On the number of modes of a Gaussian mixture

Miguel Á. Carreira-Perpiñán
Dept. of Computer Science
University of Toronto
miguel@cs.toronto.edu

Christopher K. I. Williams
School of Informatics
University of Edinburgh
c.k.i.williams@ed.ac.uk

February 7, 2003

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.

Keywords: Mode finding, Gaussian mixtures, linear Gaussian scale space, kernel density estimation, EM algorithms, mean shift algorithms.

1 Introduction

We propose a mathematical conjecture about Gaussian mixtures: 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 Gaussian mixtures are the convolution of a delta mixture with a Gaussian kernel) and statistical smoothing (since Gaussian kernel density estimates are Gaussian mixtures). 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 Gaussian mixtures 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 section 2, and review related proof approaches in section 3. We show the convergence of an algorithm that tries to find all modes in section 4 and discuss relevant applications in section 5.

2 The conjecture

Consider a Gaussian mixture 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 $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$:

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

We write $p(\mathbf{x})$ and not $p(\mathbf{x}|\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M)$ because we assume that the parameters $\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$ have been estimated previously and their values fixed.

In general, there is no analytic expression for the modes of p, since we do not even know how many modes it has. Intuitively, it seems reasonable that the number of modes of p will be smaller than or equal to the number M of components in the mixture: 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 appear. Besides, modes should always appear inside the region enclosed by the component centroids—more precisely, in their convex hull. Based on this reasoning, Carreira-Perpiñán (2001) (see also Carreira-Perpiñán, 1999) proposed the following conjecture. First, let us recall that the convex hull of the vectors $\{\boldsymbol{\mu}_m\}_{m=1}^M$ is defined as the set

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

Conjecture 2.1. 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: $\boldsymbol{\Sigma}_m=\boldsymbol{\Sigma}\ \forall m=1,\dots,M$ *(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. Also, the conditions of the conjecture are necessary, and we give examples where either using non-isotropic covariances or non-Gaussian kernels results in additional modes.

Firstly note that all stationary (critical) points of (almost) all Gaussian mixtures are isolated, i.e., for each critical point there is a ball of positive radius containing no other critical points. This rules out submanifolds of modes such as a volcano in 2D, and is guaranteed by Morse theory (which relates critical points of smooth real functions on manifolds with the topology of the manifold; Milnor, 1963). Let a *Morse function* be a smooth (infinitely differentiable) function whose critical points are isolated, and note that a Gaussian mixture is smooth. Then we have that: all critical points of a Morse function are nondegenerate (the Hessian matrix at them is not singular); Morse functions form an open subset which is dense in the vector space of all smooth functions; and the set of non-Morse functions has lower dimension. Thus, Morse functions are generic, or typical (a small perturbation turns a non-Morse function into a Morse one).

2.1 Necessity of the conditions

Figure 1 (left plot) gives a simple example of a mixture with nonisotropic, different component covariance matrices that has more modes than components and the modes lie outside the convex hull of the centroids. Clearly, it is possible to construct more complicated examples where elongated components interact to create a variety of modes (right plot).

In general, the conjecture does not hold if the kernel $p(\mathbf{x}|m)$ is not Gaussian. This may seem counterintuitive, since one may expect that localised, tapering kernels would behave like the Gaussian. However, it is easy to construct counterexamples that show this is not so, where an additional mode appears where kernels interact (see fig. 2). Consider a 1D mixture of two identical kernels K with different centroids and call x the point at which both kernels cross. For a mode to exist at x we need $\lim_{h\to 0}(K(x-h)+K(x+h)-2K(x))<0$, or $K''(x)<0$ (i.e., the kernel is concave at x). If the maximum value of the kernel is more than $2K(x)$ there will also be two modes flanking the middle one. These conditions will be satisfied by many platykurtic kernels, such as the Epanechnikov or biweight kernels.

A related problem is whether convolution of a function with a given kernel can introduce new maxima in the function. In 1D, the necessity (and uniqueness among all kernels) that the kernel be Gaussian for maxima creations not to occur has been established in scale-space theory (see section 3.2). In the context of mixtures, this means that if the individual variances of the mixture kernels are increased, creation of new modes is to be expected generically for non-Gaussian kernels. Examples of mode creations in 1D abound in the literature (e.g. Yuille and Poggio, 1986), although for some kernels such examples may be difficult to construct (e.g. for the Cauchy kernel; Minnotte and Scott, 1993; Chaudhuri and Marron, 2000). Feller (1966, p. 164) also points out that the convolution of two nonsymmetric unimodal densities need not be unimodal (though the convolution of two symmetric unimodal densities in 1D is unimodal). However, as we discuss later for 2D Gaussian kernels, the creation of modes does not necessarily mean that the total number of modes will exceed the number of components.

