Multiresolution Gaussian Mixture Models: Theory and Applications

Roland Wilson

February 28, 2000

1 Introduction

Multiresolution image representations have been successfully applied to many problems in image analysis [26, 3, 11, 5] because of their ability to trade off spatial resolution against frequency domain resolution and their symmetry with respect to translations and dilations, two of the most important sources of variation in images [4]. Not every multiresolution representation is equally effective in this regard: schemes using dyadic decimation, such as pyramids and orthonormal wavelets, sacrifice translation invariance to reduce redundancy [6, 25, 12]. While this is desirable in image compression, it is not ideal for problems in computer vision, where compactness is less important than utility in a range of problems - from segmentation to motion analysis. In those cases, it is the combination of statistical inference, usually from incomplete (eg. projected) data and symmetry under motion which are most significant.

Another interesting development in recent years has been the use of Gaussian Mixture models to cope with statistical problems for which no simple parametric model exists [20, 27]. While it is well known that algorithms such as Expectation-Maximisation can lead to effective approximations in terms of a finite number of components, the general problem of mixture modelling is difficult when the number of components is unknown [13, 17].

This paper describes a new multiresolution representation which tackles the mixture modelling problem head-on. The general approach shares important features with the Classification and Regression Trees (CART) system of [9] and its derivatives [17]. It may also be seen as a generalisation of scale-space [26], in that it uses Gaussian functions and includes spatial co-ordinates, but unlike a conventional basis set, they are adapted to the data and used statistically. Moreover, they are defined in a space whose dimension reflects the inference problem, not simply the image data. Thus in dealing with colour images, a $5-D$ space is required (two spatial and three colour dimensions); for inferring $3-D$ structure from motion, typically nine dimensions are required (three spatial, three colour and three motion axes). Yet MGMM has no difficulty in principle in moving seamlessly between these spaces. The next section of the paper outlines the theory underlying MGMM as a method of approximating an arbitrary density and shows how it can deal with smooth motions of an image. The following section contains a description of how MGMM can be implemented efficiently. This is followed by a presentation of some simple experiments, which illustrate how it may be used in important 'early vision' problems: segmentation and motion analysis. The paper is concluded with some remarks on the implications and potential of the new approach.

2 Background

Three main elements dictate the form of representation called MGMM: ability to approximate *any* probability density in a space of arbitrary dimension; closure under affine motions and a multiresolution structure, which can be used to make computation efficient.

Before outlining the theory of MGMM, it may be useful to explain the need for such an approach. It is widely recognised that statistical methods are a key element in vision and image analysis eg. [7, 22]. While Gaussian models work adequately in many cases, more general distributions pose considerable difficulties. The first is simply estimating the density, a problem which has beset statisticians for may decades [23, 24]. There are two widely used alternatives: non-parametric, or kernel estimation and parametric estimation. Kernel methods are a generalisation of histogramming, in which, rather than use a non-overlapping set of intervals, the density is estimated using a smooth kernel to 'iron out' the fluctuations which sampling inevitably causes. In effect, this represents a convolution which is the statistical analogue of anti-aliasing: it is well known that raw histograms are woefully inadequate as density estimates [23]. The alternative, which is clearly preferable *if* there is sufficient prior information or knowledge about the processes producing the data, is to use a parametric model. In recent years, this has been extended through the use of Bayesian methods for model selection and the estimation of so-called 'hyper-parameters', which characterise the model, eg. [19, 15], but while useful in particular problems, these hardly represent a general approach to the inference problem. It seems that what is needed is a method which combines the generality of kernel estimation with the power of parametric methods. This is one of the key motivations behind the MGMM approach.

To define the MGMM approach in a general way, we start by observing that any probability density function can be approximated to an arbitrary precision by a set of Gaussian functions: it is well known that the Gaussian functions are a complete set on $L^2(R^n)$. However, we want to make a stronger claim - namely that an L^1 approximation of an arbitrary density function need only involve *positive* coefficients in the expansion. To this end, we state the theorem:

Theorem 1: Let $f(.) : R^n \to R$ be any nonnegative real integrable function on R^n with

$$\int_{R^n} d\boldsymbol{x}\, f(\boldsymbol{x}) = 1 \tag{1}$$

Then for any $\delta > 0$ there exists an approximation of $f(.)$ by a strictly positive sum of Gaussian functions of the form

$$\hat{f}(\boldsymbol{x}) = \sum_i f_i g_{\Sigma_i}(\boldsymbol{x} - \boldsymbol{\mu}_i) \tag{2}$$

of means $\boldsymbol{\mu}_i$ and covariances Σ_i, such that $f_i > 0, \forall i$ and

$$\delta > \int_{R^n} d\boldsymbol{x}\, |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \tag{3}$$

Proof: We start in one dimension and then extend the result to R^n, primarily for notational simplicity. The proof is based on the approximation of a scaled characteristic function $\chi_X(x)$, for which

$$\begin{array}{rcll} \chi_X(x) & = & \frac{1}{2\epsilon} & \text{if } x \in X \\ & & 0 & \text{else} \end{array} \tag{4}$$

where the set X is the interval $\epsilon < x \leq \epsilon$. Consider the convolution of $\chi_X(.)$ with the Gaussian function $g_\sigma(x)$ of variance σ^2

$$g_\sigma \star \chi_X(x) = \int_X dy g_\sigma(y-x) \tag{5}$$

which is easily written as

$$g_\sigma \star \chi(x) = \frac{1}{4\epsilon}[\mathrm{erf}(\frac{\epsilon + x}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{\epsilon - x}{\sqrt{2}\sigma})] \tag{6}$$

where erf(.) is the error function [1]. If $\sigma = \epsilon/M$, then for any $\delta > 0$, there exists a value of M such that

$$\begin{aligned} \int dx\ |\chi_X(x) - g_{\epsilon/M} \star \chi_X(x)| \ &< \ \int_{X^M} dx\ \chi_X(x) + \\ &\int_{X-X^M} dx \exp[-M] + \int_{R-X} dx\ g_{\epsilon/M} \star \chi_X(x) \\ &< \frac{2}{M} + 2\epsilon \exp[-M] < \frac{\delta}{2} \end{aligned} \tag{7}$$

