Mode-finding for mixtures of Gaussian distributions

Miguel Á. Carreira-Perpiñán*
Dept. of Computer Science, University of Sheffield, UK
M.Carreira@dcs.shef.ac.uk

Technical Report CS–99–03

March 31, 1999 (revised August 4, 2000)

Abstract

I consider the problem of finding all the modes of a mixture of multivariate Gaussian distributions, which has applications in clustering and regression. I derive exact formulas for the gradient and Hessian and give a partial proof that the number of modes cannot be more than the number of components, and are contained in the convex hull of the component centroids. Then, I develop two exhaustive mode search algorithms: one based on combined quadratic maximisation and gradient ascent and the other one based on a fixed-point iterative scheme. Appropriate values for the search control parameters are derived by taking into account theoretical results regarding the bounds for the gradient and Hessian of the mixture. The significance of the modes is quantified locally (for each mode) by error bars, or confidence intervals (estimated using the values of the Hessian at each mode); and globally by the sparseness of the mixture, measured by its differential entropy (estimated through bounds). I conclude with some reflections about bump-finding.

Keywords: Gaussian mixtures, maximisation algorithms, mode finding, bump finding, error bars, sparseness.

1 Introduction

Gaussian mixtures (Titterington et al., 1985) are ubiquitous probabilistic models for density estimation in machine learning applications. Their popularity is due to several reasons:

- Since they are a linear combination of Gaussian densities, they inherit some of the advantages of the Gaussian distribution: they are analytically tractable for many types of computations, have desirable asymptotic properties (e.g. the central limit theorem) and scale well with the data dimensionality. Furthermore, many natural data sets occur in clusters which are approximately Gaussian.
- The family of Gaussian mixtures is a universal approximator for continuous densities. In fact, Gaussian kernel density estimation (spherical Gaussian mixtures) can approximate any continuous density given enough kernels (Titterington et al., 1985; Scott, 1992). In particular, they can model multimodal distributions.
- Many complex models result in a Gaussian mixture after some assumptions are made in order to obtain tractable models. For example, Monte Carlo approximations to integrals of the type:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \approx \frac{1}{N}\sum_{n=1}^{N} p(\mathbf{x}|\boldsymbol{\theta}_n)$$

  where $\{\boldsymbol{\theta}_n\}_{n=1}^N$ are samples from the distribution $p(\boldsymbol{\theta})$ and $p(\mathbf{x}|\boldsymbol{\theta}_n)$ is assumed Gaussian. This is the case of continuous latent variable models that sample the latent variables, such as the generative topographic mapping (GTM) (Bishop et al., 1998).
- A number of convenient algorithms for estimating the parameters (means, covariance matrices and mixing proportions) exist, such as the traditional EM for maximum-likelihood (Bishop, 1995) or more recent varieties, including Bayesian (and non-Bayesian) methods that can tune the number of components needed as well as the other parameters (e.g. Roberts et al., 1998).

*Currently at the Department of Neuroscience, Georgetown University Medical Center, Washington, DC 20007, USA. Email: miguel@giccs.georgetown.edu.

Examples of models (not only for density estimation, but also for regression and classification) that often result in a Gaussian mixture include the GTM model mentioned before, kernel density estimation (also called Parzen estimation) (Scott, 1992), radial basis function networks (Bishop, 1995), mixtures of probabilistic principal component analysers (Tipping and Bishop, 1999), mixtures of factor analysers (Hinton et al., 1997), support vector machines for density estimation (Vapnik and Mukherjee, 2000), models for conditional density estimation such as mixtures of experts (Jacobs et al., 1991) and mixture density networks (Bishop, 1995), the emission distribution of hidden Markov models for automatic speech recognition and other applications (Rabiner and Juang, 1993) and, of course, the Gaussian mixture model itself. An extensive list of successful applications of Gaussian mixtures is given in Titterington et al. (1985).

Mixture models are not the only way to combine densities, though—for example, individual components may be combined multiplicatively rather than additively, as in logarithmic opinion pools (Genest and Zidek, 1986) or in the recent product of experts model (Hinton, 1999). This may be a more efficient way to model high-dimensional data which simultaneously satisfies several low-dimensional constraints: each expert is associated to a single constraint and gives high probability to regions that satisfy it and low probability elsewhere, so that the product acts as an AND operation.

Gaussian mixtures have often been praised for their ability to model multimodal distributions, where each mode represents a certain entity. For example, in visual modelling or object detection, a probabilistic model of a visual scene should account for multimodal distributions so that multiple objects can be represented (Moghaddam and Pentland, 1997; Isard and Blake, 1998). In missing data reconstruction and inverse problems, multivalued mappings can be derived from the modes of the conditional distribution of the missing variables given the present ones (Carreira-Perpiñán, 2000). Finding the modes of posterior distributions is also important in Bayesian analysis (Gelman et al., 1995, chapter 9). However, the problem of finding the modes of this important class of densities seems to have received little attention—although the problem of finding modes in a data sample, related to clustering, has been studied (see section 8.1).

Thus, the problem approached in this paper is to find all the modes of a given Gaussian mixture (of known parameters). No direct methods exist for this even in the simplest special case of one-dimensional bi-component mixtures, so iterative numerical algorithms are necessary. Intuitively, it seems reasonable that the number of modes will be smaller or equal than the number 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.

We formalise these notions in conjecture B.1, for which we provide a partial proof. This conjecture suggests that a hill-climbing algorithm starting from every centroid will not miss any mode. The analytical tractability of Gaussian mixtures allows a straightforward application of convenient optimisation algorithms and the computation of error bars. To our knowledge, this is the first time that the problem of finding all the modes of a Gaussian mixture has been investigated, although certainly the idea of using the gradient as mode locator is not new (e.g. Wilson and Spann, 1990).

The rest of the paper is organised as follows. Sections 2–3 give the equations for the moments, gradient and Hessian of the Gaussian mixture density with respect to the independent variables. Sections 4–5 describe algorithms for locating the modes. The significance of the modes thus obtained is quantified locally by computing error bars for each mode (section 6) and globally by measuring the sparseness of the mixture via the entropy (section 7). Section 8 summarises the paper, mentions some applications and compares mode finding with bump finding. The mathematical details are complemented in the appendices.

2 Moments of a mixture of (Gaussian) distributions

Consider a mixture distribution (Titterington et al., 1985) of $M > 1$ components in $\mathbb{R}^D$ for $D \geq 1$:

$$p(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{m=1}^{M} p(m)p(\mathbf{x}|m) \stackrel{\text{def}}{=} \sum_{m=1}^{M} \pi_m p(\mathbf{x}|m) \qquad \forall \mathbf{x} \in \mathbb{R}^D \tag{1}$$

where $\sum_{m=1}^{M} \pi_m = 1$, $\pi_m \in (0,1)\ \forall m = 1, \ldots, M$ and each component distribution is a normal probability distribution in $\mathbb{R}^D$. So $\mathbf{x}|m \sim \mathcal{N}_D(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, where $\boldsymbol{\mu}_m \stackrel{\text{def}}{=} \mathrm{E}_{p(\mathbf{x}|m)}\{\mathbf{x}\}$ and $\boldsymbol{\Sigma}_m \stackrel{\text{def}}{=} \mathrm{E}_{p(\mathbf{x}|m)}\left\{(\mathbf{x}-\boldsymbol{\mu}_m)(\mathbf{x}-\boldsymbol{\mu}_m)^T\right\} > 0$ are the mean vector and covariance matrix, respectively, of component m. Note that 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. Then:

- The **mixture mean** is: $\boldsymbol{\mu} \stackrel{\text{def}}{=} \mathrm{E}_{p(\mathbf{x})}\{\mathbf{x}\} = \sum_{m=1}^{M} \pi_m \boldsymbol{\mu}_m$.

- The **mixture covariance** is:

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \mathrm{E}_{p(\mathbf{x})}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\} = \mathrm{E}_{p(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^T\right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T = \sum_{m=1}^{M} \pi_m \, \mathrm{E}_{p(\mathbf{x}|m)}\left\{\mathbf{x}\mathbf{x}^T\right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T =$$
$$\sum_{m=1}^{M} \pi_m \boldsymbol{\Sigma}_m + \sum_{m=1}^{M} \pi_m \boldsymbol{\mu}_m \boldsymbol{\mu}_m^T - \boldsymbol{\mu}\boldsymbol{\mu}^T = \sum_{m=1}^{M} \pi_m \left(\boldsymbol{\Sigma}_m + (\boldsymbol{\mu}_m - \boldsymbol{\mu})(\boldsymbol{\mu}_m - \boldsymbol{\mu})^T\right). \quad (2)$$

These results are valid for any mixture, not necessarily of Gaussian distributions.

3 Gradient and Hessian with respect to the independent random variables

Here we obtain the gradient and the Hessian of the density function p with respect to the independent variables $\mathbf{x}$ (not with respect to the parameters $\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$). Firstly let us derive the gradient and Hessian for a D-variate normal distribution. Let $\mathbf{x} \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then:

$$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

which is differentiable and nonnegative for all $\mathbf{x} \in \mathbb{R}^D$. Let $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ be the singular value decomposition of $\boldsymbol{\Sigma}$, so that $\mathbf{U}$ is orthogonal and $\boldsymbol{\Lambda} > 0$ diagonal. Calling $\mathbf{z} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$:

$$\frac{\partial p}{\partial x_d} = p(\mathbf{x}) \frac{\partial}{\partial x_d}\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) = p(\mathbf{x}) \frac{\partial}{\partial x_d}\left(-\frac{1}{2}(\mathbf{z}^T\boldsymbol{\Lambda}^{-1}\mathbf{z})\right)$$
$$= p(\mathbf{x}) \sum_{e=1}^{D} \frac{\partial}{\partial z_e}\left(-\frac{1}{2}\sum_{f=1}^{D}\frac{z_f^2}{\lambda_f}\right)\frac{\partial z_e}{\partial x_d} = p(\mathbf{x}) \sum_{e=1}^{D} -\frac{z_e}{\lambda_e} u_{de},$$

since $\partial z_e / \partial x_d = u_{de}$. The result above is the dth element of vector $-p(\mathbf{x})\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{z}$ and so the gradient is[1]:

$$\mathbf{g} \stackrel{\text{def}}{=} \nabla p(\mathbf{x}) = p(\mathbf{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x}). \quad (3)$$

Taking the second derivatives:

$$\frac{\partial}{\partial x_c}\left(\frac{\partial p}{\partial x_d}\right) = \frac{\partial p}{\partial x_c}\sum_{e=1}^{D} -\frac{z_e}{\lambda_e}u_{de} + p(\mathbf{x})\sum_{e=1}^{D} -\frac{u_{de}}{\lambda_e}\frac{\partial z_e}{\partial x_c}$$
$$= p(\mathbf{x})\left(\sum_{e=1}^{D} -\frac{z_e}{\lambda_e}u_{ce}\right)\left(\sum_{e=1}^{D} -\frac{z_e}{\lambda_e}u_{de}\right) + p(\mathbf{x})\sum_{e=1}^{D} -\frac{u_{de}}{\lambda_e}u_{ce},$$

which is the (c,d)th element of matrix $\frac{\mathbf{g}\mathbf{g}^T}{p(\mathbf{x})} - p(\mathbf{x})\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T$. So the Hessian is:

$$\mathbf{H} \stackrel{\text{def}}{=} (\nabla\nabla^T)p(\mathbf{x}) = \frac{\mathbf{g}\mathbf{g}^T}{p(\mathbf{x})} - p(\mathbf{x})\boldsymbol{\Sigma}^{-1} = p(\mathbf{x})\boldsymbol{\Sigma}^{-1}\left((\boldsymbol{\mu}-\mathbf{x})(\boldsymbol{\mu}-\mathbf{x})^T - \boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}. \quad (4)$$

It is clear that the gradient is zero at $\mathbf{x} = \boldsymbol{\mu}$ only and there $\mathbf{H} = -|2\pi\boldsymbol{\Sigma}|^{-1/2}\boldsymbol{\Sigma}^{-1} < 0$, thus being a maximum.

Consider now a finite mixture of D-variate normal distributions $p(\mathbf{x}) = \sum_{m=1}^{M} p(m)p(\mathbf{x}|m)$ where $\mathbf{x}|m \sim \mathcal{N}_D(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. By the linearity of the differential operator and defining:

$$\mathbf{g}_m \stackrel{\text{def}}{=} \nabla p(\mathbf{x}|m) = p(\mathbf{x}|m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x})$$
$$\mathbf{H}_m \stackrel{\text{def}}{=} (\nabla\nabla^T)p(\mathbf{x}|m) = \frac{\mathbf{g}_m\mathbf{g}_m^T}{p(\mathbf{x}|m)} - p(\mathbf{x}|m)\boldsymbol{\Sigma}_m^{-1} = p(\mathbf{x}|m)\boldsymbol{\Sigma}_m^{-1}\left((\boldsymbol{\mu}_m-\mathbf{x})(\boldsymbol{\mu}_m-\mathbf{x})^T - \boldsymbol{\Sigma}_m\right)\boldsymbol{\Sigma}_m^{-1}$$

we obtain:

$$\text{Gradient } \mathbf{g} \stackrel{\text{def}}{=} \nabla p(\mathbf{x}) = \sum_{m=1}^{M} p(m)\mathbf{g}_m = \sum_{m=1}^{M} p(\mathbf{x},m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x}) \quad (5)$$

$$\text{Hessian } \mathbf{H} \stackrel{\text{def}}{=} (\nabla\nabla^T)p(\mathbf{x}) = \sum_{m=1}^{M} p(m)\mathbf{H}_m = \sum_{m=1}^{M} p(\mathbf{x},m)\boldsymbol{\Sigma}_m^{-1}\left((\boldsymbol{\mu}_m-\mathbf{x})(\boldsymbol{\mu}_m-\mathbf{x})^T - \boldsymbol{\Sigma}_m\right)\boldsymbol{\Sigma}_m^{-1}. \quad (6)$$

[1] For clarity of notation, we omit the dependence on $\mathbf{x}$ of both the gradient and the Hessian, writing $\mathbf{g}$ and $\mathbf{H}$ where we should write $\mathbf{g}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x})$.