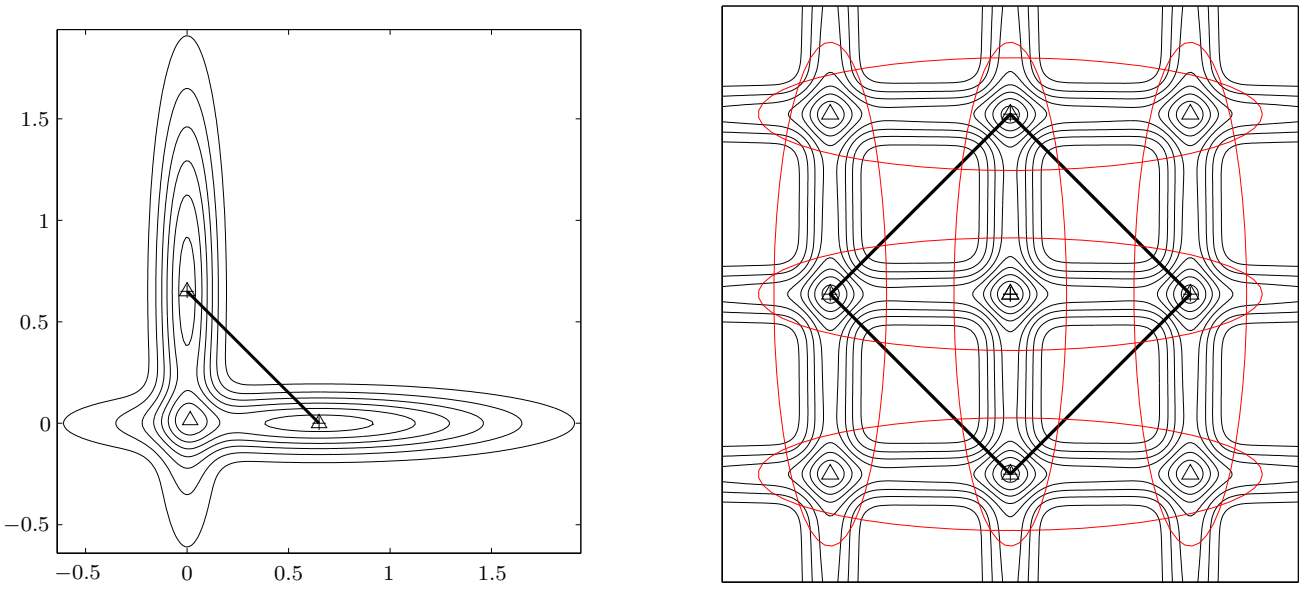

Figure 1: Mixtures in dimension $D \geq 2$ that have different, non-isotropic covariances do not generally verify conjecture 2.1. The left graph shows a contour plot for the bicomponent mixture $p(\mathbf{x}) = \sum_{m=1}^{2} \frac{1}{2} |2\pi\mathbf{\Sigma}_m|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T \mathbf{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)}$ with $\pi_1 = \pi_2 = \frac{1}{2}$, $\boldsymbol{\mu}_1 = \left(\begin{smallmatrix} 0.6 \\ 0 \end{smallmatrix}\right)$, $\boldsymbol{\mu}_2 = \left(\begin{smallmatrix} 0 \\ 0.6 \end{smallmatrix}\right)$, $\mathbf{\Sigma}_1 = \left(\begin{smallmatrix} 0.65 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$ and $\mathbf{\Sigma}_2 = \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.65 \end{smallmatrix}\right)$. This mixture 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 are possible that will result in a multiplicity of modes, as shown in the right graph (inspired by fig. 2 of Hinton, 2002): a new mode will appear where elongated components (indicated by the ellipses) intersect.

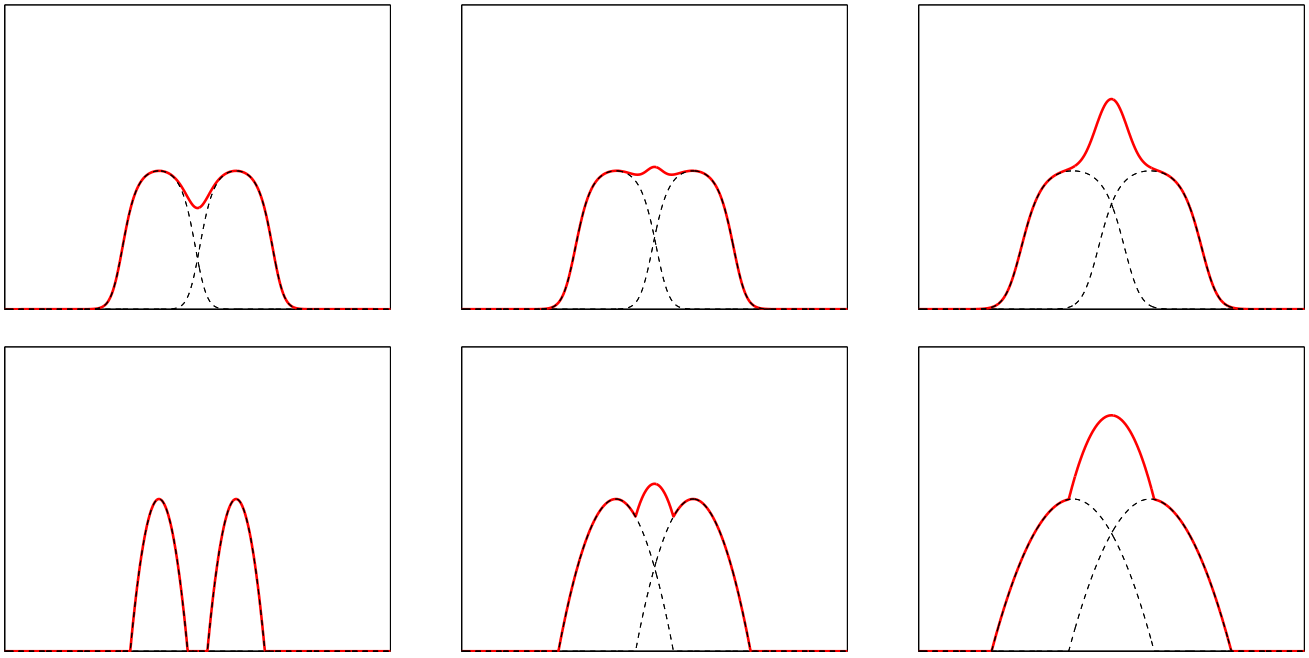

Figure 2: Mixtures of non-Gaussian kernels do not generally verify conjecture 2.1. Here, a 1D mixture $p(x) = \frac{1}{2}K\left(\frac{x-\mu_1}{\sigma}\right) + \frac{1}{2}K\left(\frac{x-\mu_2}{\sigma}\right)$ of two identical kernels K can have from one to three modes. We give two examples: top row, kernel $K(x) \propto \frac{1}{1+\frac{1}{10}e^{x^2}}$, with infinite support; bottom row, Epanechnikov kernel (eq. (12)), with compact support. The thick line represents the mixture and the dashed lines the two individual components.

2.2 The modes lie in the convex hull for any dimension D

All modes (in fact, all stationary points) lie in the convex hull of the centroids for the case of isotropic Gaussian mixtures. 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 by corollary 1 of Loog et al. (2001), stated as follows: for a nonnegative function f with compact support, all critical points of f convolved with an isotropic Gaussian are in the convex closure of the support of f. For a Gaussian mixture, f is a delta mixture, which verifies the conditions and whose support are the centroids.

Yet another proof is given by the following theorem from Alexander Heimel.