where X^M is the subset of R for which $|x \pm \epsilon| < \frac{1}{M}$. Now the convolution, which is smooth, can be approximated by a sum

$$g_\sigma \star \chi_X(x) \approx \sum_i f_i g_\sigma(x - 2^{-k}i) \tag{8}$$

where k is chosen such that there is an integral absolute approximation error less than $\frac{\delta}{2}$

$$\int dx |g_\sigma \star \chi_X(x) - \sum_i f_i g_\sigma(x - 2^{-k})| < \frac{\delta}{2} \tag{9}$$

where $f_i > 0, \forall i$. But this is the Riemann integral theorem [8]: one simply uses for f_i the scaled values of $\chi_X(2^{-k}i)$

$$f_i = 2^{-k}\chi_X(2^{-k}i) \tag{10}$$

Then for some k, the sum, which is absolutely convergent, will give an error smaller than $\delta/2$.

We have thus proved that the function $\chi_X(.)$ can be approximated to arbitrary precision by a finite set of Gaussian functions. To complete the proof, we note that

$$\sum_i f_i \to 1 \tag{11}$$

which follows from the uniform sampling of $\chi_X(.)$, the scaling of $\chi_X(.)$, (4) and the convergence of the sum. Thus if $\chi_X(.)$ represents a density function, then (8) is a

Gaussian Mixture approximating it to the specified tolerance. But *any* probability density can be approximated by a mixture of functions of the form $\chi_X(\frac{x-y_i}{\sigma_i})$: these *simple* functions are the means by which the Lebesgue integral is defined [8] and any probability distribution is measurable.

The extension to R^n is trivial: one simply replaces the $1-d$ elements by the corresponding Cartesian separable, product form and the proof goes through in the same way. This completes the proof. $\bullet$

In order to use this result, we need an efficient way of deciding how many components to use in modelling an arbitrary density, or more realistically, data drawn from an arbitrary density. To this end, consider the following simple decision: given a set of data, is it better modelled with one Gaussian component or two? This is a decision which cannot be made on the basis of likelihood alone: it would always give preference to the more complex description. The increase in likelihood must be balanced against the *cost* in terms of complexity of using two components instead of one.

This type of problem has received much attention in the literature on classification [9, 17, 19], and in system identification, in which model order has to be chosen [10]. In both areas, widespread use has been made of Akaike's *Information Criterion* (AIC), which is derived from the Kullback-Liebler *Information Divergence* between two densities:

$$d(f,g) = \int_{R^n} d\boldsymbol{x} f(\boldsymbol{x}) \ln[\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}] \tag{12}$$

A widely used alternative to AIC is the Minimum Description Length (MDL) criterion, in which an estimate is made of the 'coding cost' of the parametric description of an object [18]. Now, in many cases, quite what the MDL of an object might be is unclear, since it would appear to depend on the choice of description language and on the density induced on strings from that language by the objects of interest. In the present context, much greater precision is available, since not only the parameters, but in principle their prior and posterior densities can be estimated. Lastly, Bayesian Evidence (BE), ie. the posterior $P(M|Y)$ supporting a model M amongst a set of alternatives, has also been propounded [17, 15] as the best way of selecting a model. In the context of selecting either one or two Gaussian components to model a set of data, this involves penalising the two-component model though the prior. It is not so clear, however, just how steep the penalty should be; at least in the AIC and in the present case in MDL, this question is settled by definition. In general, evaluation of these criteria is itself a source of considerable difficulty, with resort usually being made to simple approximations, eg. based on saddle-point approximations or sample averaging over an appropriate set of simulation epochs. In the present case, *once the data are classified*, an exact solution is possible. Classification of the data is achieved by Gibbs sampling.

Because the Bayesian paradigm is an appropriate one in many applications, let us start by considering an evidential framework for the problem of 'one-or-two'. We can base the decision between $H_1:$ one component and $H_2:$ two components on the ratio

$$R_B = \frac{P(H_1)P(\boldsymbol{X}|H_1)}{P(H_2)P(\boldsymbol{X}|H_2)} \tag{13}$$

where $P(H_1) = 1 - P(H_2)$ is the prior probability of a single component and $P(\boldsymbol{X}|H_i)$ is the *evidence* for H_i: the integral over the parameter density of the likelihood

$$P(\boldsymbol{X}|H_i) = \int_{R^m} d\boldsymbol{\theta}\ f(\boldsymbol{X}|\boldsymbol{\theta}, H_i) f(\boldsymbol{\theta}|H_i) \tag{14}$$

We can use the results of the appendix to compute the Bayes Factor for two against one components. In the following, unsubscripted quantities refer to the *unsplit* data, while subscripts $1, 2$ refer to the two components of the putative mixture approximation. Thus the Bayes Factor can be written from (40) as

$$BF = \frac{|\boldsymbol{A}_1|^{\frac{d}{2}}|\boldsymbol{A}_2|^{\frac{d}{2}}|\boldsymbol{A}^*|^{\frac{d^*}{2}}}{|\boldsymbol{A}_1^*|^{\frac{d_1^*}{2}}|\boldsymbol{A}_2^*|^{\frac{d_2^*}{2}}|\boldsymbol{A}|^{\frac{d}{2}}} \prod_{i=1}^{n} \frac{\Gamma(\frac{d_1^*+1-i}{2})\Gamma(\frac{d_2^*+1-i}{2})}{\Gamma(\frac{d^*+1-i}{2})\Gamma(\frac{d+1-i}{2})} \tag{15}$$