3.1 Gradient and Hessian of the log-density

Call $L(\mathbf{x}) \stackrel{\text{def}}{=} \ln p(\mathbf{x})$. Then, the gradient and Hessian of L are related to those of p as follows:

$$\text{Gradient } \nabla L(\mathbf{x}) = \frac{1}{p}\mathbf{g} \tag{7}$$

$$\text{Hessian } (\nabla\nabla^T)L(\mathbf{x}) = -\frac{1}{p^2}\mathbf{g}\mathbf{g}^T + \frac{1}{p}\mathbf{H}. \tag{8}$$

Note from eq. (8) that, if the Hessian $\mathbf{H}$ of p is definite negative, then the Hessian of L is also definite negative, since $-\frac{1}{p^2}\mathbf{g}\mathbf{g}^T$ is either a null matrix (at stationary points) or negative definite (everywhere else).

In this paper, we will always implicitly refer to the gradient $\mathbf{g}$ or Hessian $\mathbf{H}$ of p, eqs. (5) and (6), rather than those of L, eqs. (7) and (8), unless otherwise noted.

4 Exhaustive mode search by a gradient-quadratic search

Consider a Gaussian mixture p with $M > 1$ components as in equation (1). Since the family of Gaussian mixtures is a density universal approximator, the landscape of p could be very complex. However, assuming that conjecture B.1 in appendix B is true, there are at most M modes and it is clear that every centroid $\boldsymbol{\mu}_m$ of the mixture must be near, if not coincident, with one of the modes, since the modes are contained in the convex hull of the centroids (as the mean is). Thus, an obvious procedure to locate all the modes is to use a hill-climbing algorithm starting from every one of the centroids, i.e., starting from every vertex of the convex hull.

Due to the ease of calculation of the gradient (5) and the Hessian (6), it is straightforward to use quadratic maximisation (i.e., Newton's method) combined with gradient ascent (Press et al., 1992). Assuming we are at a point $\mathbf{x}_0$, let us expand $p(\mathbf{x})$ around $\mathbf{x}_0$ as a Taylor series to second order:

$$p(\mathbf{x}) \approx p(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T\mathbf{g}(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)$$

where $\mathbf{g}(\mathbf{x}_0)$ and $\mathbf{H}(\mathbf{x}_0)$ are the gradient and Hessian of p at $\mathbf{x}_0$, respectively. The zero-gradient point of the previous quadratic form is given by:

$$\nabla p(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0) + \mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{x}_0 - \mathbf{H}^{-1}(\mathbf{x}_0)\mathbf{g}(\mathbf{x}_0) \tag{9}$$

which jumps from $\mathbf{x}_0$ to the maximum (or minimum, or saddle point) of the quadratic form in a single leap. Thus, for maximisation, the Hessian can only be used if it is negative definite, i.e., if all its eigenvalues are negative.

If the Hessian is not negative definite, which means that we are not yet in a hill-cap (defined as the region around a mode where $\mathbf{H} < 0$), we use gradient ascent:

$$\mathbf{x} = \mathbf{x}_0 + s\mathbf{g}(\mathbf{x}_0) \tag{10}$$

where $s > 0$ is the step size. That is, we jump a distance $s\,\|\mathbf{g}(\mathbf{x}_0)\|$ in the direction of the gradient (which does not necessarily point towards the maximum). For comments about the choice of the step size, see section 4.3.

Once found a point for which $\mathbf{g} = \mathbf{0}$, the Hessian (6) can confirm that the point is indeed a maximum by checking that $\mathbf{H} < 0$. Of course, both the nullity of the gradient and the negativity of the Hessian can only be ascertained to a certain numerical accuracy, but due to the simplicity of the surface of $p(\mathbf{x})$ this should not be a problem (at least for a small dimensionality D). Section 4.3 discusses the control parameters for the gradient ascent. Fig. 1 shows the pseudocode for the algorithm. Fig. 2 illustrates the case with a two-dimensional example.

Some remarks:

- It is not convenient here to use multidimensional optimisation methods based on line searches, such as the conjugate gradient method, because the line search may discard local maxima. Since we are interested in finding all the modes, we need a method that does not abandon the region of influence of a maximum. Gradient ascent with a small step followed by quadratic optimisation in a hill-cap should not miss local maxima.

- If the starting point is at or close to a stationary point, i.e., with near-zero gradient, the method will not iterate. Examination of the Hessian will determine if the point is a maximum, a minimum or a saddle point. In the latter two cases it will be discarded.

- The gradient ascent should not suffer too much in higher dimensions because the search follows a one-dimensional path. Of course, this path can twist itself in many more dimensions and thus became longer, but once it reaches a hill-cap, quadratic maximisation converges quickly. If the dimension of the space is D, computing the gradient and the Hessian is $\mathcal{O}(D)$ and $\mathcal{O}(D^2)$, respectively. Inverting the Hessian is $\mathcal{O}(D^3)$, but this may be reduced by the techniques of appendix C.
- Other optimisation strategies based on the gradient or the Hessian, such as the Levenberg-Marquardt algorithm (Press et al., 1992), can also be easily constructed.

4.1 Maximising the density p vs. maximising the log-density $L = \ln p$

Experimental results show that, when the component centroids $\boldsymbol{\mu}_m$ are used as starting points, there is not much difference in speed of convergence between using the gradient and Hessian of $p(\mathbf{x})$ and using those of $L(\mathbf{x}) = \ln p(\mathbf{x})$; although there is difference from other starting points, e.g. far from the convex hull, where $p(\mathbf{x})$ is very small. Fig. 2 illustrates this: in the top row, observe the slow search in points lying in areas of near-zero probability in the case of $p(\mathbf{x})$ and the switch from gradient to quadratic search when the point is in a hill-cap, where the Hessian is negative definite. In the middle row, observe how much bigger the areas with negative definite Hessian are for the surface of $L(\mathbf{x})$ in regions where $p(\mathbf{x})$ is small, as noted in section 3.1. This means that, for starting points in regions where $p(\mathbf{x})$ is small, quadratic steps can be taken more often and thus convergence is faster. However, the centroids are usually in areas of high $p(\mathbf{x})$ and thus there is no improvement for our mode-finding algorithm.

In any case, at each step one can compute the gradient and Hessian for both $p(\mathbf{x})$ and $\ln p(\mathbf{x})$ and choose the one for which the new point has the highest probability.

It may be argued that L is a quadratic form if p is Gaussian, in which case a quadratic optimiser would find the maximum in a single step. However, p will be far from Gaussian even near the centroids or modes if the mixture components interact strongly (i.e., if they are close enough with respect to their covariance matrices, as in fig. 2) and this will be the case when the mixture is acting as a density approximator (as in kernel estimation).

4.2 Low-probability components

Gaussian mixtures are often applied to high-dimensional data. Due to computational difficulties and to the usual lack of sufficient training data (both issues arising from the curse of the dimensionality; Scott, 1992), the estimated mixture may not be a good approximation to the density of the data. If this is the case, some of the modes found may be spurious, due to artifacts of the model. A convenient way to filter them out is to reject all modes whose probability (normalised by the probability of the highest mode) is smaller than a certain small threshold $\theta > 0$ (e.g. $\theta = 0.01$).

A similar situation arises when the mixture whose modes are to be found is the result of computing the conditional distribution of a joint Gaussian mixture given the values of certain variables. For example, if

$$p(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} p(m) p(\mathbf{x}, \mathbf{y}|m)$$

with $(\mathbf{x}, \mathbf{y}|m) \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ then it is easy to see that the conditional distribution $p(\mathbf{x}|\mathbf{y})$ is again a Gaussian mixture:

$$p(\mathbf{x}|\mathbf{y}) = \sum_{m=1}^{M} p(m|\mathbf{y}) p(\mathbf{x}|\mathbf{y}, m)$$

where:

- The components are normally distributed, $(\mathbf{x}|\mathbf{y}, m) \sim \mathcal{N}(\boldsymbol{\mu}'_m, \boldsymbol{\Sigma}'_m)$, with $\boldsymbol{\mu}'_m$ and $\boldsymbol{\Sigma}'_m$ dependent on $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$ and the value of $\mathbf{y}$.
- The new mixing proportions are

$$p(m|\mathbf{y}) = \frac{p(\mathbf{y}|m)p(m)}{p(\mathbf{y})} \propto p(m) e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_{m,y})^T \boldsymbol{\Sigma}_{m,yy}^{-1} (\mathbf{y}-\boldsymbol{\mu}_{m,y})}$$

  where $\boldsymbol{\mu}_{m,y}$ and $\boldsymbol{\Sigma}_{m,yy}$ are obtained by crossing out the columns and rows of variables $\mathbf{x}$ in $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, respectively. Thus, the means (projected in the $\mathbf{y}$ axes) of most components will be far from the value of $\mathbf{y}$ and will have a negligible mixing proportion $p(m|\mathbf{y})$. Filtering out such low-probability components will accelerate considerably the mode search without missing any important mode.

inputs

Gaussian mixture defined by $\{p(m), \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$

constants

$\sigma \leftarrow \sqrt{\min_{m=1,\ldots,M}\{\min\{\text{eigenvalues}(\boldsymbol{\Sigma}_m)\}\}}$ — Minimal standard deviation

$\epsilon \leftarrow 10^{-4}$ — Small number $0 < \epsilon \ll 1$

$\theta \leftarrow 10^{-2}$ — Rejection threshold $0 < \theta \ll 1$

control parameters

$\texttt{min_step} \leftarrow \sigma^2(2\pi\sigma^2)^{\frac{D}{2}}$

$\texttt{min_grad} \leftarrow \sigma^{-1}(2\pi\sigma^2)^{-\frac{D}{2}}\epsilon e^{-\frac{1}{2}\epsilon^2}$

$\texttt{min_diff} \leftarrow 100\epsilon\sigma$

$\texttt{max_eig} \leftarrow 0$

$\texttt{max_it} \leftarrow 1000$

initialise

$s \leftarrow 64{*}\texttt{min_step}$ — Step size

$\mathcal{M} \leftarrow \varnothing$ — Mode set

optionally

Remove all components for which $\frac{p(m)}{\max_{m=1,\ldots,M} p(m)} < \theta$ and renormalise $p(m)$

for $m = 1, \ldots, M$ — For each centroid

$i \leftarrow 0$ — Iteration counter

$\mathbf{x} \leftarrow \boldsymbol{\mu}_m$ — Starting point

$p \leftarrow p(\mathbf{x})$ — From eq. (1)

repeat — Gradient-quadratic search loop

$\mathbf{g} \leftarrow \nabla \ln p(\mathbf{x})$ — Gradient from eq. (7)

$\mathbf{H} \leftarrow (\nabla\nabla^T)\ln p(\mathbf{x})$ — Hessian from eq. (8)

$\mathbf{x}_{\text{old}} \leftarrow \mathbf{x}$

$p_{\text{old}} \leftarrow p$

if $\mathbf{H} < 0$ — Hessian negative definite?

$\mathbf{x} \leftarrow \mathbf{x}_{\text{old}} - \mathbf{H}^{-1}\mathbf{g}$ — Quadratic step

$p \leftarrow p(\mathbf{x})$ — From eq. (1)

end

if $\mathbf{H} \not< 0$ **or** $p \leq p_{\text{old}}$

$\mathbf{x} \leftarrow \mathbf{x}_{\text{old}} + s\mathbf{g}$ — Gradient step

$p \leftarrow p(\mathbf{x})$ — From eq. (1)

while $p < p_{\text{old}}$

$s \leftarrow s/2$

$\mathbf{x} \leftarrow \mathbf{x}_{\text{old}} + s\mathbf{g}$ — Gradient step

$p \leftarrow p(\mathbf{x})$ — From eq. (1)

end

end

$i \leftarrow i + 1$

until $i \geq \texttt{max_it}$ **or** $\|\mathbf{g}\| < \texttt{min_grad}$

if $\max\{\text{eigenvalues}(\mathbf{H})\} < \texttt{max_eig}$ — Update mode set

$\mathcal{N} \leftarrow \{\boldsymbol{\nu} \in \mathcal{M} : \|\boldsymbol{\nu} - \mathbf{x}\| \leq \texttt{min_diff}\} \cup \{\mathbf{x}\}$

$\mathcal{M} \leftarrow (\mathcal{M} \setminus \mathcal{N}) \cup \{\arg\max_{\boldsymbol{\nu}\in\mathcal{N}} p(\boldsymbol{\nu})\}$

end

end

return $\mathcal{M}$

Figure 1: Pseudocode of the gradient-quadratic mode-finding algorithm described in section 4. Instead of for $L(\mathbf{x}) = \ln p(\mathbf{x})$, the gradient and the Hessian can be computed for $p(\mathbf{x})$ using eqs. (5) and (6): $\mathbf{g} \leftarrow \nabla p(\mathbf{x})$, $\mathbf{H} \leftarrow (\nabla\nabla^T)p(\mathbf{x})$. Also, at each step one can compute the gradient and Hessian for both $p(\mathbf{x})$ and $\ln \mathbf{p}(\mathbf{x})$ and choose the one for which the new point has the highest probability.