Theorem 2.1 (Heimel, 2000, pers. comm.). *Let $\{f_m\}_{m=1}^M$ be a set of M functions from $\mathbb{R}^D$ to $\mathbb{R}$ where each function can be written as $f_m(\mathbf{x}) = g_m(\|\mathbf{x}-\boldsymbol{\mu}_m\|)$ for strictly monotonically decreasing functions $\{g_m\}_{m=1}^M$ and points $\{\boldsymbol{\mu}_m\}_{m=1}^M$. Then all the maxima of the function $F \stackrel{\text{def}}{=} \sum_{m=1}^M f_m$ are inside the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^M$.*

Proof. By contradiction. Call $\mathcal{H}$ the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^M$. Suppose F takes a maximum at $\mathbf{x} \notin \mathcal{H}$. Call $\mathbf{u}$ the closest point in $\mathcal{H}$ to $\mathbf{x}$. Then, it is easy to see that any point $\mathbf{x}'$ in the segment between $\mathbf{x}$ and $\mathbf{u}$ is closer to all points of $\mathcal{H}$ than $\mathbf{x}$ is. Thus $f_m(\mathbf{x}') > f_m(\mathbf{x})\ \forall m = 1,\dots,M$ and so $F(\mathbf{x}') > F(\mathbf{x})$. Since in every neighbourhood of $\mathbf{x}$ there are points for which F is larger, $\mathbf{x}$ cannot be a maximum. □

2.3 The homoscedastic case is equivalent to the homoscedastic isotropic one

The following theorem shows that the modes problem for a homoscedastic mixture with a given arbitrary covariance $\boldsymbol{\Sigma}$ is equivalent to that of another homoscedastic mixture 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 2.2. *The mixtures $p(\mathbf{x}) = \sum_{m=1}^M \pi_m |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)}$ (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), with Jacobian $\left|\frac{d\mathbf{u}}{d\mathbf{x}}\right| = |\boldsymbol{\Lambda}|^{\frac{1}{2}}$, and define $\boldsymbol{\nu}_m \stackrel{\text{def}}{=} \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\boldsymbol{\mu}_m$. The density of $\mathbf{u}$ is then $p(\mathbf{u}) = p(\mathbf{x})\left|\frac{d\mathbf{u}}{d\mathbf{x}}\right|^{-1} = \sum_{m=1}^M \pi_m (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{u}-\boldsymbol{\nu}_m\|^2}$.

The gradient of p can be written:

$$\frac{\partial p}{\partial x_d} = \sum_{e=1}^D \frac{\partial p}{\partial u_e}\frac{\partial u_e}{\partial x_d} = \sum_{e=1}^D \frac{\partial p}{\partial u_e}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\right)_{ed} \Longrightarrow \nabla_{\mathbf{x}} p = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\nabla_{\mathbf{u}} p.$$

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 2.2 shows that case 2 of conjecture 2.1 is a particular case of case 3 (case 1 is also a particular case of case 3, obviously).

2.4 The conjecture holds for $D = 1$

We can prove this using the scale-space theory proofs of non-creation of maxima with Gaussian blurring (section 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 Gaussian mixture. 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 2.3. *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 Gaussian mixture 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 $m = 1, \dots, M-1$ 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 Gaussian mixture 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 section 3.2, no new maxima can appear, and so the original mixture p has M modes at most. □

Theorem 2.3 generalises the result of Silverman (1981) described in section 3.3 which proves that a *homoscedastic* mixture in 1D has at most M modes. The following corollary results from the fact that all marginal and conditional distributions of a Gaussian mixture (of arbitrary covariances) are also Gaussian mixtures (Mardia et al., 1979, p. 63).

Corollary 2.4. *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.

3.1 Nonlinear system of equations for the stationary points of the density

Carreira-Perpiñán (2001) approached the problem by trying to determine the stationary (or critical) points of the Gaussian mixture density p as follows. Consider the case with $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma},\ m = 1, \dots, M$ (homoscedastic mixture) 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 $x_1, \dots, x_D \subset \mathbb{R}$. 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:

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

as a nonlinear system of M equations and M unknowns $\lambda_1, \dots, \lambda_M \in (0,1)$ 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

$$\lambda = \frac{1}{1 + e^{-\alpha(\lambda - \lambda_0)}} \qquad \text{with} \qquad \begin{cases} \alpha = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0 \\ \lambda_0 = \frac{1}{2} + \frac{1}{\alpha}\log\frac{1-\pi}{\pi} \in (-\infty, \infty) \end{cases}$$

which can have at most 3 roots in $(0,1)$, as can be easily seen geometrically in fig. 3, 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.

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

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

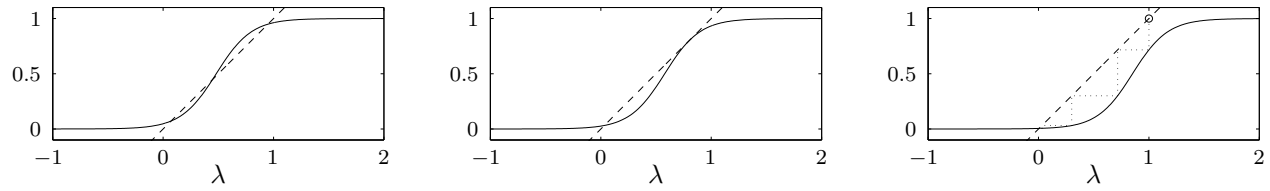

Figure 3: Three possible cases for the solutions of the equation $\lambda = f(\lambda)$, where $f(\lambda) = \frac{1}{1+e^{-\alpha(\lambda-\lambda_0)}}$, with $\alpha > 0$. The solid line corresponds to $f(\lambda)$ and the dashed one to λ. The right figure also shows in dotted line the sequence of fixed-point iterations starting from $\lambda = 1$ (marked "$\circ$"), converging to a fixed point slightly larger than 0.