In (15), quantities with a * denote posterior and without it denote the corresponding prior parameters. Thus, $\boldsymbol{A}$ is the prior covariance parameter for a single component, while $\boldsymbol{A}_1^*$ is the posterior covariance parameter for the first of the two components. The sample size, N, is hidden in the parameters $d^* = d + N, d_i^* = d + N_i, i = 1, 2$. A considerable simplification of the above occurs if we make the following choices: $\boldsymbol{A}_i = \boldsymbol{A} = \frac{d}{d^*}\boldsymbol{A}^*$. In this case, the log-Bayes Factor reduces to

$$\begin{aligned} LBF \quad = \quad & \frac{d^* + d}{2} \ln|\boldsymbol{A}^*| - \frac{d_1^*}{2} \ln|\boldsymbol{A}_1^*| - \frac{d_2^*}{2} \ln|\boldsymbol{A}_2^*| + \frac{nd}{2} \ln\frac{d^*}{d} + \\ & \sum_{i=1}^{n} \ln\Gamma(\frac{d_1^* + 1 - i}{2}) + \sum_{i} \ln\Gamma(\frac{d_2^* + 1 - i}{2}) \\ & - \sum_{i} \ln\Gamma(\frac{d^* + 1 - i}{2}) - \sum_{i} \ln\Gamma(\frac{d + 1 - i}{2}) \end{aligned} \tag{16}$$

AIC is based on likelihoods and can be written using (38) as

$$AIC = \frac{N}{2}\ln|\boldsymbol{S}| - \frac{N_1}{2}\ln|\boldsymbol{S}_1| - \frac{N_2}{2}\ln|\boldsymbol{S}_2| - n(n+3) \tag{17}$$

Comparing (16) and (17), there is an obvious similarity in the dependence on the data, except for the use of priors in the former. Moreover, in the LBF, the 'penalty' term increases with both the dimension of the problem and the sample size, whereas the AIC penalty depends only on the dimension. Note that although it is possible to use Bayesian estimates in AIC or ML estimates in the BF, it makes little sense to do so.

Figures 1(a)-(d) show an interesting feature of the two. All four figures show contour plots of the AIC or LBF figures for a test of two $2 - D$ Gaussian components against one, as a function of separation either in scale or position (horizontal axis) against population size (vertical). The exact values are used for a population made up of two such Gaussian components, with 50% of each in the mixture. As has been noted by others, it does appear that AIC is more supportive of splitting as the population increases; on the other hand, for very small populations, it is the LBF which is more inclined to split, ie is positive. This tendency of LBF to favour splitting of small populations is not a good one and shows that neither criterion dsiplays all of the features one would like, even in such a simple case as a two-vs-one decision. Nevertheless, the LBF has clear advantages, as long as the population size is large enough to give meaningful results.

A different perspective can be given to the BF if we rewrite it using Bayes's Rule, viz.

$$LBF = \ln \frac{P(\boldsymbol{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{P(\boldsymbol{X}|\boldsymbol{\theta})} - \ln \frac{P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{X})}{P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} + \ln \frac{P(\boldsymbol{\theta}|\boldsymbol{X})}{P(\boldsymbol{\theta})} \tag{18}$$

While the first term represents the gain in likelihood from using two components instead of one, the remainder represents the cost of this, in terms of the mutual information between parameters and data (cf. Appendix I, (41)).

2.1 Effects of Motion

The other major property, a crucial one for motion analysis, is the closure of the set of $n - D$ Gaussian functions $\mathcal{G}^n$ under affine maps $\boldsymbol{A} : R^n \mapsto R^n$

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{L}\boldsymbol{x} + \boldsymbol{a} \tag{19}$$

where $\boldsymbol{L}$ is an invertible matrix and $\boldsymbol{a}$ a translation. Again, it is obvious that the action of $\boldsymbol{A}$ on $\mathcal{G}^n$ is closed, since

$$g_\Sigma(\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) = g_{\Sigma_A}(\boldsymbol{x} - \boldsymbol{\mu}_A) \tag{20}$$

where

$$\boldsymbol{\mu}_A = \boldsymbol{\mu} - \boldsymbol{a} \tag{21}$$

and

$$\boldsymbol{\Sigma}_A = \boldsymbol{L}^T \boldsymbol{\Sigma} \boldsymbol{L} \tag{22}$$

But now we are in a position to prove a rather interesting result, summarised as **Theorem 2**: Let $f(.) \geq 0$ be an integrable function $f : R^n \to R$, as above and let $\boldsymbol{t}(.) : R^n \to R^n$ be a smooth, invertible map from R^n to itself, having bounded first and second derivatives. Then for any $\delta > 0$, f has an approximation as a Gaussian mixture of the form of (2), with integral absolute error no greater than $\delta/2$ and there exists an approximation of the transformed function $\boldsymbol{T}f$

$$\boldsymbol{T}f(\boldsymbol{x}) = f(\boldsymbol{t}^{-1}(\boldsymbol{x})) \tag{23}$$

of the same form, where each Gaussian component g_{Σ_i} is transformed according to a local affine approximation of the flow field $\boldsymbol{t}$, with error no greater than δ.

Proof: The proof is based on the smoothness of the mapping and the local nature of the MGMM approximation. First, we observe that *any* smooth flow can be approximated using

$$\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{t}\boldsymbol{x}_0 + \nabla_x(\boldsymbol{t})(\boldsymbol{x} - \boldsymbol{x}_0) + O(\|\boldsymbol{x}\|^2) \tag{24}$$

via Taylor's theorem, where $\nabla_x(.)$ is the gradient. Now, let $g_\Sigma(\boldsymbol{x} - \boldsymbol{\mu})$ be an arbitrary Gaussian function on R^n. Then it follows from (24) that there exists a scale parameter $\sigma(\delta) > 0$ such that

$$\int d\boldsymbol{x} |g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{x} - \boldsymbol{\mu}) - g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{\mu}) + \nabla_x(\boldsymbol{t})(\boldsymbol{x} - \boldsymbol{\mu}))| < \frac{\delta}{2} \tag{25}$$

We can achieve this by choosing $\sigma(\delta) > 0$ so that the error $\epsilon(\boldsymbol{x})$ - the integrand in (25) - satisfies

1. $\epsilon(\boldsymbol{x}) < \frac{\delta}{4}$, if $\|\boldsymbol{x} - \boldsymbol{t}(\boldsymbol{\mu})\| < r$. This follows because the error in the Taylor expansion for $\boldsymbol{t}$ is of order $\sigma^2(\delta)$ if $r = M\sigma(\delta)$.