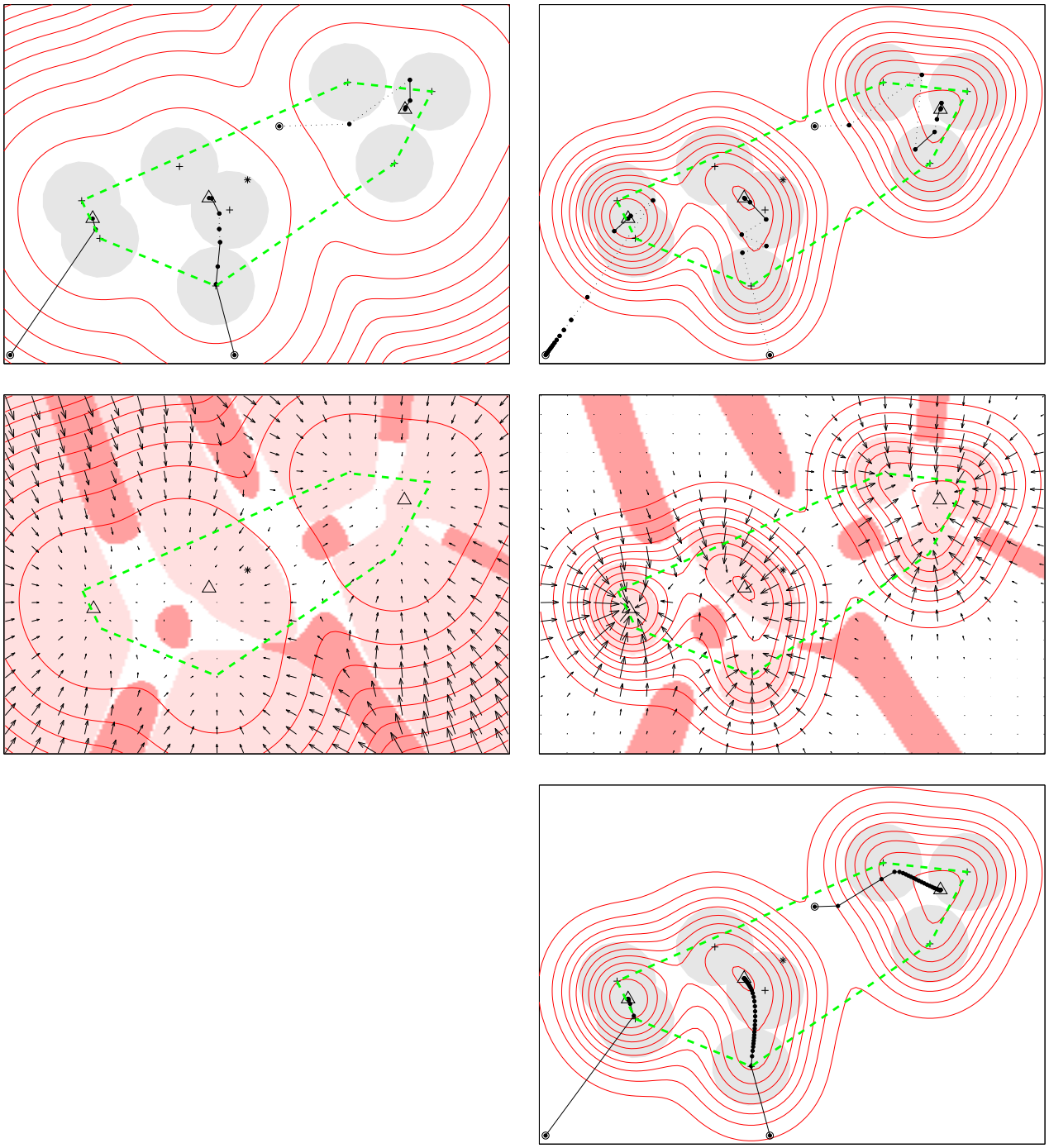

Figure 2: Example of mode searching in a two-dimensional Gaussian mixture. In this example, the mixing proportions are equal and the components are isotropic with equal covariance $\sigma^2\mathbf{I}_2$. The surface has 3 modes and various other features (saddle points, ridges, plateaux, etc.). The mixture modes are marked "△" and the mixture mean "∗". The dashed, thick-line polygon is the convex hull of the centroids. The left column shows the surface of $L(\mathbf{x}) = \ln p(\mathbf{x})$ and the right column the surface of $p(\mathbf{x})$. *Top row*: contour plot of the objective function. Each original component is indicated by a grey disk of radius σ centred on the corresponding mean vector $\boldsymbol{\mu}_m$ (marked "+"). A few search paths from different starting points (marked "∘") are given for illustrative purposes (paths from the centroids are much shorter); continuous lines indicate gradient steps and dotted lines quadratic steps. *Middle row*: plot of the gradient (arrows) and the Hessian character (dark colour: positive definite; white: indefinite; light colour: negative definite). *Bottom row*: like the top row, but here the fixed-point iterative algorithm was used.

4.3 Control parameters for the gradient-quadratic mode-finding algorithm

We consider here a single D-dimensional isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_D)$ and derive some control parameters for the gradient-quadratic algorithm which may be transferrable to the mixture case: `min_step`, `min_grad`, `min_diff` and `max_eig`.

- For the isotropic Gaussian, the gradient always points to the mode, by symmetry. Thus, $\mathbf{g} = \|\mathbf{g}\| \frac{\boldsymbol{\mu}-\mathbf{x}}{\|\boldsymbol{\mu}-\mathbf{x}\|}$. Consider a point at a normalised distance $\rho = \left\| \frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma} \right\|$ from the mode $\boldsymbol{\mu}$. Then $\mathbf{g} = \frac{\|\mathbf{g}\|}{\rho\sigma}(\boldsymbol{\mu} - \mathbf{x})$. Thus, a step $s = \frac{\rho\sigma}{\|\mathbf{g}\|}$ would jump directly to the mode: $\mathbf{x}_{\text{new}} = \mathbf{x} + s\mathbf{g} = \mathbf{x} + \frac{\rho\sigma}{\|\mathbf{g}\|}\frac{\|\mathbf{g}\|}{\rho\sigma}(\boldsymbol{\mu} - \mathbf{x}) = \boldsymbol{\mu}$. From eq. (3) $\|\mathbf{g}\| = p(\mathbf{x}) \left\| \frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma^2} \right\| = \sigma^{-1}(2\pi\sigma^2)^{-\frac{D}{2}} e^{-\frac{1}{2}\rho^2} \rho \leq \sigma^{-1}(2\pi\sigma^2)^{-\frac{D}{2}}\rho$ and so $s \geq \sigma^2(2\pi\sigma^2)^{\frac{D}{2}}$. Therefore, a step of $\texttt{min_step} = \sigma^2(2\pi\sigma^2)^{\frac{D}{2}}$ times the gradient would never overshoot, that is, would climb up the hill monotonically. However, it would be too small for points a few normalised distances away from the mode. Moreover, theorem D.3 shows that the gradient norm for the mixture is never larger than that of any isolated component, which suggests using even larger step sizes. Thus, our gradient ascent algorithm starts with a step size of several (64, corresponding to a point at 2.88 normalised distances away from the mode) times the previous step size and halves it every time the new point has a worse probability.

- To determine when a gradient norm is considered numerically zero, we choose the points for which the normalised distance is less than a small value ϵ (set to 10^{-4}). This gives a minimum gradient norm $\texttt{min_grad} = \sigma^{-1}(2\pi\sigma^2)^{-\frac{D}{2}}\epsilon e^{-\frac{1}{2}\epsilon^2} \approx \sigma^{-1}(2\pi\sigma^2)^{-\frac{D}{2}}\epsilon$. Since for the mixtures the gradient will be smaller than for the components, points with a gradient norm of `min_grad` may be farther from the mode than they would be in the case of a single mixture, so ϵ should be small compared to 1.

- Thus, since the algorithm will not get closer to the mode than a normalised distance of ϵ, the normalised distance below which two points are considered the same may be taken as several times larger than ϵ. We take $\texttt{min_diff} = 100\epsilon\sigma$ as the minimum absolute difference between two modes to be assimilated as one.

  Updating the mode set $\mathcal{M}$ in the algorithm after a new mode $\mathbf{x}$ has been found is achieved by identifying all modes previously found that are closer to $\mathbf{x}$ than a distance `min_diff`, including $\mathbf{x}$ (the $\mathcal{N}$ set in figures 1 and 3), removing them from the mode set $\mathcal{M}$ and adding to $\mathcal{M}$ the mode in $\mathcal{N}$ with highest probability p.

- The Hessian will be considered negative definite if its algebraically largest eigenvalue is less than a nonnegative parameter `max_eig`. We take $\texttt{max_eig} = 0$, since theorem D.4 shows that not too far from a mode (for the case of one component, inside a radius of one normalised distance), the Hessian will already be negative definite. This strict value of `max_eig` will rule out all minima (for which $\|\mathbf{g}\| = 0$ but $\mathbf{H} > 0$) and should not miss any maximum.

- To limit the computation time, we define `max_it` as the maximum number of iterations to be performed.

These results can be easily generalised to a mixture of full-covariance Gaussians. Theorems D.1 and D.3 show that the mixture gradient anywhere is bounded above and depends on the smallest eigenvalue of the covariance matrix of any of the components. Calling this minimal eigenvalue σ^2, the control parameter definitions given above remain the same.

In high dimensions, these parameters may require manual tuning (perhaps using knowledge of the particular problem being tackled), specially if too many modes are obtained—due to the nature of the geometry of high-dimensional spaces (Scott, 1992). For example, both `min_step` and `min_grad` depend exponentially on D, which can lead to very large or very small values depending on the value of σ. For `min_diff`, consider the following situation: vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ differ in a small value δ in each component, so that $\|\mathbf{x}_1 - \mathbf{x}_2\| = \sqrt{D}\delta$. For high D, $\|\mathbf{x}_1 - \mathbf{x}_2\|$ will be large even though one would probably consider $\mathbf{x}_1$ and $\mathbf{x}_2$ as the same vector. However, if that difference was concentrated in a single component, one would probably consider them as very different vectors.

Finally, we remark that there is a lower bound in the precision achievable by any numerical algorithm due to the finite-precision arithmetic (Press et al., 1992, pp. 398–399), so that in general we cannot get arbitrarily close to a scalar value μ: at best, our estimate x we will get to $|x - \mu| \sim \mu\sqrt{\epsilon_m}$, where ϵ_m is the machine accuracy (usually $\epsilon_m \approx 3 \times 10^{-8}$ for simple precision and $\epsilon_m \approx 10^{-15}$ for double precision). This gives a limit in how small to make all the control parameters mentioned. Furthermore, converging to many decimals is a waste, since the mode is at best only a (nonrobust) statistical estimate based on our model—whose parameters were also estimated to some precision.

5 Exhaustive mode search by a fixed-point search

Equating the gradient expression (5) to zero we obtain immediately a fixed-point iterative scheme:

$$\mathbf{g} = \sum_{m=1}^{M} p(\mathbf{x}, m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x}) = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{f}(\mathbf{x}) \text{ with } \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} \left(\sum_{m=1}^{M} p(m|\mathbf{x})\boldsymbol{\Sigma}_m^{-1}\right)^{-1} \sum_{m=1}^{M} p(m|\mathbf{x})\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m \quad (11)$$

where we have used Bayes' theorem. Note that using the gradient of $L(\mathbf{x}) = \ln p(\mathbf{x})$ makes no difference here. A fixed point $\mathbf{x}$ of the mapping $\mathbf{f}$ verifies by definition $\mathbf{x} = \mathbf{f}(\mathbf{x})$. The fixed points of $\mathbf{f}$ are thus the stationary points of the mixture density p, including maxima, minima and saddle points. An iterative scheme $\mathbf{x}^{(n+1)} = \mathbf{f}(\mathbf{x}^{(n)})$ will converge to a fixed point of $\mathbf{f}$ under certain conditions, e.g. if $\mathbf{f}$ is a contractive mapping in an environment of the fixed point (Isaacson and Keller, 1966). Unfortunately, the potential existence of several fixed points in the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^M$ and the complexity of eq. (11) make a convergence analysis of the method difficult. However, in a number of experiments it has found exactly the same modes as the gradient-quadratic method. Thus, as in section 4, iterating from each centroid should find all maxima, since at least some of the centroids are likely to be near the modes. Checking the eigenvalues of the Hessian of p with eq. (6) will determine whether the point found is actually a maximum.

The fixed-point iterative algorithm is much simpler than the gradient-quadratic one, but it also requires many more iterations to converge inside the convex hull of $\{\boldsymbol{\mu}_m\}_{m=1}^M$, as can be seen experimentally (observe in fig. 2 (bottom) the quick jump to the area of high probability and the slow convergence thereafter). The inverse matrix $\left(\sum_{m=1}^{M} p(m|\mathbf{x})\boldsymbol{\Sigma}_m^{-1}\right)^{-1}$ may be trivially computed in some cases (e.g. if all the components are diagonal). In the particular[2] case where $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ for all $m = 1, \ldots, M$, the fixed-point scheme reduces to the extremely simple form

$$\mathbf{x}^{(n+1)} = \sum_{m=1}^{M} p(m|\mathbf{x}^{(n)})\boldsymbol{\mu}_m,$$

i.e., the new point $\mathbf{x}^{(n+1)}$ is the conditional mean of the mixture under the current point $\mathbf{x}^{(n)}$. This is formally akin to EM algorithms for parameter estimation of mixture distributions (Dempster et al., 1977), to clustering by deterministic annealing (Rose, 1998) and to algorithms for finding pre-images in kernel-based methods (Schölkopf et al., 1999).

Whether this algorithm is faster than the gradient-quadratic one has to be determined for each particular case, depending on the values of D and M and the numerical routines used for matrix inversion.

5.1 Control parameters for the fixed-point mode-finding algorithm

A theoretical advantage of the fixed-point scheme over gradient ascent is that no step size is needed. As in section 4.3, call σ^2 the smallest eigenvalue of the covariance matrix of any of the components. A new tolerance parameter $\texttt{tol} = \epsilon\sigma$ is defined for some small ϵ (10^{-4}), so that if the distance between two successive points is smaller than `tol`, we stop iterating. Alternatively, we could use the `min_grad` control parameter of section 4.3. The following control parameters from section 4.3 remain unchanged: $\texttt{min_diff} = 100\epsilon\sigma$, `max_eig` and `max_it`. Note that `min_diff` should be several times larger than `tol`.

6 Error bars for the modes

In this section we deal with the problem of deriving error bars, or confidence intervals, for a mode of a mixture of Gaussian distributions. That is, the shape of the distribution around that mode (how peaked or how spread out) contains information about the certainty of the value of the mode. These error bars are not related in any way to the numerical precision with which that mode was found by the iterative algorithm; they are related to the statistical dispersion around it.