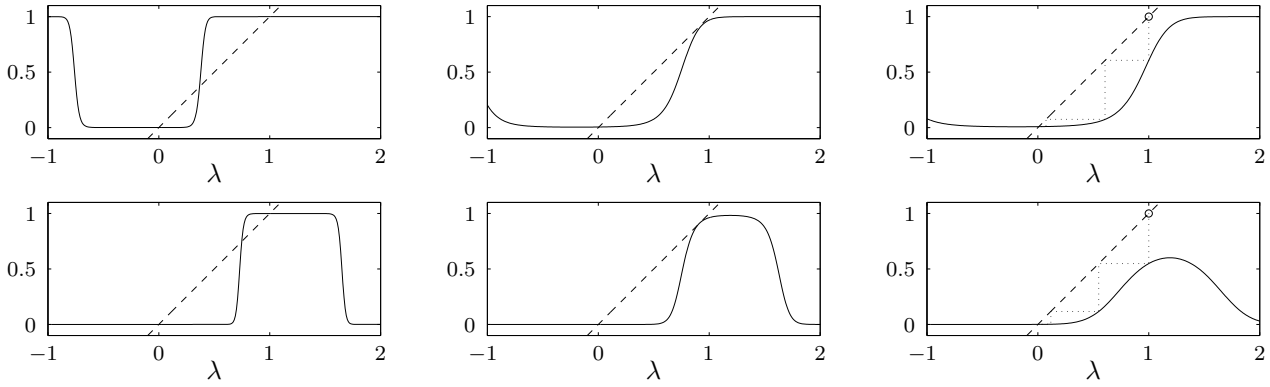

Figure 4: Six possible cases for the solutions of the equation $\lambda = f(\lambda)$, where $f(\lambda) = \frac{1}{1+\beta e^{-\alpha(\lambda-\lambda_0)^2}}$, with $\alpha \neq 0$, $\beta > 0$ and $\lambda_0 \notin [0, 1]$. The solid line corresponds to $f(\lambda)$ and the dashed one to λ. The top row is for $\alpha > 0$ and the bottom row for $\alpha < 0$. The right figures also show in dotted line the sequence of fixed-point iterations starting from $\lambda = 1$ (marked "$\circ$"), converging to a fixed point slightly larger than 0.

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. For the case $M = 2$, the transcendental equation for λ to which (3) reduces is:

$$\lambda = \frac{1}{1 + \beta e^{-\alpha(\lambda-\lambda_0)^2}} \qquad \text{with} \qquad \begin{cases} s = \frac{\sigma_1}{\sigma_2} > 0, \neq 1 \text{ (ignoring } \sigma_1 = \sigma_2) \\ \alpha = \frac{1}{2} \left\| \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{\sigma_2} \right\|^2 \left(\frac{s^2-1}{s^2} \right) \neq 0 \text{ (ignoring } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2) \\ \beta = \left(\frac{1-\pi}{\pi} \right) s^2 e^{\frac{\alpha s^2}{(1-s^2)^2}} > 0 \\ \lambda_0 = \frac{1}{1-s^2} \notin [0, 1] \end{cases}$$

and so there are 3 stationary points at most (see fig. 4), but again the system becomes difficult in general.

A further problem with this approach is that the equations apply to all stationary points (maxima, minima and saddle-points) 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 $q(m|\mathbf{x}) < 1$ (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 (Lindeberg, 1994) 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. The family of functions is the starting point for the analysis and segmentation of the image, for example by determining the lifetime in scale

space of blobs, each associated with a maximum of $I(\mathbf{x}; s)$. As the scale increases, fine detail should disappear, and so the maxima in the image tend to decrease due to merging with other critical points. In order to track the maxima from small to large scales, and generally to have a non-committed visual front-end, it is desirable to use a convolution kernel that does not introduce new maxima as the scale increases—since these would be spurious detail due to the kernel rather than reflecting true structure in the image.

In this context, several researchers (including among others[1] Koenderink, 1984; Babaud et al., 1986; Yuille and Poggio, 1986; Roberts, 1997) 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 $I(\mathbf{x}; 0) = I(\mathbf{x})$:

$$\frac{\partial I}{\partial s} = \frac{1}{2}\nabla^2_{\mathbf{x}} I = \frac{1}{2}\sum_{d=1}^{D} \frac{\partial^2 I}{\partial x_d^2}.$$

(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[2] 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 (1990): 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 (2001, 2002), 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 (1995) 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). However, it cannot be applied directly because the family is not unconstrained, but must obey the diffusion equation. Damon showed that maxima 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 Gaussian mixture 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 Gaussian mixture 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 2.3 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. 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 Lifshitz and Pizer (1990). And even if new modes can appear when blurring a Gaussian mixture, 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 mixture with only a few modes before and after the creation. Thus, the truth of our conjecture does not necessarily imply the non-creation of maxima upon Gaussian blurring in dimension larger than 1. A brute-force search has failed to find counterexamples of the conjecture (see section 4.5). Perhaps an approach based on catastrophe theory but restricted to initial images which are delta mixtures would resolve the question.

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 section 2.

[1] Note that Roberts (1997) specifically dealt with Gaussian mixtures (i.e., blurring of delta mixtures).

[2] These points are called degenerate critical points, top-points, bifurcation points or catastrophes.

3.3 Kernel density estimation in 1D

Given a data sample $\{x_n\}_{n=1}^N \subset \mathbb{R}$, Silverman (1981) considers the 1D Gaussian kernel density estimate $p(x;h)$ (section 5.2). This is, of course, a homoscedastic isotropic Gaussian mixture 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.

3.4 Other results

Related results have been proven, in a different way, in the literature. Behboodian (1970) shows that for $M = 2$ and $D = 1$, with no restriction on σ_1 and σ_2, $p(x)$ has one, two or three stationary points which all lie in the convex hull of $\{\mu_1, \mu_2\}$. Konstantellos (1980) gives a necessary condition for unimodality for two cases: $M = 2$, $\pi_1 = \pi_2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ and $D > 1$; and $M = 2$ and $D = 2$, with no restriction on π_m, $\mathbf{\Sigma}_m$. These results are very particular and do not seem easily generalisable.

4 Algorithms for finding all the modes

We now turn to the algorithmic question of finding all the modes of a Gaussian mixture. This is practically important in applications. Given that no direct solution exists, we need to use numerical iterative methods. Carreira-Perpiñán (2000a) suggested starting a mode-seeking algorithm from every centroid to locate all the modes. He gave two hill-climbing algorithms applicable to Gaussian mixtures with components of arbitrary covariance: a gradient-quadratic one and a fixed-point iteration one. Here we deal only with the latter because it defines 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) and restate it as a dynamical system.