2. $\int_{\|\boldsymbol{x}\|>r} d\boldsymbol{x}|\epsilon(\boldsymbol{x})| < \int_{\|\boldsymbol{x}\|>r} g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{x} - \boldsymbol{\mu}) + g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{\mu}) + \nabla_x(\boldsymbol{t})(\boldsymbol{x} - \boldsymbol{\mu})) \leq \frac{\delta}{4}$. This follows from the observation that the tails of the two Gaussians can be made of arbitrarily low weight by a similar choice of r.

Putting the two errors together, we get a total error

$$\int d\boldsymbol{x} |g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{x} - \boldsymbol{\mu}) - g_{\sigma(\delta)\Sigma}(\boldsymbol{t}(\boldsymbol{\mu}) + \nabla_x(\boldsymbol{t})(\boldsymbol{x} - \boldsymbol{\mu}))| < \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2} \tag{26}$$

Since this applies to an arbitrary Gaussian, apply it to each component in a mixture representation of $f(\boldsymbol{t}(\boldsymbol{x}))$

$$f(\boldsymbol{t}(\boldsymbol{x})) = \sum_i f_i g_{\Sigma_i}(\boldsymbol{x} - \boldsymbol{\mu}_i) \tag{27}$$

for which (i) the absolute integral error is less than $\delta/2$ and (ii) each component has a variance Σ small enough that its transformation by $\boldsymbol{t}$ introduces an error no bigger than $\delta/2$. Such a representation exists, by Theorem 1. But then the error in the representation of $f(\boldsymbol{x})$ by the affine approximation of $\boldsymbol{t}$ is no greater than δ. This completes the proof.•

3 MGMM

The key to applying these ideas is to use a sequential approach, which leads to a multiresolution tree structure: Multiresolution Gaussian Mixture Modelling (MGMM). We start from the natural assumption that we do not know the density of interest, but wish to estimate it from some data, such as an image or set of images. If these data are denoted $\boldsymbol{X}_i, 1 \leq i \leq n$, we can compute the sample mean and covariance [21] and thence infer a single multivariate Gaussian model $g_{\Sigma_0}(\boldsymbol{x} - \boldsymbol{\mu}_0)$, where $\boldsymbol{\mu}_0, \Sigma_0$ are the sample (Maximum Likelihood) estimates for the data. Now, in rare cases this will model the data adequately. If not, then we split the data into two parts and model each with a Gaussian. This can be done using the Markov Chain Monte Carlo (MCMC) sampling technique described in [19], which treats the inference as one containing *hidden variables*, namely the class Z_i to which each datum belongs and samples from the posteriors for the population size, means and covariances, assuming *conjugate* priors, whose parameters are simply those of the population as a whole. Thus the prior for the means of the two classes are Gaussian, while the covariances are drawn from an inverse Wishart density: a Normal-Inverse Wishart (NIW) density. Because the NIW is the natural conjugate prior for this problem, the posterior is also NIW and so a Gibbs sampler can be used [19, 14]. Sampling for (i) the hidden variables and (ii) the corresponding sub-population densities gives the two Gaussians, based on the posterior estimates from the sampler for $g_{\Sigma_{1j}}(\boldsymbol{x} - \boldsymbol{\mu}_{1j}), i \in [0, 1]$. This Gibbs sampler is described fully in Appendix II. Clearly, we have the basis of a recursion here: we can build a tree representation, in which each leaf is a subset of the population found by successive binary splits from the root node, which represents the whole population. Because the NIW is a conjugate prior, the same computations can be used at every node in the tree:

1. Select class j and test its normal density approximation for 'goodness-of-fit'. If the fit is adequate, terminate, *else*

2. Split class j into two components: class $j1$ and class $j2$, by sampling the hidden variables $Z_{ji}, 1 \leq i \leq N_j$ and thence obtaining a Bayesian estimate of the class means and covariances by sampling from the posteriors, given the prior $g_{\Sigma_j}(\boldsymbol{x} - \boldsymbol{\mu}_j)$.

This needs a criterion for splitting. In the previous section, we reviewed two popular choices: LBF and AIC, either of which can be expressed in the present case as

$$C = \ln \frac{f(\boldsymbol{X}|\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2})}{f(\boldsymbol{X}|\boldsymbol{\theta}_j)} - k(N_j, n) \tag{28}$$

where the log-likelihood ratio is subject to a 'complexity' penalty which is a function of the size of the population at j and the dimensions of the space. In the context of the MGMM tree, however, the LBF given in (18) does not have the right form: it is based on the comparison of *either* a single component or two components as a model of the data. Any node in the tree, other than the root, is the result of a series of dichotomies and may itself be split: splitting a node *adds* complexity to the existing description, rather than acting as a *de novo* alternative. Moreover, the parameter estimates at node j are the obvious hyperparameters for the two resulting classes. Indeed, if we wish to store or transmit the parameters $\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}$, we would need the values of $\boldsymbol{\theta}_j$ also, since these appear in the prior. Finally, in keeping with its roots in rate-distortion theory [2], we add a *rate control* parameter λ, which allows us to trade-off the information in a given MGMM tree against its accuracy. If we take this into account, we replace (18) by the Minimum Information Criterion (MIC):

$$MIC(\lambda) = \ln \frac{f(\boldsymbol{X}|\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2})}{f(\boldsymbol{X}|\boldsymbol{\theta}_j)} - \lambda \ln \frac{f(\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}|\boldsymbol{X}, \boldsymbol{\theta}_j)}{f(\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}|\boldsymbol{\theta}_j)} \tag{29}$$

Another way of viewing the constant λ is that it accounts for the *prior* $P(H_1)$ in (13). If we do not wish to take this communication-oriented approach, then setting $\lambda = 1$ amounts to setting $P(H_1) = 0.5$, ie using the LBF. Although the aim of MIC might sound similar to Rissanen's MDL, two observations might be made about this: (i) the aim here is to provide a trade-off between the information cost of the parameters and the information gained by providing a better fit to the data; (ii) in the NIW case, these criteria have such closely related forms that they all amount to some form of Bayes Evidence, leaving space for one's prejudice in the definiton of the prior.