The confidence interval, or in higher dimensions, the confidence hyperrectangle, means here a hyperrectangle containing the mode and with a probability under the mixture distribution of value P fixed in advance. For example, for $P = 0.9$ in one dimension we speak of a 90% confidence interval. Since computing error bars for the mixture distribution is analytically difficult, we follow an approximate computational approach: we replace the mixture distribution around the mode by a normal distribution centred in that mode and with a certain covariance matrix. Then we compute symmetric error bars for this normal distribution, which is

[2] Important models fall in this case, such as Gaussian kernel density estimation (Parzen estimators) (Scott, 1992) or the generative topographic mapping (GTM) (Bishop et al., 1998), as well as Gaussian radial basis function networks.

inputs

Gaussian mixture defined by $\{p(m), \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$

constants

$\sigma \leftarrow \sqrt{\min_{m=1,\ldots,M}\{\min\{\text{eigenvalues}(\boldsymbol{\Sigma}_m)\}\}}$ — Minimal standard deviation

$\epsilon \leftarrow 10^{-4}$ — Small number $0 < \epsilon \ll 1$

$\theta \leftarrow 10^{-2}$ — Rejection threshold $0 < \theta \ll 1$

control parameters

$\texttt{tol} \leftarrow \epsilon\sigma$

$\texttt{min_diff} \leftarrow 100\epsilon\sigma$

$\texttt{max_eig} \leftarrow 0$

$\texttt{max_it} \leftarrow 1000$

initialise

$\mathcal{M} \leftarrow \varnothing$ — Mode set

optionally

Remove all components for which $\frac{p(m)}{\max_{m=1,\ldots,M} p(m)} < \theta$ and renormalise $p(m)$

for $m = 1, \ldots, M$ — For each centroid

$i \leftarrow 0$ — Iteration counter

$\mathbf{x} \leftarrow \boldsymbol{\mu}_m$ — Starting point

repeat — Fixed-point iteration loop

$\mathbf{x}_{\text{old}} \leftarrow \mathbf{x}$

$\mathbf{x} \leftarrow \left(\sum_{m=1}^M p(m|\mathbf{x})\boldsymbol{\Sigma}_m^{-1}\right)^{-1} \sum_{m=1}^M p(m|\mathbf{x})\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m$ — From eq. (11)

$i \leftarrow i + 1$

until $i \geq \texttt{max_it}$ **or** $\|\mathbf{x} - \mathbf{x}_{\text{old}}\| < \texttt{tol}$

$\mathbf{H} \leftarrow (\nabla\nabla^T)p(\mathbf{x})$ — Hessian from eq. (6)

if $\max\{\text{eigenvalues}(\mathbf{H})\} < \texttt{max_eig}$ — Update mode set

$\mathcal{N} \leftarrow \{\boldsymbol{\nu} \in \mathcal{M} : \|\boldsymbol{\nu} - \mathbf{x}\| \leq \texttt{min_diff}\} \cup \{\mathbf{x}\}$

$\mathcal{M} \leftarrow (\mathcal{M} \setminus \mathcal{N}) \cup \{\arg\max_{\boldsymbol{\nu}\in\mathcal{N}} p(\boldsymbol{\nu})\}$

end

end

return $\mathcal{M}$

Figure 3: Pseudocode of the fixed-point mode-finding algorithm described in section 5.

easy, as we show in section 6.1. Section 6.2 deals with the problem of selecting the covariance matrix of the approximating normal.

Ideally we would like to have asymmetric bars, accounting for possible skewness of the distribution around the mode, but this is difficult in several dimensions. Also, note that in high dimensions the error bars become very wide, since due to the curse of the dimensionality the probability contained in a fixed hypercube decreases exponentially with the dimension (Scott, 1992).

6.1 Confidence intervals at the mode of a normal distribution

Consider a D-variate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ is the singular value decomposition[3] of its covariance matrix $\boldsymbol{\Sigma}$. Given $\rho > 0$, $\mathcal{R} = \|\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})\|_\infty \leq \rho$ represents a hyperrectangle[4] with its centre on $\mathbf{x} = \boldsymbol{\mu}$. Its sides are aligned with the principal axes of $\boldsymbol{\Sigma}$, that is, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D)$, and have lengths $2\rho\sqrt{\lambda_1}, \ldots, 2\rho\sqrt{\lambda_D}$ (see fig. 4a). The probability $P(\rho)$ contained in this hyperrectangle $\mathcal{R}$ can be computed as

[3] In general, the singular value decomposition of a rectangular matrix $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ with $\mathbf{U}$, $\mathbf{V}$ orthogonal and $\mathbf{S}$ diagonal, but for a symmetric square matrix $\mathbf{U} = \mathbf{V}$.

[4] Considering a hyperellipse $\mathcal{E} = \|\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})\|_2 \leq \rho$ instead of a hyperrectangle simplifies the analysis, but we are interested in separate intervals along each direction.

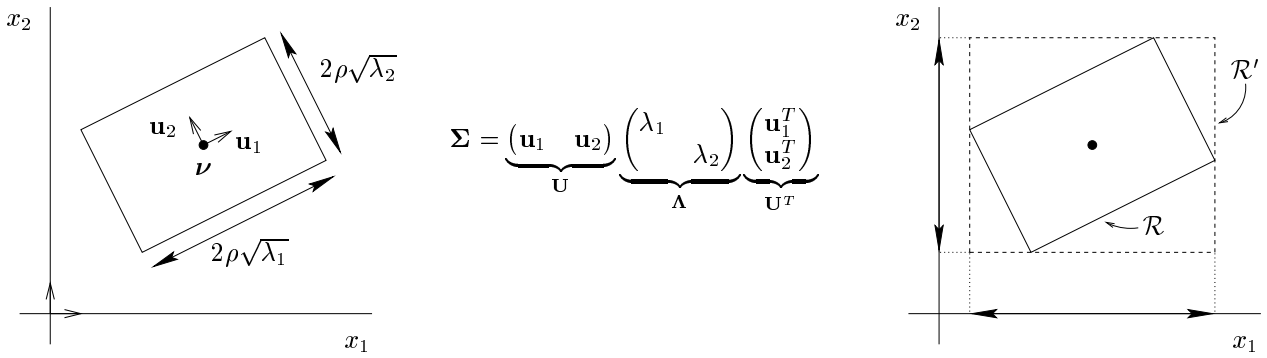

Figure 4: Left: schematic of the error bars, or hyperrectangle $\mathcal{R}$, in two dimensions. Right: how to obtain error bars in the original axes by circumscribing another hyperrectangle $\mathcal{R}'$ to $\mathcal{R}$. It is clear that $P(\mathcal{R}') \geq P(\mathcal{R})$.

$P^{1/D}$	ρ
1	∞
0.9973	3
0.9545	2
0.6827	1
0.5	0.6745

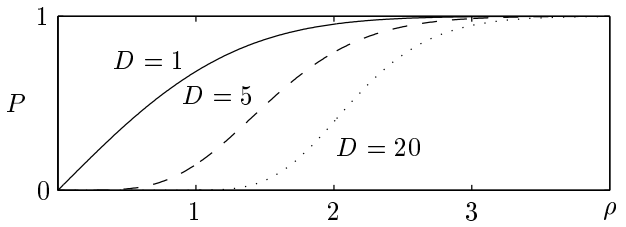

Figure 5: Probability P of a D-dimensional hypercube of side 2ρ centred in the mode of a D-dimensional normal distribution, from eq. (12). Note that to obtain P from the table one has to raise the numbers on the left column to power D.

follows:

$$P(\rho) = \int_{\mathcal{R}} |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\, d\mathbf{x} = \int_{\|\mathbf{z}\|_\infty \leq \rho} |2\pi\boldsymbol{\Lambda}|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} |\boldsymbol{\Lambda}|^{\frac{1}{2}}\, d\mathbf{z} = \prod_{d=1}^{D} P_{\mathcal{N}(0,1)}\{z_d \in [-\rho, \rho]\} = \left(\operatorname{erf}\left(\frac{\rho}{\sqrt{2}}\right)\right)^D \quad (12)$$

where we have changed $\mathbf{z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$, with Jacobian $|\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T| = |\boldsymbol{\Lambda}|^{-1/2}$, and the error function erf is defined as

$$\operatorname{erf}(x) \stackrel{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt = P_{\mathcal{N}(0,1)}\left\{[-\sqrt{2}x, \sqrt{2}x]\right\}.$$

From (12), $\rho = \sqrt{2}\operatorname{arg\,erf}(P^{1/D})$, so that for a desired confidence level given by the probability P we can obtain the appropriate interval (see table in figure 5). The natural confidence intervals, or error bars, that $\mathcal{R}$ gives us are of the form $\left|\sum_{c=1}^{D} u_{cd}(x_c - \mu_c)\right| \leq \rho\sqrt{\lambda_d}$. They follow the directions of the principal axes of $\boldsymbol{\Sigma}$, which in general will not coincide with the original axes of the $x_1, \ldots, x_D$ variables. Of course, we can obtain a new rectangle $\mathcal{R}'$ aligned with the $x_1, \ldots, x_D$ axes by taking intervals ranging from the minimal to the maximal corner of $\mathcal{R}$ in each direction, but obviously $P(\mathcal{R}') \geq P(\mathcal{R})$ and it is difficult to find $P(\mathcal{R}')$ exactly or to bound the error (see fig. 4b).

Note that for the mixture with $\boldsymbol{\Sigma}_m = \sigma^2\mathbf{I}_D$, we have $\boldsymbol{\Lambda} - \sigma^2\mathbf{I}_D \geq 0$ always and so the error bars have a minimal length of $2\rho\sigma$ in each principal direction.

6.2 Approximation by a normal distribution near a mode of the mixture

If the mixture distribution is unimodal, then the best estimate of the mixture distribution using a normal distribution has the mean and the covariance equal to those of the mixture. However, since we want the mode to be contained in the confidence interval, we estimate the normal mean with the mixture mode. This will be a good approximation unless the distribution is very skewed. Thus, the covariance $\boldsymbol{\Sigma}$ of the approximating normal is given by eq. (2).

If the mixture distribution is multimodal, we can use local information to obtain the covariance $\boldsymbol{\Sigma}$ of the approximating normal. In particular, at each mode of the mixture, the value of the Hessian contains

information of how flat or how peaked the distribution $p(\mathbf{x})$ is around that mode. Consider a mode of the mixture at $\mathbf{x} = \boldsymbol{\nu}$ with Hessian $\mathbf{H}$. Since it is a maximum, $\mathbf{H} < 0$, and so we can write the singular value decomposition of $\mathbf{H}$ as $\mathbf{H} = -\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ where $\boldsymbol{\Lambda} > 0$. Since the Hessian of a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ at its mode is equal to $-|2\pi\boldsymbol{\Sigma}|^{-1/2}\,\boldsymbol{\Sigma}^{-1}$, from eq. (4), equating this to our known mixture Hessian $\mathbf{H}$ we obtain $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{S}^{-1}\mathbf{U}^T$ where $\mathbf{S} = (2\pi)^{\frac{D}{D+2}}\,|\boldsymbol{\Lambda}|^{-\frac{1}{D+2}}\,\boldsymbol{\Lambda} = |2\pi\boldsymbol{\Lambda}^{-1}|^{\frac{1}{D+2}}\,\boldsymbol{\Lambda}$, as can be easily confirmed by substitution. So $\boldsymbol{\Sigma} = |2\pi(-\mathbf{H})^{-1}|^{-\frac{1}{D+2}}\,(-\mathbf{H})^{-1}$. Thus, we can approximate the mixture probability to second order in a neighbourhood near its mode $\boldsymbol{\nu}$ as

$$p(\mathbf{x}) \approx p(\boldsymbol{\nu}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\nu})^T\mathbf{H}(\mathbf{x}-\boldsymbol{\nu}) = p(\boldsymbol{\nu}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\nu})^T\left(|2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\,\boldsymbol{\Sigma}^{-1}\right)(\mathbf{x}-\boldsymbol{\nu}).$$

Using the log-density $L(\mathbf{x}) = \ln p(\mathbf{x})$, call $\mathbf{H}'$ the Hessian of L at a mode $\boldsymbol{\nu}$. The second order approximation near the mode $\boldsymbol{\nu}$

$$L(\mathbf{x}) \approx L(\boldsymbol{\nu}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\nu})^T\mathbf{H}'(\mathbf{x}-\boldsymbol{\nu}) \Longrightarrow p(\mathbf{x}) = e^{L(\mathbf{x})} \approx p(\boldsymbol{\nu})e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\nu})^T\mathbf{H}'(\mathbf{x}-\boldsymbol{\nu})}$$

gives $\boldsymbol{\Sigma} = (-\mathbf{H}')^{-1}$. Note that, from eq. (8), $\mathbf{H}' = \frac{1}{p(\boldsymbol{\nu})}\mathbf{H}$.

6.3 Error bars at the mode of the mixture

From the previous discussion we can derive the following algorithm. Choose a confidence level $0 < P < 1$ and compute $\rho = \sqrt{2}\,\mathrm{arg\,erf}\,(P^{1/D})$. Given a vector $\boldsymbol{\nu}$ and a negative definite matrix $\mathbf{H}$ representing a mode of the mixture and the Hessian at that mode, respectively, and calling $\boldsymbol{\Sigma}$ the covariance matrix of the mixture:

- If the mixture is unimodal, then decompose $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ with $\mathbf{U}$ orthogonal and $\boldsymbol{\Lambda} > 0$ diagonal. The D principal error bars are centred in $\boldsymbol{\nu}$, directed along the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_D$ and with lengths $2\rho\sqrt{\lambda_1}, \ldots, 2\rho\sqrt{\lambda_D}$.
- If the mixture is multimodal, then:
  - If $\mathbf{H}$ is the Hessian of $p(\mathbf{x})$, then decompose $-\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ with $\mathbf{U}$ orthogonal and $\boldsymbol{\Lambda} > 0$ diagonal. Compute $\mathbf{S} = |2\pi\boldsymbol{\Lambda}^{-1}|^{\frac{1}{D+2}}\,\boldsymbol{\Lambda}$. The D principal error bars are centred in $\boldsymbol{\nu}$, directed along the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_D$ and with lengths $\frac{2\rho}{\sqrt{s_1}}, \ldots, \frac{2\rho}{\sqrt{s_D}}$.
  - If $\mathbf{H}$ is the Hessian of $L(\mathbf{x}) = \ln p(\mathbf{x})$, then decompose $-\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ with $\mathbf{U}$ orthogonal and $\boldsymbol{\Lambda} > 0$ diagonal. The D principal error bars are centred in $\boldsymbol{\nu}$, directed along the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_D$ and with lengths $\frac{2\rho}{\sqrt{\lambda_1}}, \ldots, \frac{2\rho}{\sqrt{\lambda_D}}$.

