79–87
[22] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York, Oxford (1995)
[23] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Number 26 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York (1986)
[24] Carreira-Perpiñán, M.Á.: Reconstruction of sequential data with probabilistic models and continuity constraints. In Solla, S.A., Leen, T.K., Müller, K.R., eds.: Advances in Neural Information Processing Systems. Volume 12, MIT Press, Cambridge, MA (2000) 414–420
[25] Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The generative topographic mapping. Neural Computation **10** (1998) 215–234
[26] Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with application in pattern recognition. IEEE Trans. Inf. Theory **IT–21** (1975) 32–40
[27] Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. on Pattern Anal. and Machine Intel. **17** (1995) 790–799
[28] Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. on Pattern Anal. and Machine Intel. **24** (2002) 603–619
[29] Wong, Y.: Clustering data by melting. Neural Computation **5** (1993) 89–104
[30] Chakravarthy, S.V., Ghosh, J.: Scale-based clustering using the radial basis function network. IEEE Trans. Neural Networks **7** (1996) 1250–1261
[31] Leung, Y., Zhang, J.S., Xu, Z.B.: Clustering by scale-space filtering. IEEE Trans. on Pattern Anal. and Machine Intel. **22** (2000) 1396–1410
[32] Minnotte, M.C., Scott, D.W.: The mode tree: A tool for visualization of nonparametric density features. Journal of Computational and Graphical Statistics **2** (1993) 51–68