We conclude by observing that an alternative view of the MGMM description is as a patchwork of affine models, each leaf node being the result of a linear regression on the data [9]. For example, in the case of a grey level image, the MGMM description gives for each class a Gaussian model, which is directly related to a least-squares approximation of the form

$$z_i(\boldsymbol{x}) = \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{x}_i) + z_{i0} + \nu_i(\boldsymbol{x}) \tag{30}$$

where $z_i(.)$ is the grey level as a function of the spatial coordinate $\boldsymbol{x}$ for the ith class and $\nu_i(.)$ is the *residual*. The matrix $\boldsymbol{A}_i$ is easily found from the covariance matrix Σ_i for that class and $\boldsymbol{x}_i, z_{i0}$ from the mean.

4 Experiments

The first set of experiments is designed to test the sampler on data taken from a known density: a pair of $2-D$ normals, with varying distance between means, covariance and populations:

1. 40 points drawn from a normal $\boldsymbol{\mu}^T = (0,0), \boldsymbol{\Sigma} = \boldsymbol{I}$ density and 40 from a normal $\boldsymbol{\mu}^T = (4,0), \boldsymbol{\Sigma} = \boldsymbol{I}$ density.

2. First component as in first case and second shifted to $\boldsymbol{\mu}^T = (2,0)$.

3. First component as above and second shifted to $\boldsymbol{\mu}^T = (0,0)$, with $\boldsymbol{\Sigma} = 16\boldsymbol{I}$.

4. First component as above and second shifted to $\boldsymbol{\mu}^T = (0,0)$, with $\boldsymbol{\Sigma} = 4\boldsymbol{I}$.

In addition, subsets consisting of $(40,20)$, $(40,10)$ and $(20,20)$ data from the two samples respectively were used, giving 12 combinations in all. The data in each case were obtained from the same set, by simply translating and scaling the second set appropriately. This was done to ensure that the variations seen in results were caused by the sampling algorithm. The sampler used the sample mean and covariance of the whole population as prior parameters for the NIW distribution, with the hyperparameters set to their minimum values, ie. $c = 1, d = n+2$ in (31). The error rates for the hidden variable classification taken from 10 independent runs with different simulation seeds are shown in Table 4. The classification used the sample mean from the last 100 iterations of the sampler. Investigations showed that the sampler had settled into a stationary distribution by then. Although this is a small number of steps, these problems are comparatively well posed, especially in the first two cases. Correspondingly, the error rates for these two cases are small and the estimated parameters are close to their true values in every simulation. Although high error rates are observed for the concentric cases, it should be obvious that in such cases, classification of the central points in either case is essentially arbitrary. It was noted that in virtually every case, the sampler setttled in a state corresponding to two components of differing variance, located approximately concentrically. Figures 2(a)-(d) show typical results from the simulation.

Note that in these examples and in those that follow, to prevent singularities and to reflect the finite precision of the image measurements, each datum consists of both a location and a variance parameter. Correspondingly, the sample covariance has to be modified to take account of this.

The method has also been tested on some more realistic applications in image analysis. In all the cases described below, the MGMM algorithm was limited to 50 iterations and typically converged in $10-20$. The Bayes Factor was used as the criterion for splitting a component. The first results show the segmentation of a well known image - Lena - using the MGMM approach. In this case, the data are $3-d$: two spatial dimensions and one for intensity. The MGMM tree for this 256×256 pixel image, produced using the Gibbs Sampler, are shown schematically in figure 5 and are superposed on the corresponding least squares approximations in figure 4, showing the improvement in the reconstruction as the number of leaves increases. In

Population	Case 1	Case 2	Case 3	Case 4
40,40	0.0213	0.212	0.341	0.411
40,20	0.050	0.138	0.273	0.395
40,10	0.018	0.120	0.222	0.378
20,20	0.073	0.20	0.338	0.448

Table 1: Missclassification rates for samples of sizes shown in leftmost column from two normal densities as listed in text. The classification used the sample means over the last 100 of 200 iterations of the sampler.

fig. 5, the sample mean vector, in the order (x, y, z) and the population probability are shown next to each leaf vertex in the tree. Figures 5(a)-(d) show the representation of the image by 8 regions, each corresponding to a leaf of the MGMM tree. The classification 5(a) is just the MAP classification of pixels based on all 3 coordinates (spatial and intensity), while fig. 5(b) shows the MGMM tree nodes and the ellipses corresponding to the spatial covariance of the Gaussian component at each leaf node. The lines indicate the tree structure, with thickness indicating height in the tree: the thickest lines are those from the root; the thinnest those to the lowest level in this tree - level 4, numbering from 0 at the root. In other words, the tree is not especially balanced, nor should it be, if it is properly adapted to the data. The two reconstructions use the least-squares approximation based on the estimated mean and covariance associated with that class at that pixel, as in (30). The left image, (c), shows the reconstruction using only spatial information, using a 'soft' decision: each pixel is treated as a mixture, with weights given by the relative magnitudes of the 8 Gaussian components at that position. In effect, this is making an inference from $2-D$ to $3-D$ using the MGMM representation, since grey level is ignored in classifying the pixels. The reconstruction signal-noise ratio for this case is $15.1dB$ (peak-rms). Although this is a poor reconstruction visually, it represents an extreme paucity of information: only 8 sets of $3-D$ Gaussian parameters are required, ie. 72 values. On the other hand, when a full MAP classification is performed using all 3 co-ordinates, the reconstruction SNR climbs to $24.3dB$ and indeed the visual appearance is surprisingly good. Figure 6(a) shows the approximation of the probablility density based on the MGMM and extracting the marginal density for the intensity on level 3 of a gray level pyramid, which has just 1024 pixels. This is a trivial computation from the MGMM representation, which gives a better approximation of the density than can be obtained by simple histogramming. For comparison, the histogram of the 256×256 pixel image is shown in figure 6(b).