6.4 Discussion

Since we are only using second order local information, namely the Hessian at the mode, the best we can do is to approximate the mixture quadratically, as we have shown. This approximation is only valid in a small neighbourhood around the mode, which can lead to poor estimates of the error bars, as fig. 6(left) shows. In this case, the one-dimensional mixture looks like a normal distribution but with its top flattened. Thus, the Hessian there is very small, which in turn gives a very large (co)variance $\boldsymbol{\Sigma}$ for the normal with the same Hessian (dashed line). In this particular case, one can find a better normal distribution giving more accurate error bars, for example the one in dotted line (which has the same variance as the mixture). But when the mixture is multimodal, finding a better normal estimate of the mixture would require a more complex procedure (even more so in higher dimensions). At any rate, a small Hessian indicates a flat top and some uncertainty in the mode.

Observe that, while the directions of the bars obtained from $L(\mathbf{x}) = \ln p(\mathbf{x})$ coincide with those from $p(\mathbf{x})$ always, the lengths are different in general (except when $p(\mathbf{x})$ is Gaussian). Figure 6 illustrates the point.

7 Quantifying the sparseness of a Gaussian mixture

Besides finding the modes, one may also be interested in knowing whether the density is sparse—sharply peaked around the modes, with most of the probability mass concentrated around a small region around each mode—or whether its global aspect is flat. For example, if the Gaussian mixture under consideration represents the conditional distribution of variables $\mathbf{x}$ given the values of other variables $\mathbf{y}$, the modes could be taken as possible values of the mapping $\mathbf{y} \to \mathbf{x}$ provided that the distribution is sparse (Carreira-Perpiñán, 2000). That is, a sparse distribution would correspond to a functional relationship (perhaps multivalued) while

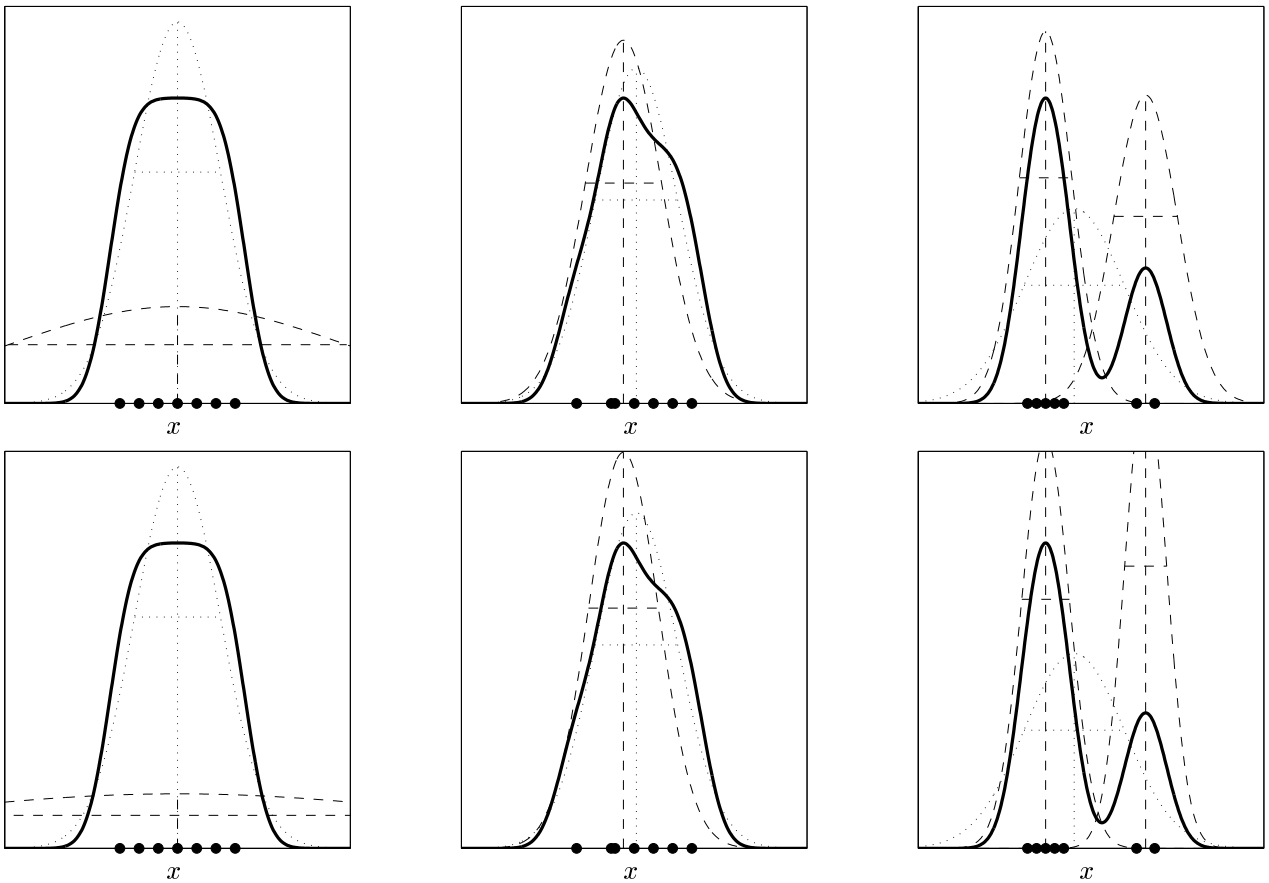

Figure 6: Error bars for a one-dimensional mixture of Gaussian distributions. The top graphs correspond to the bars obtained from $p(x)$ and the bottom ones to those obtained from $L(x) = \ln p(x)$. In each graph, the line codes are as follows. Thick solid line: the mixture distribution $p(x) = \sum_{m=1}^{M} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma}\right)^2}$ with $M = 7$ components, where $\sigma = 1$ and the component centroids $\{\mu_m\}_{m=1}^{M}$ are marked on the horizontal axis. The dashed lines indicate the approximating normals using the Hessian method and the dotted ones the approximating normals using the mixture covariance method. In both cases, the vertical line(s) indicate the location of the mode(s) or mixture mean, respectively, and the horizontal one the error bars for a confidence of 68%, i.e., the interval is two standard deviations long. On the left graph, the Hessian gives too broad an approximation because the mixture top is very flat. On the centre graph, both methods give a similar result. On the right graph, the mixture covariance method breaks down due to the bimodality of the mixture. Observe how the bars obtained from $p(x)$ do not coincide with those from $\ln p(x)$, being sometimes narrower and sometimes wider.

a flat distribution would correspond to independence (fig. 7). Of course, these are just the two extremes of a continuous spectrum.

While the error bars locally characterise the peak widths, we can globally characterise the degree of sparseness of a distribution $p(\mathbf{x})$ by its differential entropy $h(p) \stackrel{\text{def}}{=} \mathrm{E}\{-\ln p\}$: high entropy corresponds to flat distributions, where the variable $\mathbf{x}$ can assume practically any value in its domain (fig. 7, right); and low entropy corresponds to sparse distributions, where $\mathbf{x}$ can only assume a finite set of values (fig. 7, left). This differential entropy value should be compared to the differential entropy of a reference distribution, e.g. a Gaussian distribution of the same covariance or a uniform distribution on the same range as the inputs.

There are no analytical expressions for the entropy of a Gaussian mixture, but in appendix E we derive

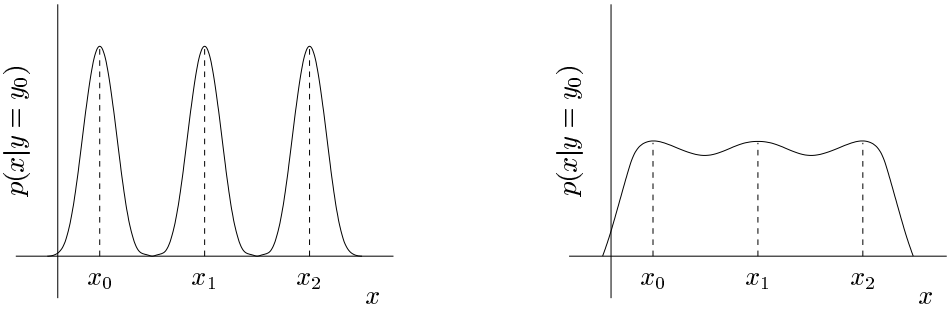

Figure 7: Shape of the conditional distribution of variable x given value y_0 of variable y. Left: sparse (multiply peaked), low entropy; x is almost functionally dependent on y for $y = y_0$, with $f(y_0) = x_0$ or x_1 or x_3. Right: flat, high entropy; x is almost independent of y for $y = y_0$.

the following upper (UB_1) and lower bounds (LB_1, LB_2):

$$\mathsf{LB}_1 \stackrel{\text{def}}{=} \frac{1}{2} \log \left\{ (2\pi e)^D \prod_{m=1}^{M} |\mathbf{\Sigma}_m|^{\pi_m} \right\}$$

$$\mathsf{LB}_2 \stackrel{\text{def}}{=} -\log \left\{ \sum_{m,n=1}^{M} p(m)p(n) \, |2\pi(\mathbf{\Sigma}_m + \mathbf{\Sigma}_n)|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{\mu}_m - \boldsymbol{\mu}_n)^T (\mathbf{\Sigma}_m + \mathbf{\Sigma}_n)^{-1} (\boldsymbol{\mu}_m - \boldsymbol{\mu}_n)} \right\}$$

$$\mathsf{UB}_1 \stackrel{\text{def}}{=} \frac{1}{2} \log \left((2\pi e)^D \, |\mathbf{\Sigma}| \right).$$

where $\mathbf{\Sigma}$ is the covariance matrix of the mixture, given in eq. (2). Other approximations to the differential entropy of an arbitrary continuous distribution are available. These are usually based on an Edgeworth polynomial expansion of the density, which leads to the use of cumulants, such as skew and kurtosis, and have been often used in the context of projection pursuit and independent component analysis (Jones and Sibson, 1987; Hyvärinen, 1998). However, cumulant-based approximations have the disadvantage of being much more sensitive to structure in the tails than in the centre of the distribution. Besides, the kurtosis is not meaningful for multimodal distributions.

Kurtosis, the fourth-order cumulant of a distribution, has been proposed and used by Field (1994) and others as a measure of sparseness in the context of neural codes, based on the experimental observation that the receptive fields of neurons in primary visual cortex typically have positive kurtosis—although this fact has been debated both theoretically and experimentally (e.g. Baddeley, 1996). However, kurtosis is not a good measure of sparseness in the sense we have described above, since we are not interested in the shape (or skew) of a peak, but in its narrowness and in the narrowness of the other peaks. In fact, we can have a distribution depending on one variance parameter (uniform, normal, double exponential) with constant kurtosis independent of that parameter (negative, zero, positive) but whose width can vary from zero (maximum sparseness, entropy $= -\infty$) to infinity (minimum sparseness, entropy $= +\infty$). As for the variance, it would be an appropriate measure of sparseness for unimodal distributions, but not for multimodal ones, since in this case the variance depends not just on the peaks' width but also on their separations (this also applies to the kurtosis). In fact, by taking the extreme case of sparse distribution, a mixture of delta functions, we can obtain any desired value for its variance and kurtosis by varying the separation between individual deltas, while its differential entropy remains constant at $-\infty$. In all these cases, the entropy gives a more natural measure of sparseness[5].

8 Conclusions

We have presented algorithms to find all the modes of a given Gaussian mixture based on the intuitive conjecture that the number of modes is upper bounded by the number of components and that the modes are contained in the convex hull of the component centroids. While our proof of this conjecture is only partial (for a mixture of two components of arbitrary dimension and equal covariance), no counterexample has been found in a number of simulations. The only other related works we are aware of (Behboodian, 1970; Konstantellos, 1980) also provided partial proofs under various restricted conditions. All algorithms have been extensively tested for the case of spherical components in simulated mixtures, in an inverse kinematics problem for a robot arm (unpublished results) and in the context of a missing data reconstruction algorithm applied to a speech

[5] However, the entropy is still not ideal, since for a mixture of a delta function and a nonzero variance Gaussian it will still give a value of $-\infty$, which does not seem appropriate. We are investigating other quantitative measures of sparseness.

inverse problem, the acoustic-to-articulatory mapping (Carreira-Perpiñán, 2000), which requires finding all the modes of a Gaussian mixture of about 1000 components in over 60 dimensions for every speech frame in an utterance.

Given the current interest in the machine learning and computer vision literature in probabilistic models able to represent multimodal distributions (specially Gaussian mixtures), these algorithms could be of benefit in a number of applications or as part of other algorithms. Specifically, they could be applied to clustering and regression problems. An example of clustering application is the determination of subclustering within galaxy systems from the measured position (right ascension and declination) and redshifts of individual galaxies (Pisani, 1993). The density of the position-velocity distribution is often modelled as a Gaussian mixture (whether parametrically or nonparametrically via kernel estimation) whose modes correspond in principle to gravitationally bound galactic structures. An example of regression application is the representation of multivalued mappings (which are often the result of inverting a forward mapping) with a Gaussian mixture (Carreira-Perpiñán, 2000). In this approach, all variables (arguments $\mathbf{x}$ and results $\mathbf{y}$ of a mapping) are jointly modelled by a Gaussian mixture and the mapping $\mathbf{x} \rightarrow \mathbf{y}$ is defined as the modes of the conditional distribution $p(\mathbf{y}|\mathbf{x})$, itself a Gaussian mixture.

The algorithms described here can be easily adapted to find minima of the mixture rather than maxima. However one must constrain them to search only for proper minima and avoid following the improper minima at $p(\mathbf{x}) \rightarrow 0$ when $\|\mathbf{x}\| \rightarrow \infty$.