4.1 The fixed-point iteration algorithm as an EM algorithm

By equating the gradient of the Gaussian mixture density to zero, using Bayes' theorem and rearranging we obtain a fixed-point iterative scheme (Carreira-Perpiñán, 2000a):

$$\begin{aligned} \mathbf{g} &= \sum_{m=1}^{M} p(\mathbf{x},m)\mathbf{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x}) = \mathbf{0} \\ &\Longrightarrow \mathbf{x}^{(\tau+1)} = \mathbf{f}(\mathbf{x}^{(\tau)}) \text{ with } \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} \left(\sum_{m=1}^{M} p(m|\mathbf{x})\mathbf{\Sigma}_m^{-1} \right)^{-1} \sum_{m=1}^{M} p(m|\mathbf{x})\mathbf{\Sigma}_m^{-1}\boldsymbol{\mu}_m. \end{aligned} \tag{4}$$

Following a suggestion from the second author, Carreira-Perpiñán (2001) showed that this algorithm can also be derived as an *expectation-maximisation (EM) algorithm* (Dempster et al., 1977; McLachlan and Krishnan, 1997) as follows[3]. Consider the following density model with parameters $\mathbf{v} = (v_1, \ldots, v_D)^T$ and fixed $\{\pi_m, \boldsymbol{\mu}_m, \mathbf{\Sigma}_m\}_{m=1}^M$:

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

That is, $\mathbf{x}|\mathbf{v}$ is a D-dimensional Gaussian mixture where component m has mixing proportion π_m (fixed), mean vector $\boldsymbol{\mu}_m - \mathbf{v}$ ($\boldsymbol{\mu}_m$ fixed) and covariance matrix $\mathbf{\Sigma}_m$ (fixed). Varying $\mathbf{v}$ results in a rigid translation of the whole mixture 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 $z_n \in \{1, \ldots, M\}$ the (unknown) index of the mixture component that generated data point $\mathbf{x}_n$. Then:

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

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

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

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

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:

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

As a function of $\mathbf{v}$, $p(\mathbf{x}_n|z_n,\mathbf{v})$ is a Gaussian density of mean $\boldsymbol{\mu}_{z_n} - \mathbf{x}_n$ and covariance $\boldsymbol{\Sigma}_{z_n}$, so we get

$$\frac{\partial p(\mathbf{x}_n|z_n,\mathbf{v})}{\partial \mathbf{v}} = p(\mathbf{x}_n|z_n,\mathbf{v}) \boldsymbol{\Sigma}_{z_n}^{-1}(\boldsymbol{\mu}_{z_n} - \mathbf{x}_n - \mathbf{v})$$

and so 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).

The EM algorithm for Gaussian mixtures converges to a local optimum from any starting point (Dempster et al., 1977; Redner and Walker, 1984; McLachlan and Krishnan, 1997). Specifically, at every iteration τ, the iterative scheme (6) will either increase the log-likelihood or leave it unchanged. The log-likelihood is

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

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 $\mathbf{x}$. In principle, though, convergence can occur to a saddle point or to a minimum as well as to a mode (in the very unlikely case where the initial value is at a minimum, the scheme will remain stuck at it). 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.

These issues can be demonstrated with the following example. Consider an isotropic Gaussian mixture with $M = 2$ components in $D = 2$ dimensions, with $\pi_1 = \pi_2 = \frac{1}{2}$, $\sigma_1 = \sigma_2 = 1$ and $\boldsymbol{\mu}_1 = \left(\begin{smallmatrix} -1 \\ 0 \end{smallmatrix}\right) = -\boldsymbol{\mu}_2$. This mixture has two modes and a saddle point at the origin, to which any starting point of the form $\left(\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right)$ will converge, because the EM iteration of eq. (7) becomes $x_1^{(\tau+1)} = 0$ and $x_2^{(\tau+1)} = f(x_2^{(\tau)}) \stackrel{\text{def}}{=} \tanh(x_2^{(\tau)})$ $\forall \tau = 1, 2, 3, \ldots$. A small perturbation of x_1 takes it away to one of the two modes. The convergence is first-order since $f'(0) \neq 0$ (see below). The same mixture in 1D (discarding the x_2 variable) has a minimum at the origin.

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

4.2 Particular cases

In the important case of homoscedastic mixtures, the fixed-point scheme reduces to the extremely simple form

$$\mathbf{x}^{(\tau+1)} = \sum_{m=1}^{M} p(m|\mathbf{x}^{(\tau)})\boldsymbol{\mu}_m \qquad p(m|\mathbf{x}) = \frac{\pi_m e^{-\frac{1}{2}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}_m}{\sigma}\right\|^2}}{\sum_{m'=1}^{M}\pi_{m'}e^{-\frac{1}{2}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}_{m'}}{\sigma}\right\|^2}} \tag{7}$$

where $p(m|\mathbf{x})$ is the posterior probability or responsibility of component m given point $\mathbf{x}$. Thus, the new point $\mathbf{x}^{(\tau+1)}$ is the conditional mean of the mixture under the current point $\mathbf{x}^{(\tau)}$. This is formally akin to clustering by deterministic annealing (Rose, 1998), to algorithms for finding pre-images in kernel-based methods (Schölkopf et al., 1999) and to mean-shift algorithms (section 5.2).

In the case of isotropic mixtures we obtain another very simple form:

$$\mathbf{x}^{(\tau+1)} = \sum_{m=1}^{M} q(m|\mathbf{x}^{(\tau)})\boldsymbol{\mu}_m \qquad q(m|\mathbf{x}) = \frac{p(m|\mathbf{x})\sigma_m^{-2}}{\sum_{m'=1}^{M}p(m'|\mathbf{x})\sigma_{m'}^{-2}} \tag{8}$$

where the $q(m|\mathbf{x})$ values are the responsibilities $p(m|\mathbf{x})$ reweighted by the inverse variance and renormalised.

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 (Redner and Walker, 1984, eq. (5.3)).