The second illustration uses two frames of the 'Miss America' sequence in the segmentation, which is based on the $5-d$ data consisting of spatial co-ordinates, intensity and $2-d$ motion. The raw motion estimates were obtained using a multiresolution correlation algorithm, again from 256×256 pixel images. In this example, a tree description with 66 leaves was obtained from frame 15 of the sequence and the $2-d$ motions. The affine motions extracted by least squares (cf. (30) from the covariance matrix for these data were then applied to the spatial co-ordinates of each blob and the result used to reconstruct frame 16 of the sequence. The reconstructions of frame 15-16, figures 7(a)-(d) show that the MGMM approach can deal easily with non-rigid motions, the peak-rms signal-noise ratio (PSNR) being $20.0dB$ for this reconstruction, as it is for the original frame.

As a final example, the two images of a breakfast table in Fig. 8(a)-(b) were used in the same multiresolution cross-correlation algorithm to compute a stereoscopic disparity map. This was input to the MGMM tree algorithm as $4-D$ data: spatial co-ordinates, grey level and horizontal disparity. The original images were also 256×256 pixels, but 64×64 images were input to the MGMM algorithm. Reconstruction of the disparity also used all 4 dimensions in classifying each pixel, the depths of the classified pixels being estimated using the least squares approximation based on their spatial position, as in (30). The resulting approximation is shown in figure 8(c) and is superimposed on the left image in figure 8(d), showing a reasonable correlation with the expected heights of the larger objects, such as milk and cereal packets. There were 150 leaves in the MGMM tree for this example, due to the complexity of the scene, giving a reconstruction SNR of $18.8dB$.

5 Conclusions

In this paper, a novel and versatile statistical image representation has been presented. The theory of MGMM as a general form of density approximation was outlined and its use in describing images and sequences illustrated. The key properties which mark MGMM out as a representation are as follows: although it is a statistical model, it incorporates spatial relationships; it is 'auto-scaling', ie. it classifies data based on their likelihood, rather than on simple Euclidean distance; because it is multiresolution, it allows efficient computation; because of the closure of the Gaussian functions under affine motions, it deals directly with the problem of image motion. This combination of properties makes MGMM a unique approach to image modelling.

It should be clear that the tree structure of MGMM implies that as an approximation of an arbitrary density with a *fixed* number of components, it is most likely not optimal: tree searches have their limitations. But this is not the real point of MGMM: what MGMM offers is a way of approximating an *arbitrary* distribution to an *arbitrary* precision in a computationally efficient way. In this respect too it is unique.

The experiments on synthetic data show that a sampling approach to estimation of the Gaussian parameters at each stage can be computationally effective, while the results on real image data illustrate that accurate reconstructions are possible with very compact MGMM representations - even in the most complex of the above examples, only 125 classes were needed for 256×256 image data. All of the examples show that MAP estimation combined with MGMM is an effective tool in image analysis and vision.

Of course, this report represents preliminary work, which requires further development, both theoretical and practical. For example, the work on motion needs to be extended to take proper account of temporal structure and this in turn requires a proper theoretical basis. It is clear that none of the selection criteria is ideal in all cases. Finally, the essential problem in vision of moving from a $2-D$ representation to a $3-D$ one has not been tackled yet. These are all under active investigation at the time of this writing.

Acknowledgement

This work was supported by the UK EPSRC.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1965.

[2] T. Berger. *Rate-Distortion Theory: a Mathematical Basis for Data Compression.* Prentice-Hall, Englewood Cliffs, 1971.

[3] P. J. Burt and E. H. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Commun.*, COM(31):532–540, 1983.

[4] I. Daubechies. Orthogonal Bases of Compactly Supported Wavelets. *Comm. on Pure and Appl. Math.*, XLI:909–996, 1988.

[5] I. Daubechies. The Wavelet Transform, Time-Frequency Localisation and Signal Analysis. *IEEE Trans. Information Theory*, 36:961–1005, 1990.

[6] E. H. Adelson E. P. Simoncelli, W. T. Freeman and D. J. Heeger. Shiftable Multiscale Transforms. *IEEE Trans. Information Theory*, 38:587–607, 1992.

[7] A. K. Jain. *Fundamentals of Digital Image Processing.* Englewood Cliffs, New Jersey, 1989.

[8] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis.* Dover, 1975.

[9] R. A. Ohlsen L. Breiman, J. H. Friedman and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.

[10] L. J. Ljung. *System Identification: Theory for the User.* Englewood Cliffs, Prentice-Hall, 1987.

[11] S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Patt. Anal. Machine Intell.*, 11(7):674–693, July 1989.

[12] S. G. Mallat. Multifrequency Channel Decompositions of Images and Wavelet Models. *IEEE Trans. Acous. Speech Sig. Proc.*, 37(12):2091–2110, December 1989.

[13] G. J. McLachlan. *Mixture Models: Inference and Applications to Inference and Clustering.* New York, M. Dekker, 1988.

[14] A. O'Hagan. *Kendall's Advanced Theory of Statistics, vol. 2B.* London, Edward Arnold, 1994.

[15] J. J. K. O'Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* New York, Springer-Verlag, 1996.

[16] S. J. Press. *Applied Multivariate Analysis.* Florida, Robert E. Krieger, 1982.

[17] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge U.P., 1996.

[18] J. Rissanen. Stochastic Complexity. *Jnl. Royal Stat. Soc., B*, 49:223–239, 1987.

[19] C. P. Robert. Mixtures of Distributions: Inference and Estimation. In S. Richardson W. R. Gilks and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

[20] I. Rezek S. J. Roberts, D. Husmeier and W. Perry. Bayesian Approaches to Gaussian Mixture Modelling. *IEEE Trans. Patt. Anal. Machine Intell.*, 20(11):1133–1141, 1998.

[21] G. A. F. Seber. *Multivariate Observations.* New York, Wiley, 1984.

[22] R. Szeliski. *Bayesian Modeling of Uncertainty in Low-level Vision.* Boston, Kluwer, 1989.

[23] R. A. Tapia and J. R. Thompson. *Non-parametric Probability Density Estimation.* Baltimore, Johns Hopkins Pr., 1978.