8.1 Bump-finding rather than mode-finding

If we want to pick representative points of an arbitrary density $p(\mathbf{x})$, not necessarily a Gaussian mixture, using a mode as a reconstructed point is not appropriate in general because the optimal value (in the L_2 sense) is the mean. That is, summarising the whole distribution $p(\mathbf{x})$ in a single point $\hat{\mathbf{x}}$ is optimised by taking the mean of the distribution, $\mathrm{E}_{p(\mathbf{x})}\{\mathbf{x}\}$, rather than the (global) mode, $\arg\max_{\mathbf{x}} p(\mathbf{x})$, since the mean minimises the average squared error $\mathrm{E}_{p(\mathbf{x})}\left\{\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right\}$, as can easily be seen. Besides, the modes are very sensitive to the idiosyncrasies of the training data and in particular to outliers—they are not robust statistics. This suggests that, when the conditional distribution is multimodal, we should look for *bumps*[6] associated to the correct values and take the means of these bumps as reconstructed values instead of the modes. If these bumps are symmetrical then the result would coincide with picking the modes, but if they are skewed, they will be different.

How to select the bumps and their associated probability distribution is a difficult problem not considered here. A possible approach would be to decompose the distribution $p(\mathbf{x})$ as a mixture:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k) \tag{13}$$

where $p(\mathbf{x}|k)$ is the density associated to the k-th bump. This density should be localised in the space of $\mathbf{x}$ but can be asymmetrical. If $p(\mathbf{x})$ is modelled by a mixture of Gaussians (as is the case in this paper) then the decomposition (13) could be attained by regrouping Gaussian components. What components to group together is the problem; it could be achieved by a clustering algorithm—but this is dangerous if one does not know the number of clusters (or bumps) to be found. Computing then the mean of each bump would be simple, since each bump is a Gaussian mixture itself.

This approach would avoid the exhaustive mode finding procedure, replacing it by a grouping and averaging procedure—probably much faster. And again, in the situations of section 4.2, low-probability components from the Gaussian mixture may be discarded to accelerate the procedure.

These ideas operate exclusively with the functional form of a Gaussian mixture as starting point, as do our mode-finding algorithms. Bump-finding methods that work directly with a data sample exist, such as algorithms that partition the space of the $\mathbf{x}$ variables into boxes where $p(\mathbf{x})$ (or some arbitrary function of $\mathbf{x}$) takes a large value on the average compared to the average value over the entire space, e.g. PRIM (Friedman and Fisher, 1999); or some nonparametric and parametric clustering methods, e.g. scale-space clustering (Wilson and Spann, 1990; Roberts, 1997) or methods based on morphological transformations (Zhang and Postaire, 1994).

[6] Bumps of a density function $p(\mathbf{x})$ are usually defined as continuous regions where $p''(\mathbf{x}) < 0$, while modes are points where $p'(\mathbf{x}) = 0$ and $p''(\mathbf{x}) < 0$ (Scott, 1992). However, there does not seem to be agreement on the definition of "bumps," "modes" or even "peaks" in the literature. Titterington, in his comments to Friedman and Fisher (1999), claims that "bump-hunting" has tended to be used specifically for identifying and even counting the number of modes in a density function, rather than for finding fairly concentrated regions where the density is comparatively high.

9 Internet files

A Matlab implementation of the mode-finding algorithms and error bars computation is available in the WWW at `http://www.dcs.shef.ac.uk/~miguel/papers/cs-99-03.html`.

Acknowledgements

We thank Steve Renals for useful conversations and for comments on the original manuscript.

A Symbols used

$\mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	D-dimensional normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\mathbf{I}_D$	$D \times D$ identity matrix.
$L(\mathbf{x})$	The logarithm of the density $p(\mathbf{x})$, $L(\mathbf{x}) \stackrel{\text{def}}{=} \ln p(\mathbf{x})$.
$\mathbf{g}$, $\mathbf{g}(\mathbf{x})$	Gradient vector (of the density $p(\mathbf{x})$ with respect to the independent variable $\mathbf{x}$).
$\mathbf{H}$, $\mathbf{H}(\mathbf{x})$	Hessian matrix (of the density $p(\mathbf{x})$ with respect to the independent variable $\mathbf{x}$).
$\|\cdot\|$ or $\|\cdot\|_2$	The Euclidean norm: $\|\mathbf{x}\|_2^2 \stackrel{\text{def}}{=} \sum_{d=1}^D x_d^2$; or, for a density p: $\|p\|_2^2 \stackrel{\text{def}}{=} \int p^2(\mathbf{x})\, d\mathbf{x} = \mathrm{E}_{p(\mathbf{x})}\{p(\mathbf{x})\}$.
$\|\cdot\|_\infty$	The maximum norm: $\|\mathbf{x}\|_\infty \stackrel{\text{def}}{=} \max_{d=1,\ldots,D}\{|x_d|\}$.
$h(p)$	Differential entropy of density p: $h(p) \stackrel{\text{def}}{=} \mathrm{E}\{-\ln p\} = -\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x}$.
$\langle p, q \rangle$	Scalar product of densities p, q: $\langle p, q \rangle \stackrel{\text{def}}{=} \int p(\mathbf{x}) q(\mathbf{x})\, d\mathbf{x}$.
$\mathrm{E}_{p(\mathbf{x})}\{\mathbf{f}(\mathbf{x})\}$	Mean of $\mathbf{f}(\mathbf{x})$ with respect to the distribution of $\mathbf{x}$: $\mathrm{E}_{p(\mathbf{x})}\{\mathbf{f}(\mathbf{x})\} \stackrel{\text{def}}{=} \int \mathbf{f}(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}$.

B Modes of a finite mixture of normal distributions

There is no analytical expression for the modes of a finite mixture of normal distributions. However, numerical computation of all the modes is straightforward using gradient ascent, because the number of modes cannot be more than the number of components in the mixture, if we believe conjecture B.1. 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 B.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 $\{\boldsymbol{\mu}_m\}_{m=1}^M$.

Proof. The following proof is only valid for the particular case where $M = 2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. I have not been able to prove this conjecture in general.

Let us prove that, for $M = 2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the gradient becomes null only in one, two or three points, which lie in the convex hull of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Assume without loss of generality that the centroids are all different. For $\{\lambda_m\}_{m=1}^M \subset \mathbb{R}$ and $\sum_{m=1}^M \lambda_m = 1$, the set of the points

$$\mathbf{x} = \sum_{m=1}^M \lambda_m \boldsymbol{\mu}_m = \sum_{m=1}^{M-1} \lambda_m \boldsymbol{\mu}_m + \left(1 - \sum_{m=1}^{M-1} \lambda_m\right) \boldsymbol{\mu}_M = \boldsymbol{\mu}_M + \sum_{m=1}^{M-1} \lambda_m (\boldsymbol{\mu}_m - \boldsymbol{\mu}_M)$$

is the minimal linear manifold containing $\{\boldsymbol{\mu}_m\}_{m=1}^M$, i.e., the hyperplane passing through all the centroids. Note that this is not necessarily a vector subspace, because it may not contain the zero vector. However, the set

$$\left\{ \mathbf{y} : \mathbf{y} = \sum_{m=1}^{M-1} \lambda_m (\boldsymbol{\mu}_m - \boldsymbol{\mu}_M) \text{ with } \{\lambda_m\}_{m=1}^{M-1} \subset \mathbb{R} \right\}$$

is a vector subspace, namely the one spanned by $\{\boldsymbol{\mu}_m - \boldsymbol{\mu}_M\}_{m=1}^{M-1}$. Then, an arbitrary point $\mathbf{x} \in \mathbb{R}^D$ can be decomposed as $\mathbf{x} = \sum_{m=1}^M \lambda_m \boldsymbol{\mu}_m + \mathbf{w}$, where $\{\lambda_m\}_{m=1}^M \subset \mathbb{R}$ with $\sum_{m=1}^M \lambda_m = 1$ and $\mathbf{w}$ is a vector orthogonal to that manifold, i.e., orthogonal to $\{\boldsymbol{\mu}_m - \boldsymbol{\mu}_M\}_{m=1}^{M-1}$. Let us now compute the zero-gradient points of $p(\mathbf{x})$ from eq. (5):

$$\mathbf{g}(\mathbf{x}) = p(\mathbf{x}) \sum_{m=1}^M p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{\mu}_m - \mathbf{x}) = p(\mathbf{x}) \left(\sum_{m=1}^M p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} \left(\boldsymbol{\mu}_m - \sum_{n=1}^M \lambda_n \boldsymbol{\mu}_n \right) - \sum_{m=1}^M p(m|\mathbf{x}) \boldsymbol{\Sigma}_m^{-1} \mathbf{w} \right) = \mathbf{0}.$$

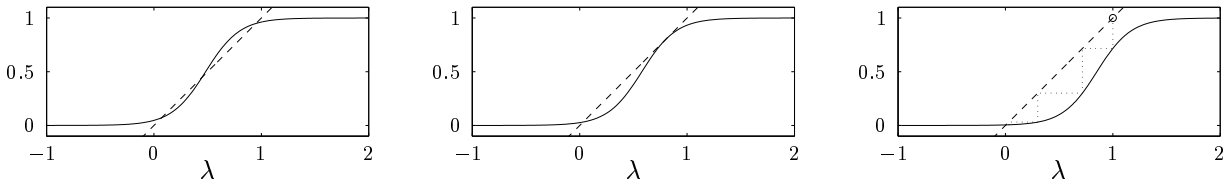

Figure 8: Three possible cases for the solutions of the equation $\lambda = f(\lambda)$, where $f(\lambda) = \frac{1}{1+e^{-\alpha(\lambda-\lambda_0)}}$. The solid line corresponds to $f(\lambda)$ and the dashed one to λ. The right figure also shows in dotted line the sequence of fixed-point iterations starting from $\lambda = 1$ (marked "$\circ$"), converging to a fixed point slightly larger than 0.

For $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}\ \forall m = 1, \ldots, M$ this becomes

$$\mathbf{g}(\mathbf{x}) = p(\mathbf{x})\boldsymbol{\Sigma}^{-1}\left(\sum_{m=1}^{M} p(m|\mathbf{x})\left(\boldsymbol{\mu}_m - \sum_{n=1}^{M}\lambda_n\boldsymbol{\mu}_n\right) - \mathbf{w}\right) = \mathbf{0} \Longrightarrow \mathbf{v} - \mathbf{w} = \mathbf{0}$$

where $\mathbf{v} = \sum_{m=1}^{M} p(m|\mathbf{x})\left(\boldsymbol{\mu}_m - \sum_{n=1}^{M}\lambda_n\boldsymbol{\mu}_n\right)$ clearly lies in the linear manifold spanned by $\{\boldsymbol{\mu}_m\}_{m=1}^{M}$ and therefore is orthogonal to $\mathbf{w}$. So $\mathbf{v} = \mathbf{w} = 0$. This proves that $\mathbf{x} = \sum_{m=1}^{M}\lambda_m\boldsymbol{\mu}_m$, i.e., all stationary points must lie in the the linear manifold spanned by the centroids.

Now

$$\begin{aligned}\mathbf{v} = \sum_{m=1}^{M} p(m|\mathbf{x})\left(\boldsymbol{\mu}_m - \sum_{n=1}^{M}\lambda_n\boldsymbol{\mu}_n\right) &= \sum_{m=1}^{M} p(m|\mathbf{x})\boldsymbol{\mu}_m - \sum_{n=1}^{M}\lambda_n\boldsymbol{\mu}_n \\ &= \sum_{m=1}^{M}(p(m|\mathbf{x}) - \lambda_m)\boldsymbol{\mu}_m = \sum_{m=1}^{M-1}(p(m|\mathbf{x}) - \lambda_m)(\boldsymbol{\mu}_m - \boldsymbol{\mu}_M) = \mathbf{0}.\end{aligned}$$

is a nonlinear system of D equations with unknowns $\lambda_1, \ldots, \lambda_{M-1}$, very difficult to study in general. For $M = 2$, call $\lambda = \lambda_1$, so that $\lambda_2 = 1 - \lambda$, and $\pi = p(1)$, so that $p(2) = 1 - \pi$. Using Bayes' theorem we get:

$$(p(1|\mathbf{x}) - \lambda)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0} \Longrightarrow \lambda = p(1|\mathbf{x}) = \frac{p(1)p(\mathbf{x}|1)}{p(1)p(\mathbf{x}|1) + p(2)p(\mathbf{x}|2)}$$

which reduces to the transcendental equation

$$\lambda = \frac{1}{1+e^{-\alpha(\lambda-\lambda_0)}} \qquad \alpha = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in (0, \infty) \qquad \lambda_0 = \frac{1}{2} + \ln\frac{1-\pi}{\pi} \in (-\infty, \infty). \tag{14}$$

It is easy to see geometrically (see fig. 8) that this equation can only have one, two or three roots in $(0, 1)$, which proves that the stationary points $\mathbf{x} = \lambda\boldsymbol{\mu}_1 + (1-\lambda)\boldsymbol{\mu}_2$ of p lie in the convex hull of $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. Further, since for $M = 2$ the convex hull is a line segment, in the case with three stationary points one of them cannot be a maximum, and so the number of modes is $M = 2$ at the most.

In fact, solving eq. (14) gives these stationary points (e.g. by fixed-point iteration, see fig. 8 right), and using eq. (6) to compute the Hessian will determine whether they are a maximum, a minimum or a saddle-point. □

Remark. 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 $\boldsymbol{\Sigma}_m$, $p(x)$ 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: $M = 2$, $\pi_1 = \pi_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $D > 1$; and $M = 2$ and $D = 2$, with no restriction on π_m, $\boldsymbol{\Sigma}_m$.

Remark. Although the previous proof makes use of formula (5), which is only valid for normal components, intuitively there is nothing special about the normal distribution here. In fact, the conjecture should hold for components of other functional forms (not necessarily positive), as long as they are bounded, piecewise continuous and have a unique maximum and no minima (*bump* functions). Note that in some cases an infinite number of stationary points may exist (as is easily seen for e.g. triangular bumps).