4.3 Speed of convergence

If a sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ converges to $\mathbf{x}^*$ then the convergence rate is said to be of order p if there exists a positive constant r such that, for all τ sufficiently large, $\left\|\mathbf{x}^{(\tau+1)} - \mathbf{x}^*\right\| / \left\|\mathbf{x}^{(\tau)} - \mathbf{x}^*\right\|^p \leq r$ and, for $p = 1$, $r < 1$ (Nocedal and Wright, 1999, pp. 28ff). General properties of the EM algorithm (McLachlan and Krishnan, 1997) show that the scheme (4) quickly moves to an area of high probability inside the convex hull of the centroids, but that its convergence is linear (first-order) and so is very slow. Both facts were experimentally confirmed by Carreira-Perpiñán (2000a) for homoscedastic mixtures; fig. 5 demonstrates it for isotropic mixtures. Note the slow crawl along ridges of the density, as well as the fact that the iterates may be attracted to saddle points, to then deviate towards a mode. Since convergence near the modes is slow, it will be more convenient to switch to the quadratic algorithm of Carreira-Perpiñán (2000a) as soon as the Hessian becomes negative definite, or the quotient of the distances between successive iterates becomes approximately constant.

We can also give a direct indication of the linear rate of convergence of the algorithm for the homoscedastic 1D case as follows. Consider

$$f(x) = \sum_{m=1}^{M} p(m|x)\mu_m = \frac{M_1(x)}{M_0(x)} \qquad M_p(x) \stackrel{\text{def}}{=} \sum_{m=1}^{M} \pi_m e^{-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma}\right)^2}\mu_m^p$$

the fixed-point iterative mapping from eq. (7), assumed to converge to x^*. Then, the quotient of the successive error magnitudes is:

$$\frac{\left|x^{(\tau+1)} - x^*\right|}{\left|x^{(\tau)} - x^*\right|} = \frac{\left|f(x^{(\tau)}) - f(x^*)\right|}{\left|x^{(\tau)} - x^*\right|} = |f'(\xi)|$$

where we have Taylor-expanded f around x^* up to first order and ξ is a point between $x^{(\tau)}$ and x^*. Now, after some algebra, we obtain

$$f'(\xi) = \frac{M_0(\xi)M_2(\xi) - M_1^2(\xi)}{\sigma^2 M_0^2(\xi)}$$

which in general is not zero, and so the rate of convergence is linear. However, in the case where the mixture components are very separated, we can disregard the effect of all components except for the closest one to x^* (say it is the mth), so that

$$M_p(\xi) \approx \pi_m e^{-\frac{1}{2}\left(\frac{\xi-\mu_m}{\sigma}\right)^2}\mu_m^p \Longrightarrow f'(\xi) \approx 0$$

and so the method becomes superlinear. This is similar to a result from Xu and Jordan (1996) for the EM algorithm when used to estimate the parameters π_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ from a data sample.

4.4 Dynamical systems view

If we consider the iteration index as a continuous variable, the fixed-point iterative algorithm of eq. (4) can also be written as a dynamical system:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) - \mathbf{x} = \left(\sum_{m=1}^{M} p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} \right)^{-1} \nabla \log p(\mathbf{x}) \tag{9}$$

where the dot above a variable will indicate differentiation with respect to the time τ, i.e., $\dot{\mathbf{x}} \stackrel{\text{def}}{=} \frac{d\mathbf{x}}{d\tau}$. For the case where $\boldsymbol{\Sigma}_m = \sigma^2 \mathbf{I}$ for all m this becomes

$$\dot{\mathbf{x}} = \nabla \big(\sigma^2 \log p(\mathbf{x}) \big). \tag{10}$$

Since maximising $p(\mathbf{x})$ is equivalent to minimising $-\sigma^2 \log p(\mathbf{x})$, we can define a Lyapunov function (Wiggins, 1990)

$$V(\mathbf{x}) = \sigma^2 \log \frac{p(\mathbf{x}^*)}{p(\mathbf{x})}$$

in an open neighbourhood U of every minimum $\mathbf{x}^*$ of $-\sigma^2 \log p(\mathbf{x})$ (i.e., every fixed point of $\mathbf{f}$). V verifies:

1. $V(\mathbf{x}^*) = 0$ and $V(\mathbf{x}) > 0 \; \forall \mathbf{x} \in U \setminus \{\mathbf{x}^*\}$, i.e., V is positive definite in $U \setminus \{\mathbf{x}^*\}$.
2. $\dot{V} = \sum_{d=1}^{D} \frac{\partial V}{\partial x_d} \dot{x_d} = \dot{\mathbf{x}}^T \nabla V = \nabla \big(\sigma^2 \log p(\mathbf{x}) \big) \big(-\nabla \big(\sigma^2 \log p(\mathbf{x}) \big) \big) < 0 \; \forall \mathbf{x} \in U \setminus \{\mathbf{x}^*\}$ and equal to 0 at $\mathbf{x}^*$.

So, for the dynamical system of eq. (10), V is a strict Lyapunov function and $\mathbf{x}^*$ is an asymptotically stable point. Thus, the dynamical system converges from any starting point $\mathbf{x}$ in the neighbourhood U to the fixed point $\mathbf{x}^*$ (i.e., it has no cycles or chaotic behaviour). Besides, the convergence near the fixed point is exponential.

Unfortunately, finding a Lyapunov function for the general case (9) is more difficult.

4.5 Brute-force search for counterexamples

Whether starting the algorithm from each centroid can indeed find all modes depends, of course, 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, since they do not have to 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 saddle-points, 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[4]. We found 3 differences in the homoscedastic mixture case (all false alarms) and 10 in the isotropic case (7 genuine, 3 false alarms). One of the genuine differences is shown in fig. 5 (right column): note how the green basin of attraction at the top right contains no centroids. The associated mode lies in a very flat area of the density, as indicated by the lack of contours[5]; the same happened in all other cases. We then conclude that, in the isotropic case, the claim that every mode can be reached from some centroid does not hold, but that such occurrences are rare.

[4] 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 (Carreira-Perpiñán, 2000a). This helps to remove duplicated modes, but can occasionally fail.

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

In all of 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.

Fig. 6 shows[6] that, unlike a Voronoi tessellation, the basins of attraction need be neither convex (left plot) nor connected sets (right plot). More generally, it is a neat demonstration of this fact for EM algorithms (see pp. 94ff in McLachlan and Krishnan (1997) for another example with the log-likelihood of a Student-t distribution). 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 takes several hours in a state-of-the-art workstation because the algorithm must be started from every grid point. This makes its extension to 3D or higher 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 nonparametric) 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 $g(x) = x^2$. In this case, we want the mapping $\mathbf{f}$ to assign one or more values to any given $\mathbf{x}$ (multivalued regression). 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 (Jacobs et al., 1991; see also the mixture density networks of Bishop, 1995, §6.4), where the number of mixtures is chosen in some fashion. Also, the Nadaraya-Watson estimator gives rise to an N-component mixture for $p(\mathbf{y}|\mathbf{x})$ (though the estimator itself is defined by a unique value, the conditional mean).