[24] G. S. Watson and M. R. Leadbetter. On the Estimation of the Probability Density. In *Ann. Math. Stat.*, volume 34, pages 480–491, 1962.

[25] R. Wilson, A. Calway, and E.R.S. Pearson. A Generalized Wavelet Transform for Fourier Analysis: The Multiresolution Fourier Transform and Its Application to Image and Audio Signal Analysis. *IEEE Trans. Information Theory*, 38(2):674–690, March 1992.

[26] A. Witkin. Scale-Space Filtering. In *Proc. of IEEE ICASSP-84*, 1984.

[27] K. Palaniappan X. Zhuang, Y. Huang and Y. Zhao. Gaussian Mixture Density Modeling, Decomposition and Applications. *IEEEIP*, 5:1293–1301, 1996.

Appendix I: Analysis of the Multivariate Normal Distribution

In using the MVN distribution as a *universal approximator*, it is possible to make use of the properties of this widely studied distribution. The development here follows that in [14] and [16]. A Bayesian analysis starts from the choice of priors and in the present application, it is obvious to use the so called *natural conjugate prior* distribution: the Normal-Inverse-Wishart (NIW) distribution, for multivariate normal data with unknown mean and covariance. This is defined for $n-D$ data as

$$f(\boldsymbol{\mu},\boldsymbol{\Sigma}) = k^{-1}\ |\boldsymbol{A}|^{\frac{d}{2}}|\ |\boldsymbol{\Sigma}|^{-\frac{d+n+2}{2}} \exp[-\frac{c}{2}(\boldsymbol{\mu}-\boldsymbol{a})^T\boldsymbol{\Sigma}^{-1}((\boldsymbol{\mu}-\boldsymbol{a}) - \frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{A})] \quad (31)$$

or, briefly $f(\boldsymbol{\mu},\boldsymbol{\Sigma}) = NIW(\boldsymbol{A}, d, \boldsymbol{a}, c)$, a function of the parameters $d > n+1$, the *degrees of freedom*, c, the prior weight on the mean, $\boldsymbol{a}$, the prior mean and $\boldsymbol{A}$, the prior covariance. Increasing c and d attaches relatively more weight to the prior, for a given sample size N. The normalising constant k is given by

$$k = 2^{\frac{n(d+1)}{2}}\pi^{\frac{n(n+1)}{4}}\prod_{i=1}^{n}\Gamma(\frac{d+1-i}{2}) \quad (32)$$

The likelihood for the data $\boldsymbol{X}_i, 1 \leq i \leq N$, is then simply a product of normals, assuming that data are independently sampled, ie.

$$f(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{N}{2}}|\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp[-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{X}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})] \quad (33)$$

and because the NIW is the natural conjugate prior, the posterior is also NIW, with parameters

$$\boldsymbol{a}^* = \frac{c\boldsymbol{a} + N\bar{\boldsymbol{X}}}{c+N} \quad (34)$$

where $\bar{\boldsymbol{X}} = N^{-1}\sum_i\ \boldsymbol{X}_i$ is the sample mean and

$$\boldsymbol{A}^* = \boldsymbol{A} + N\boldsymbol{S} + \frac{cN}{c+N}(\boldsymbol{a}-\bar{\boldsymbol{X}})(\boldsymbol{a}-\bar{\boldsymbol{X}})^T \quad (35)$$

where $\boldsymbol{S} = N^{-1}\sum_i(\boldsymbol{X}_i-\bar{\boldsymbol{X}})((\boldsymbol{X}_i-\bar{\boldsymbol{X}})^T$ is the sample covariance. It is not hard to see that in the posterior, the scalar parameters are $c^* = c+N, d^* = d+N$.

The classical analysis of MVN data is based on the likelihood (33), from which it is easy to find the ML estimates

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \quad (36)$$

and

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} \quad (37)$$

Evaluating the likelihood at these estimates gives

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} f(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{S}|^{-\frac{N}{2}}\exp[-\frac{Nn}{2}] \quad (38)$$

This may be compared with the likelihood evaluated at the MAP estimates $\tilde{\boldsymbol{\mu}} = \boldsymbol{a}^*, \tilde{\boldsymbol{\Sigma}} = \frac{1}{d^*}\boldsymbol{A}^*$

$$f(\boldsymbol{X}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) = |2\pi \boldsymbol{A^*}|^{-\frac{N}{2}} {d^*}^{\frac{nN}{2}} \exp[-\frac{d^*}{2}(n - tr\boldsymbol{A}^{*-1}\boldsymbol{A})] \tag{39}$$

The above results can be used to evaluate the Bayes evidence for one component in the mixture, by integrating over the parameters

$$f(Y) = \frac{|\boldsymbol{A}|^{\frac{d}{2}}}{\pi^{\frac{nN}{2}}|\boldsymbol{A}^*|^{\frac{d^*}{2}}} \prod_{i=1}^{n} \frac{\Gamma(\frac{d^*+1-i}{2})}{\Gamma(\frac{d+1-i}{2})} \tag{40}$$

The last quantity of interest is the mutual information between the data and parameters, which is given by the ratio

$$i(\boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \frac{f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Y})}{f(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \tag{41}$$

The expectation of this quantity represents the amount of information which, on average, the parameters yield about the data, or vice versa. Since both prior and posterior are NIW, it is straightforward to evaluate (41)

$$\begin{aligned} i(\boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad & \frac{d^*}{2}\ln|\boldsymbol{A}^*| - \frac{d}{2}\ln|\boldsymbol{A}| - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{nN}{2}\ln 2 + \\ & \sum_i \ln\Gamma(\frac{d^*+1-i}{2}) - \sum_i \ln\Gamma(\frac{d+1-i}{2}) - \\ & \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}(c^*(\boldsymbol{\mu}-\boldsymbol{a}^*)(\boldsymbol{\mu}-\boldsymbol{a}^*)^T - c(\boldsymbol{\mu}-\boldsymbol{a})(\boldsymbol{\mu}-\boldsymbol{a})^T + \boldsymbol{A}^* - \boldsymbol{A})] \end{aligned} \tag{42}$$