C Efficient operations with the Hessian

The Hessian and the gradient of a Gaussian mixture are a linear superposition of terms. This makes possible, in some particular but useful cases (e.g. GTM or Gaussian kernel density estimation), to perform efficiently and exactly (i.e., with no approximations involved) various operations required by the optimisation procedure of section 4:

- If the matrix $\mathbf{S}$ defined below is easily invertible (e.g., if each covariance matrix $\mathbf{\Sigma}_m$ is diagonal, a common situation in many engineering applications), the quadratic step of eq. (9) can be performed without inverting the Hessian (theorem C.2 and observation C.3).
- If the number of mixture components is smaller than the dimensionality of the space, $M < D$, the inverse Hessian can be computed inverting an $M \times M$ matrix rather than a $D \times D$ matrix (theorem C.1).

Define the following matrices:

$$\mathbf{S}_{D\times D} \stackrel{\text{def}}{=} -\sum_{m=1}^{M} p(\mathbf{x}, m)\mathbf{\Sigma}_m^{-1}$$

$$\mathbf{R}_{D\times M} = (\mathbf{r}_1, \ldots, \mathbf{r}_M) \text{ where } \mathbf{r}_m \stackrel{\text{def}}{=} \sqrt{p(\mathbf{x}, m)}\mathbf{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x})$$

$$\mathbf{T}_{M\times M} \stackrel{\text{def}}{=} \mathbf{I}_M + \mathbf{R}^T\mathbf{S}^{-1}\mathbf{R}$$

so that $\mathbf{R}\mathbf{R}^T = \sum_{m=1}^{M} \mathbf{r}_m\mathbf{r}_m^T$ and $\mathbf{H} = \mathbf{S} + \mathbf{R}\mathbf{R}^T$, from eq. (6).

In the sequel, proofs will often make use of the Sherman-Morrison-Woodbury (SMW) formula (Press et al., 1992):

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}.$$

Theorem C.1. $\mathbf{H}^{-1} = \mathbf{S}^{-1}(\mathbf{S} - \mathbf{R}\mathbf{T}^{-1}\mathbf{R}^T)\mathbf{S}^{-1}$.

Proof. By the SMW formula. □

Theorem C.2. $\mathbf{H}^{-1}\mathbf{g} = \mathbf{S}^{-1}\mathbf{H}\mathbf{S}^{-1}\mathbf{g}$. *If* $\mathbf{\Sigma}_m = \sigma^2\mathbf{I}_D$ *then* $\mathbf{H}^{-1}\mathbf{g} = \left(\frac{\sigma^2}{p(\mathbf{x})}\right)^2 \mathbf{H}\mathbf{g}$.

Proof. Define an $M \times 1$ vector $\mathbf{v}$ with components $v_m \stackrel{\text{def}}{=} \sqrt{p(\mathbf{x}, m)}$, so that $\mathbf{R}\mathbf{v} = \mathbf{g}$ (see eq. (5)). Then:

$$\mathbf{H}^{-1}\mathbf{g} = \mathbf{H}^{-1}\mathbf{R}\mathbf{v} = \mathbf{S}^{-1}\mathbf{R}\mathbf{T}^{-1}\mathbf{v} = \mathbf{S}^{-1}\mathbf{R}(\mathbf{v} + \mathbf{R}^T\mathbf{S}^{-1}\mathbf{g}) =$$
$$\mathbf{S}^{-1}\mathbf{g} + \mathbf{S}^{-1}\mathbf{R}\mathbf{R}^T\mathbf{S}^{-1}\mathbf{g} = \mathbf{S}^{-1}(\mathbf{S} + \mathbf{R}\mathbf{R}^T)\mathbf{S}^{-1}\mathbf{g} = \mathbf{S}^{-1}\mathbf{H}\mathbf{S}^{-1}\mathbf{g}$$

where we have used $\mathbf{H}^{-1}\mathbf{R} = \mathbf{S}^{-1}\mathbf{R}\mathbf{T}^{-1}$, which again can be proved using the SMW formula. □

Observation C.3. *Since* $\mathbf{H} = \mathbf{S} + \sum_{m=1}^{M} \mathbf{r}_m\mathbf{r}_m^T$, *the Hessian can be inverted by repeatedly using the following particular case of the SMW formula:*

$$\mathbf{A}_{D\times D}, \mathbf{v}_{D\times 1}: \quad (\mathbf{A} + \mathbf{v}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{v}}.$$

The only operation not covered here that is required by the optimisation procedure is the determination of whether the Hessian is negative definite. This can be accomplished by checking the signs of its principal minors (Mirsky, 1955, p. 403) or by numerical methods, such as Cholesky decomposition (Press et al., 1992, p. 97).

D Bounds for the gradient and the Hessian

Theorem D.1. *For a normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, *the norm of the gradient is bounded as* $0 \leq \|\mathbf{g}\| \leq (e\lambda_{min}\,|2\pi\mathbf{\Sigma}|)^{-1/2}$, *where* λ_{min} *is the smallest eigenvalue of* $\mathbf{\Sigma}$.

Proof. Obviously $\|\mathbf{g}\| \geq 0$, with $\mathbf{g} = \mathbf{0}$ attained at $\mathbf{x} = \boldsymbol{\mu}$. For a one-dimensional distribution $\mathcal{N}(\mu, \lambda)$, the maximum gradient norm, of value $(2\pi\lambda^2 e)^{-1/2}$, happens at the inflexion points $\left|\frac{x-\mu}{\sqrt{\lambda}}\right| = 1$. Using these results, let us prove the theorem for the D-dimensional normal $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. Here $\mathbf{g} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x})p(\mathbf{x})$ and $\|\mathbf{g}\| = \|\mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x})\|\,p(\mathbf{x})$. Change $\mathbf{z} = \mathbf{\Lambda}^{-1}\mathbf{U}^T(\boldsymbol{\mu} - \mathbf{x})$ where $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the singular value decomposition of $\mathbf{\Sigma}$:

$$\|\mathbf{g}\| = \left\|\mathbf{U}\mathbf{\Lambda}^{1-}\mathbf{U}^T(\boldsymbol{\mu} - \mathbf{x})\right\| p(\mathbf{x}) = \|\mathbf{U}\mathbf{z}\|\,|2\pi\mathbf{\Sigma}|^{-\frac{1}{2}}\,e^{-\frac{1}{2}\mathbf{z}^T\mathbf{\Lambda}\mathbf{z}} = \|\mathbf{z}\|\,|2\pi\mathbf{\Lambda}|^{-\frac{1}{2}}\,e^{-\frac{1}{2}\mathbf{z}^T\mathbf{\Lambda}\mathbf{z}}$$

since $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D)$ is orthonormal. Consider ellipses where $\mathbf{z}^T\mathbf{\Lambda}\mathbf{z} = \rho^2$ with $\rho > 0$. In any of those ellipses, the point with maximum norm (i.e., maximum distance to the origin) lies along the direction of the smallest $\lambda_d = \lambda_{\min}$, for some[7] $d \in \{1, \ldots, D\}$. Thus, using Dirac delta notation, $z_c = \pm\delta_{cd}\rho\lambda_d^{-1/2}$ and $\|\mathbf{g}\| = \rho\lambda_d^{-1/2}\,|2\pi\mathbf{\Lambda}|^{-1/2}\,e^{-\frac{1}{2}\rho^2}$ there. Now, from the result for the one-dimensional case, this expression is maximum at $\rho = 1$. So the gradient norm is maximum at $\mathbf{z}^*$ with components $z_c^* = \pm\delta_{cd}\rho\lambda_d^{-1/2}$ for $c = 1, \ldots, c$, or $\mathbf{x}^* = \boldsymbol{\mu} - \mathbf{U}\mathbf{\Lambda}\mathbf{z}^* = \boldsymbol{\mu} \pm \lambda_d^{1/2}\mathbf{u}_d$ and $\|\mathbf{g}^*\| = (e\lambda_{\min}\,|2\pi\mathbf{\Lambda}|)^{-1/2} = (e\lambda_{\min}\,|2\pi\mathbf{\Sigma}|)^{-1/2}$ there. □

[7] There may be several values of d with the same $\lambda_d = \lambda_{\min}$, but this is irrelevant for the proof.

Corollary D.2. *If $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_D$, then $\|\mathbf{g}\|$ is maximum at the hypersphere $\left\|\frac{\boldsymbol{\mu}-\mathbf{x}}{\sigma}\right\| = 1$ and there $\|\mathbf{g}_{max}\| = (e\sigma^2(2\pi\sigma^2)^D)^{-1/2}$.*

Theorem D.3. *For a mixture of Gaussians, the gradient norm is smaller or equal than the maximum gradient norm achievable by any of the components. If $\boldsymbol{\Sigma}_m = \sigma^2 \mathbf{I}_D$ for all $m = 1, \ldots, M$, then $\|\mathbf{g}\| \leq (e\sigma^2(2\pi\sigma^2)^D)^{-1/2}$.*

Proof. For the mixture and using the triangle inequality:

$$\|\mathbf{g}\| = \left\| \sum_{m=1}^{M} p(m)\mathbf{g}_m \right\| \leq \sum_{m=1}^{M} p(m) \|\mathbf{g}_m\| \leq \sum_{m=1}^{M} p(m) \|\mathbf{g}_{m,\max}\| \leq \sum_{m=1}^{M} p(m) \|\mathbf{g}\|_{\max} = \|\mathbf{g}\|_{\max}$$

where $\|\mathbf{g}_{m,\max}\|$ is given by theorem D.1 for each component m and $\|\mathbf{g}\|_{\max} = \max_{m=1\ldots,M} \|\mathbf{g}_{m,\max}\|$. Geometrically, this is a result of the coalescence of the individual components. □

The following theorem gives a sufficient condition for the Hessian of the mixture of isotropic Gaussians to be negative definite.

Theorem D.4. *If $\boldsymbol{\Sigma}_m = \sigma^2 \mathbf{I}_D$, then $\mathbf{H} < 0$ if $\sum_{m=1}^{M} p(m|\mathbf{x}) \left\|\frac{\boldsymbol{\mu}_m - \mathbf{x}}{\sigma}\right\|^2 < 1$.*

Proof. Call $\boldsymbol{\rho}_m = \frac{\boldsymbol{\mu}_m - \mathbf{x}}{\sigma}$, $\rho_m = \|\boldsymbol{\rho}_m\|$ and $\mathbf{H}_0 = -\mathbf{I}_D + \sum_{m=1}^{M} p(m|\mathbf{x}) \boldsymbol{\rho}_m \boldsymbol{\rho}_m^T$. From eq. (6) $\mathbf{H} = \frac{p(\mathbf{x})}{\sigma^2}\mathbf{H}_0$. From lemma D.5, each eigenvalue of $\mathbf{H}_0$ is $\lambda_d = -1 + \sum_{m=1}^{M} \pi_{md} p(m|\mathbf{x}) \rho_m^2$ where $\sum_{d=1}^{D} \pi_{md} = 1$ for each $m = 1, \ldots, M$. A worst-case analysis gives $\lambda_d \leq -1 + \sum_{m=1}^{M} p(m|\mathbf{x}) \rho_m^2$ for a certain $d \in \{1, \ldots, D\}$. Thus the Hessian will be negative definite if $\sum_{m=1}^{M} p(m|\mathbf{x}) \rho_m^2 < 1$. □

The following lemma shows the effect on the eigenvalues of a symmetric matrix of a series of unit-rank perturbations and is necessary to prove theorem D.4.

Lemma D.5. *Let $\mathbf{A}$ be a symmetric $D \times D$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_D$ and $\{\mathbf{u}_m\}_{m=1}^{M}$ a set of M vectors in $\mathbb{R}^D$. Then, the eigenvalues of $\mathbf{A} + \sum_{m=1}^{M} \mathbf{u}_m \mathbf{u}_m^T$ are $\lambda'_d = \lambda_d + \sum_{m=1}^{M} \pi_{md} \mathbf{u}_m^T \mathbf{u}_m$ for some unknown coefficients π_{md} satisfying $\pi_{md} \in [0, 1]$ for $m = 1, \ldots, M$, $d = 1, \ldots, D$, and $\sum_{d=1}^{D} \pi_{md} = 1$ for each $m = 1, \ldots, M$. Thus, every eigenvalue is shifted by an amount which lies between 0 and the squared norm of the perturbing vector, for each perturbation.*

Proof. A proof for the case $M = 1$ is given in (Wilkinson, 1965, pp. 97–98). The case of arbitrary M follows by repeated application of the case $M = 1$. □

E Bounds for the entropy of a Gaussian mixture

Theorem E.1. *The entropy and L_2-norm for a normal distribution of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are:*

- $h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2} \log |2\pi e \boldsymbol{\Sigma}|$.
- $\|\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\|_2 = |4\pi\boldsymbol{\Sigma}|^{-1/4}$.

Theorem E.2 (Information theory bounds on the entropy of a mixture). *For a finite mixture $p(\mathbf{x})$ not necessarily Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$: $\sum_{m=1}^{M} \pi_m h(p(\mathbf{x}|m)) \leq h(p(\mathbf{x})) \leq \frac{1}{2} \log\left((2\pi e)^D |\boldsymbol{\Sigma}|\right)$. Equality can only be obtained in trivial mixtures where $M = 1$ or $\mathbf{x}$ and m are independent, i.e., all components are equal.*

Proof.

- LHS inequality: by the fact that conditioning reduces the entropy (or by the concavity of the entropy) (Cover and Thomas, 1991), $h(p(\mathbf{x})) \geq h(p(\mathbf{x}|m)) = -\mathrm{E}_{p(\mathbf{x},m)} \{p(\mathbf{x}|m)\} = \sum_{m=1}^{M} \pi_m h(p(\mathbf{x}|m))$.
- RHS inequality: by the fact that, for fixed mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the maximum entropy distribution is the normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, whose entropy is $\frac{1}{2} \log\left((2\pi e)^D |\boldsymbol{\Sigma}|\right)$ and the logarithm is in any base (Cover and Thomas, 1991). □

Theorem E.3 (L_2-norm lower bound on the entropy). [8] *For any density p: $h(p) \geq -2 \log \|p\|_2$.*

Proof. Since the function $-\log x$ is convex, Jensen's inequality gives $h(p) \stackrel{\text{def}}{=} \mathrm{E}_p \{-\log p(\mathbf{x})\} \geq -\log \mathrm{E}_p \{p(\mathbf{x})\} = -2 \log \|p\|_2$. □

[8] I am grateful to Chris Williams for suggesting to me the use of the L_2-norm.