Carreira-Perpiñán (2000b) 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 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 Gaussian mixture[7], since then computing $p(\mathbf{y}|\mathbf{x})$ or any other conditional distribution is trivial, and we can use the algorithms of Carreira-Perpiñán (2000a) (such as that of section 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. Carreira-Perpiñán (2001) applied this method 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

[6] Figures 5 and 6 may require to be viewed in colour to appreciate the different basins.

[7] Or any other model that results in it, such as the generative topographic mapping (Bishop et al., 1998) or kernel density estimation.

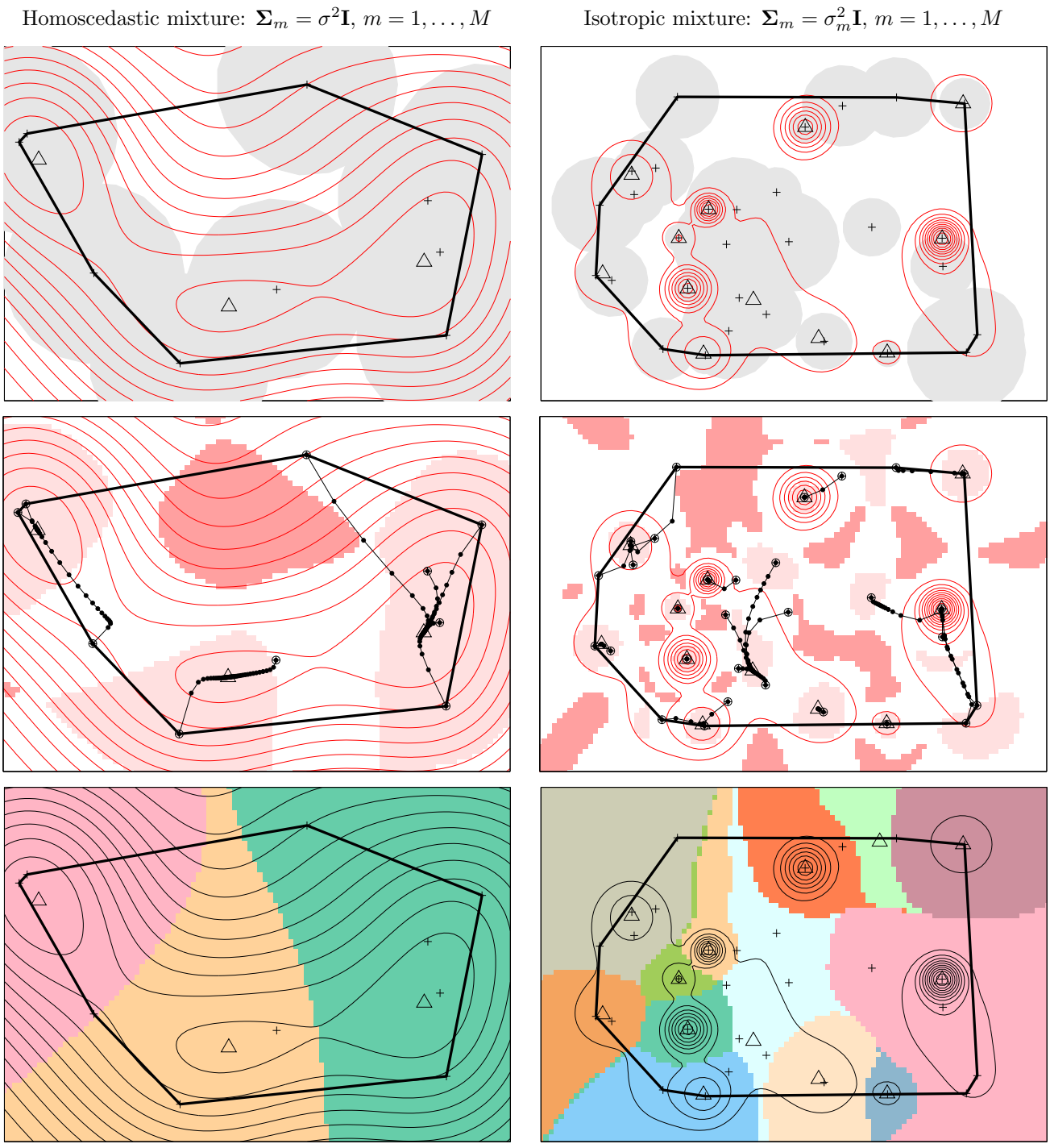

Figure 5: The fixed-point iterative algorithm for exhaustive mode finding in 2D. The left column shows an example of homoscedastic mixture ($\mathbf{\Sigma}_m = \sigma^2\mathbf{I}$) and the right one an example of isotropic mixture ($\mathbf{\Sigma}_m = \sigma_m^2\mathbf{I}$). The latter is a very rare 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 mixture modes are marked "△" and the mixture centroids "+". The thick-line polygon is the convex hull of the centroids. *Top row*: contour plot of the Gaussian mixture 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).