Theorem E.4. *For a Gaussian mixture* $p(\mathbf{x}) = \sum_{m=1}^{M} p(m)p(\mathbf{x}|m)$*:*

$$\|p\|_2^2 = \sum_{m,n=1}^{M} p(m)p(n)\,|2\pi(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_n)|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)^T(\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)^{-1}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)}.$$

Proof. $\|p\|_2^2 = \langle p, p\rangle = \sum_{m,n=1}^{M} p(m)p(n)\,\langle \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\rangle$. Computing the scalar product is tedious and is summarised as follows:

$$\begin{aligned}
\langle \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\rangle &= \int_{\mathbb{R}^D} p(\mathbf{x}|m)p(\mathbf{x}|n)\,d\mathbf{x} = \\
&\int_{\mathbb{R}^D} |2\pi\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}\,|2\pi\boldsymbol{\Sigma}_n|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)+(\mathbf{x}-\boldsymbol{\mu}_n)^T\boldsymbol{\Sigma}_n^{-1}(\mathbf{x}-\boldsymbol{\mu}_n)\right)}\,d\mathbf{x} = \\
|2\pi\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}\,|2\pi\boldsymbol{\Sigma}_n|^{-\frac{1}{2}} &\int_{\mathbb{R}^D} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})(\mathbf{x}-\boldsymbol{\mu})-(\boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_n^{-1}+\boldsymbol{\mu}_n^T\boldsymbol{\Sigma}_m^{-1})(\mathbf{x}-\boldsymbol{\mu})-\frac{1}{2}(\boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_m+\boldsymbol{\mu}_n^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_n)}\,d\mathbf{x} = \\
&|2\pi\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}\,|2\pi\boldsymbol{\Sigma}_n|^{-\frac{1}{2}}\,\left|2\pi(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})^{-1}\right|^{\frac{1}{2}}\times \\
e^{-\frac{1}{2}(\boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_m+\boldsymbol{\mu}_n^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_n)}&e^{\frac{1}{2}(\boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_n^{-1}+\boldsymbol{\mu}_n^T\boldsymbol{\Sigma}_m^{-1})(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})^{-1}(\boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_n^{-1}+\boldsymbol{\mu}_n^T\boldsymbol{\Sigma}_m^{-1})^T} = \\
&|2\pi(\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)^T(\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)^{-1}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)}
\end{aligned}$$

where we have used the following facts:

$$\int_{\mathbb{R}^D} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a})+\mathbf{b}^T(\mathbf{x}-\mathbf{a})}\,d\mathbf{x} = \left|\frac{\mathbf{A}}{2\pi}\right|^{-\frac{1}{2}} e^{\frac{1}{2}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}}$$

$$\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})^{-1}\boldsymbol{\Sigma}_n^{-1} = (\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)^{-1}$$

$$\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})^{-1}\boldsymbol{\Sigma}_m^{-1} = (\boldsymbol{\Sigma}_m(\boldsymbol{\Sigma}_m^{-1}+\boldsymbol{\Sigma}_n^{-1})\boldsymbol{\Sigma}_m^{-1})^{-1} = (\boldsymbol{\Sigma}_m(\mathbf{I}+\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\Sigma}_m))^{-1} = \boldsymbol{\Sigma}_m^{-1} - (\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)^{-1}$$

and the SMW formula. □

Corollary E.5. *For a finite Gaussian mixture with components* $\mathbf{x}|m \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$*,* $m = 1, \ldots, M$ *and with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *given in section 2:* $\max(\mathsf{LB}_1, \mathsf{LB}_2) \leq h(p(\mathbf{x})) \leq \mathsf{UB}_1$*, where:*

$$\begin{aligned}
\mathsf{LB}_1 &\stackrel{\text{def}}{=} \frac{1}{2}\log\left\{(2\pi e)^D \prod_{m=1}^{M} |\boldsymbol{\Sigma}_m|^{\pi_m}\right\} \\
\mathsf{LB}_2 &\stackrel{\text{def}}{=} -\log\left\{\sum_{m,n=1}^{M} p(m)p(n)\,|2\pi(\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)^T(\boldsymbol{\Sigma}_m+\boldsymbol{\Sigma}_n)^{-1}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_n)}\right\} \\
\mathsf{UB}_1 &\stackrel{\text{def}}{=} \frac{1}{2}\log\left((2\pi e)^D\,|\boldsymbol{\Sigma}|\right).
\end{aligned}$$

Equality $\mathsf{LB}_1 = h(p(\mathbf{x})) = \mathsf{UB}_1$ *can only be obtained in trivial mixtures where* $M = 1$ *or* $\boldsymbol{\mu}_m = \boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ *for all* $m = 1, \ldots, M$*, in which case* $\mathsf{LB}_2 = \frac{1}{2}\log|4\pi\boldsymbol{\Sigma}| < \frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}| = \mathsf{LB}_1 = h(p(\mathbf{x})) = \mathsf{UB}_1$.

Note the limit cases:

- $\Sigma_m \to 0$: the upper and lower bounds and the entropy tend to $-\infty$.
- $\Sigma_m \to \infty$: the upper and lower bounds and the entropy tend to ∞.

Observe that LB_1 can be smaller or greater than LB_2 depending on the values of $\{\pi_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$.

F Additional results

Theorem F.1. $\operatorname{tr}(\mathbf{H}) = \sum_{m=1}^{M} p(\mathbf{x}, m)\left((\boldsymbol{\mu}_m - \mathbf{x})^T\boldsymbol{\Sigma}_m^{-2}(\boldsymbol{\mu}_m - \mathbf{x}) - \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right)\right)$.

Proof.

$$\begin{aligned}
\operatorname{tr}(\mathbf{H}) = \operatorname{tr}&\left(\sum_{m=1}^{M} p(\mathbf{x}, m)\boldsymbol{\Sigma}_m^{-1}((\boldsymbol{\mu}_m - \mathbf{x})(\boldsymbol{\mu}_m - \mathbf{x})^T - \boldsymbol{\Sigma}_m)\boldsymbol{\Sigma}_m^{-1}\right) = \\
&\sum_{m=1}^{M} p(\mathbf{x}, m)\left(\operatorname{tr}\left((\boldsymbol{\mu}_m - \mathbf{x})(\boldsymbol{\mu}_m - \mathbf{x})^T\boldsymbol{\Sigma}_m^{-2}\right) - \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right)\right) = \\
&\sum_{m=1}^{M} p(\mathbf{x}, m)\left((\boldsymbol{\mu}_m - \mathbf{x})^T\boldsymbol{\Sigma}_m^{-2}(\boldsymbol{\mu}_m - \mathbf{x}) - \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right)\right).
\end{aligned}$$

□

Corollary F.2. *If* $\boldsymbol{\Sigma}_m = \sigma^2 \mathbf{I}_D$ *for all* $m = 1, \ldots, M$, *then* $\operatorname{tr}(\mathbf{H}) = -\frac{D}{\sigma^2} p(\mathbf{x}) + \frac{1}{\sigma^2} \sum_{m=1}^{M} p(\mathbf{x}, m) \left\| \frac{\boldsymbol{\mu}_m - \mathbf{x}}{\sigma} \right\|^2$.

Theorem F.3. $\int_{\mathbb{R}^D} \operatorname{tr}(\mathbf{H})\, d\mathbf{x} = 0$.

Proof. From theorem F.1:

$$\int_{\mathbb{R}^D} \operatorname{tr}(\mathbf{H})\, d\mathbf{x} = \int_{\mathbb{R}^D} \sum_{m=1}^{M} -p(\mathbf{x}, m) \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right) d\mathbf{x} + \int_{\mathbb{R}^D} \sum_{m=1}^{M} p(\mathbf{x}, m)(\boldsymbol{\mu}_m - \mathbf{x})^T \boldsymbol{\Sigma}_m^{-2}(\boldsymbol{\mu}_m - \mathbf{x})\, d\mathbf{x} =$$
$$- \sum_{m=1}^{M} p(m) \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right) + \sum_{m=1}^{M} p(m) \int_{\mathbb{R}^D} p(\mathbf{x}|m)(\boldsymbol{\mu}_m - \mathbf{x})^T \boldsymbol{\Sigma}_m^{-2}(\boldsymbol{\mu}_m - \mathbf{x})\, d\mathbf{x}.$$

The latter integral can be solved by changing $\mathbf{z}_m = \mathbf{U}_m^T \boldsymbol{\Sigma}_m^{-1/2}(\boldsymbol{\mu}_m - \mathbf{x})$, where $\boldsymbol{\Sigma}_m = \mathbf{U}_m \boldsymbol{\Lambda}_m \mathbf{U}_m^T$ is the singular value decomposition of $\boldsymbol{\Sigma}_m$, and introducing the reciprocal of the determinant of the Jacobian, $\left|\mathbf{U}_m^T \boldsymbol{\Sigma}_m^{-1/2}\right|^{-1} = |\boldsymbol{\Sigma}_m|^{1/2}$ (since $\mathbf{U}_m$ is orthogonal):

$$\int_{\mathbb{R}^D} |2\pi\boldsymbol{\Sigma}_m|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)} (\boldsymbol{\mu}_m - \mathbf{x})^T \boldsymbol{\Sigma}_m^{-2}(\boldsymbol{\mu}_m - \mathbf{x})\, d\mathbf{x} =$$
$$\int_{\mathbb{R}^D} |2\pi\boldsymbol{\Sigma}_m|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{z}_m^T \mathbf{z}_m} \mathbf{z}_m^T \boldsymbol{\Lambda}_m^{-1} \mathbf{z}_m |\boldsymbol{\Sigma}_m|^{\frac{1}{2}}\, d\mathbf{z}_m = \mathrm{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I}_D)} \left\{ \sum_{d=1}^{D} \frac{z_{md}^2}{\lambda_{md}} \right\} = \sum_{d=1}^{D} \frac{1}{\lambda_{md}} \mathrm{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I}_D)} \left\{ z_{md}^2 \right\} =$$
$$\sum_{d=1}^{D} \frac{1}{\lambda_{md}} = \operatorname{tr}\left(\boldsymbol{\Lambda}_m^{-1}\right) = \operatorname{tr}\left(\boldsymbol{\Lambda}_m^{-1} \mathbf{U}_m^T \mathbf{U}_m\right) = \operatorname{tr}\left(\mathbf{U}_m \boldsymbol{\Lambda}_m^{-1} \mathbf{U}_m^T\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right).$$

Thus:

$$\int_{\mathbb{R}^D} \operatorname{tr}(\mathbf{H})\, d\mathbf{x} = -\sum_{m=1}^{M} p(m) \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right) + \sum_{m=1}^{M} p(m) \operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\right) = 0. \quad \square$$

Theorem F.4. $\int_{\mathbb{R}^D} \mathbf{g}\, d\mathbf{x} = \mathbf{0}$.

Proof. From eq. (5):

$$\int_{\mathbb{R}^D} \mathbf{g}\, d\mathbf{x} = \int_{\mathbb{R}^D} \sum_{m=1}^{M} p(\mathbf{x}, m) \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x})\, d\mathbf{x} = \sum_{m=1}^{M} p(m) \boldsymbol{\Sigma}_m^{-1} \int_{\mathbb{R}^D} p(\mathbf{x}|m)(\boldsymbol{\mu}_m - \mathbf{x})\, d\mathbf{x} =$$
$$\sum_{m=1}^{M} p(m) \boldsymbol{\Sigma}_m^{-1} \mathrm{E}_{\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)} \{\boldsymbol{\mu}_m - \mathbf{x}\}\, d\mathbf{x} = \mathbf{0}. \quad \square$$

References

R. J. Baddeley. Searching for filters with "interesting" output distributions: An uninteresting direction to explore? *Network: Computation in Neural Systems*, 7(2):409–421, 1996.

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. Reconstruction of sequential data with probabilistic models and continuity constraints. In Solla et al. (2000), pages 414–420.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, London, Sydney, 1991.

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.

D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, July 1994.

J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2): 123–143 (with discussion, pp. 143–162), Apr. 1999.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Texts in Statistical Science. Chapman & Hall, London, New York, 1995.

C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135 (with discussion, pp. 135–148), Feb. 1986.

G. E. Hinton. Products of experts. In *Proc. of the Ninth Int. Conf. on Artificial Neural Networks (ICANN99)*, pages 1–6, Edinburgh, UK, Sept. 7–10 1999. The Institution of Electrical Engineers.

G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Networks*, 8(1):65–74, Jan. 1997.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 273–279. MIT Press, Cambridge, MA, 1998.

E. Isaacson and H. B. Keller. *Analysis of Numerical Methods.* John Wiley & Sons, New York, London, Sydney, 1966.

M. Isard and A. Blake. CONDENSATION — conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

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.

M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society, A*, 150(1): 1–18 (with comments, pp. 19–36), 1987.

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

L. Mirsky. *An Introduction to Linear Algebra.* Clarendon Press, Oxford, 1955. Reprinted in 1982 by Dover Publications.

B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 19(7):696–710, July 1997.

A. Pisani. A nonparametric and scale-independent method for cluster-analysis. 1. the univariate case. *Monthly Notices of the Royal Astronomical Society*, 265(3):706–726, Dec. 1993.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, U.K., second edition, 1992.

L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Signal Processing Series. Prentice-Hall, Englewood Cliffs, N.J., 1993.

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

S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 20(11):1133–1142, 1998.

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.

D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1992.

S. A. Solla, T. K. Leen, and K.-R. Müller, editors. *Advances in Neural Information Processing Systems*, volume 12, 2000. MIT Press, Cambridge, MA.

M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, Feb. 1999.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1985.

V. N. Vapnik and S. Mukherjee. Support vector method for multivariate density estimation. In Solla et al. (2000), pages 659–665.

J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, New York, Oxford, 1965.

R. Wilson and M. Spann. A new approach to clustering. *Pattern Recognition*, 23(12):1413–1425, 1990.

R. D. Zhang and J.-G. Postaire. Convexity dependent morphological transformations for mode detection in cluster-analysis. *Pattern Recognition*, 27(1):135–148, 1994.