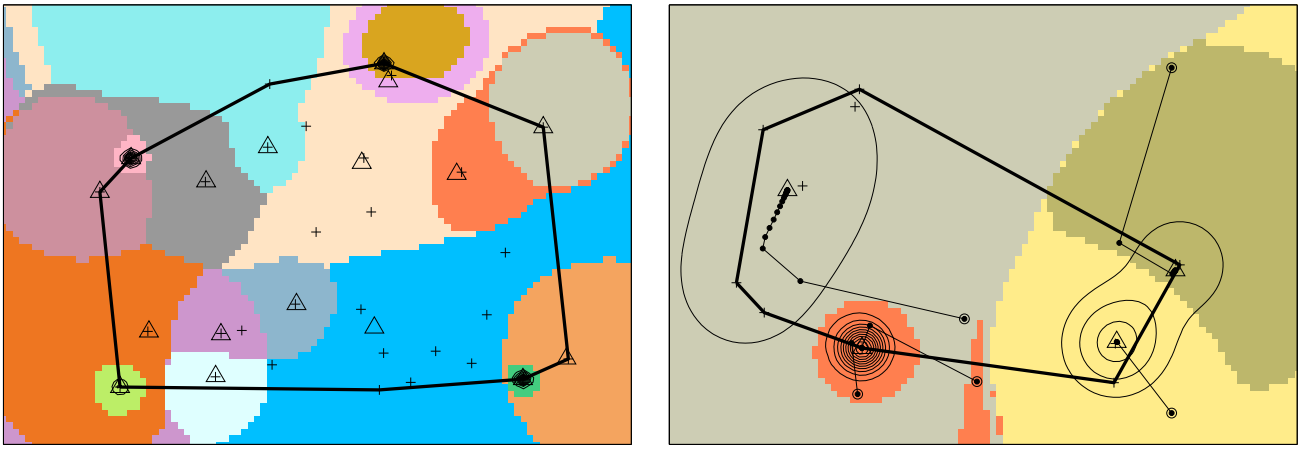

Figure 6: Basins of attraction of each mode for the fixed-point iterative algorithm in 2D. The plots are like those in the bottom row of fig. 5. Both examples are heteroscedastic mixtures with isotropic components. *Left*: basins may not be convex sets. *Right*: basins may not be connected sets (note the sample search paths).

data, of kernel K and window width $h > 0$ (which controls the amount of smoothing; Silverman, 1986):

$$p(\mathbf{x}; h) = \frac{1}{Nh^D} \sum_{n=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right). \tag{11}$$

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 (1975), recently extended by Cheng (1995) and Comaniciu and Meer (2002). 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

$$K(\mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} 1 - \|\mathbf{x}\|^2, & \|\mathbf{x}\| < 1 \\ 0, & \|\mathbf{x}\| \geq 1 \end{cases} \tag{12}$$

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 mixtures of eq. (7)—thus, *the mean-shift algorithm with the Gaussian kernel (and probably other kernels) is an EM algorithm*, which proves it has global convergence of first order. Comaniciu and Meer (2002), 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. 6) could be undesirable for some applications. Related clustering methods have been proposed by Wong (1993); Chakravarthy and Ghosh (1996); Roberts (1997); Leung et al. (2000). Minnotte and Scott (1993) used the mode trajectories in the scale space of h as a tool for data visualisation.

In scale-space clustering the mode-finding algorithms of Carreira-Perpiñán (2000a) 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 that the number of modes of a Gaussian mixture in any dimension where all components are isotropic cannot exceed the number of components. This may hold even if

Gaussian blurring of a delta mixture can occasionally create modes. A possible approach to resolve the question is to particularise Morse theory to Gaussian blurring of delta mixtures. Practically though it seems that the conjecture will hold almost always 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 for mode finding (which can also be seen as a mean-shift algorithm) as an EM algorithm guarantees it has global convergence of first order.

References

J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda. Uniqueness of the Gaussian kernel for scale-space filtering. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 8(1):26–33, Jan. 1986.

J. Behboodian. On the modes of a mixture of two normal distributions. *Technometrics*, 12(1):131–139, Feb. 1970.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, Oxford, 1995.

C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, Jan. 1998.

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

M. Á. Carreira-Perpiñán. Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22(11):1318–1323, Nov. 2000a.

M. Á. Carreira-Perpiñán. Reconstruction of sequential data with probabilistic models and continuity constraints. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 414–420. MIT Press, Cambridge, MA, 2000b.

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

S. V. Chakravarthy and J. Ghosh. Scale-based clustering using the radial basis function network. *IEEE Trans. Neural Networks*, 7(5):1250–1261, Sept. 1996.

P. Chaudhuri and J. S. Marron. Scale space view of curve estimation. *Annals of Statistics*, 28(2):408–428, Apr. 2000.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 17(8): 790–799, Aug. 1995.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 24(5):603–619, May 2002.

J. Damon. Local Morse theory for solutions to the heat equation and Gaussian blurring. *J. Diff. Equations*, 115 (2):368–401, Jan. 20 1995.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

W. Feller. *An Introduction to Probability Theory and Its Applications*. Number II in Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1966.

K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Trans. Inf. Theory*, IT–21(1):32–40, Jan. 1975.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8): 1771–1800, Aug. 2002.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

J. J. Koenderink. The structure of images. *Biol. Cybern.*, 50:363–370, 1984.

A. C. Konstantellos. Unimodality conditions for Gaussian sums. *IEEE Trans. Automat. Contr.*, AC–25(4): 838–839, Aug. 1980.

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

A. Kuijper and L. M. J. Florack. The relevance of non-generic events in scale space models. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. 7th European Conf. Computer Vision (ECCV'02)*, Copenhagen, Denmark, May 28–31 2002.

Y. Leung, J.-S. Zhang, and Z.-B. Xu. Clustering by scale-space filtering. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22(12):1396–1410, Dec. 2000.

L. M. Lifshitz and S. M. Pizer. A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 12(6):529–540, June 1990.

T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers Group, Dordrecht, The Netherlands, 1994.

M. Loog, J. J. Duistermaat, and L. M. J. Florack. On the behavior of spatial critical points under Gaussian blurring. A folklore theorem and scale-space constraints. In M. Kerckhove, editor, *Scale-Space and Morphology in Computer Vision*, volume 2106 of *Lecture Notes in Computer Science*, pages 183–192. Springer-Verlag, Berlin, 2001.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics Series. Academic Press, New York, 1979.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1997.

J. Milnor. *Morse Theory*. Annals of Mathematics Studies. Princeton University Press, Princeton, 1963.

M. C. Minnotte and D. W. Scott. The mode tree: A tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, 2:51–68, 1993.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, New York, 1999.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, Apr. 1984.

S. J. Roberts. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272, Feb. 1997.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, Nov. 1998.

B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola. Input space vs. feature space in kernel-based methods. *IEEE Trans. Neural Networks*, 10(5):1000–1017, Sept. 1999.

B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B*, 43(1):97–99, 1981.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Number 26 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1986.

S. R. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Texts in Applied Mathematics. Springer-Verlag, New York, 1990.

Y. Wong. Clustering data by melting. *Neural Computation*, 5(1):89–104, Jan. 1993.

L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, Jan. 1996.

A. L. Yuille and T. A. Poggio. Scaling theorems for zero crossings. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 8(1):15–25, Jan. 1986.