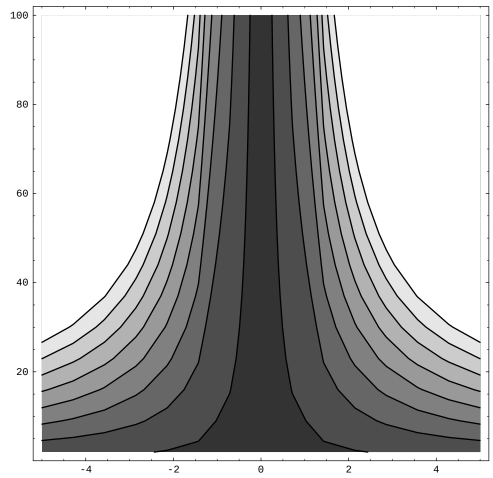
Appendix II: a Gibbs Sampler for NIW variates with hidden variables

This sampler follows exactly the model proposed by Robert [19]. It uses NIW as the prior for the sample data and a Dirichlet prior for the population size (experience has shown that this is not always helpful, however). Thus, given the estimate $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ for the mean and covariance at a node j, the sampler performs the following steps:

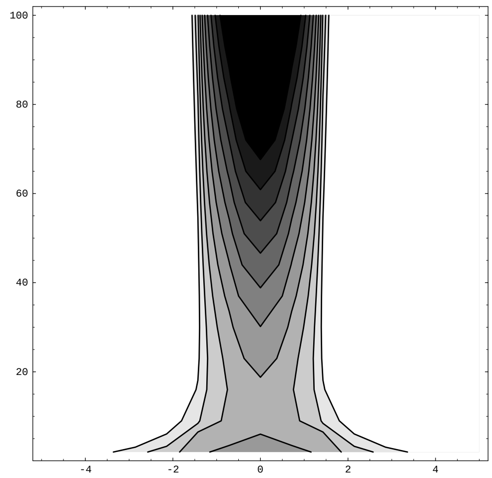
1. Sample $\boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}, i = 1, 2, \sim \mathrm{NIW}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, c_i^*, d_i^*)$.
2. Sample for the population size using the Dirichlet distribution.
3. Sample the hidden variables $Z_k \in [1, 2], 1 \leq k \leq N_j$ using a Gibbs sampler and the current estimates of $P, \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}, i = 1, 2$.
4. Calculate the posterior NIW parameters, based on the classified data and (34)-(35).

This is repeated, starting with $c_i^* = c_i = 1, d_i^* = d_i = n + 2$, the minimum values, to keep the prior vague.

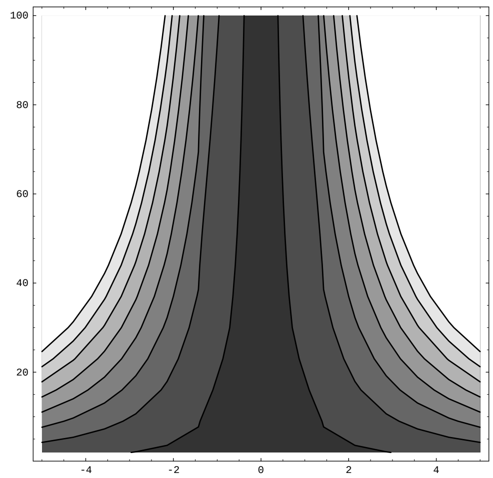
It may be noted here that a deterministic algorithm, based on $2 - means$, but using the log-likelihood rather than simple Euclidean distance, works adequately on some simpler problems, for example where classes are convex. In general, however, it is little cheaper to implement $2 - means$ and much more likely to settle in a local minimum.



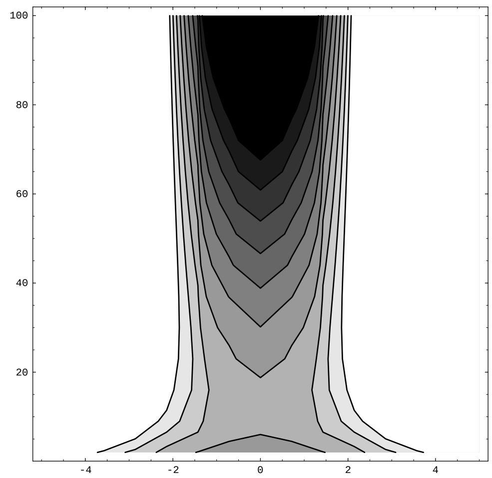
(a)



(b)

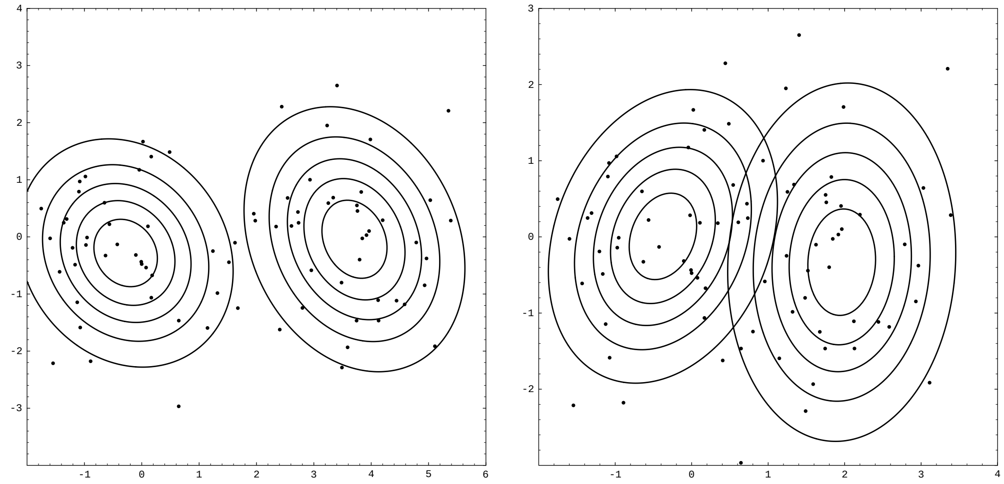


(c)

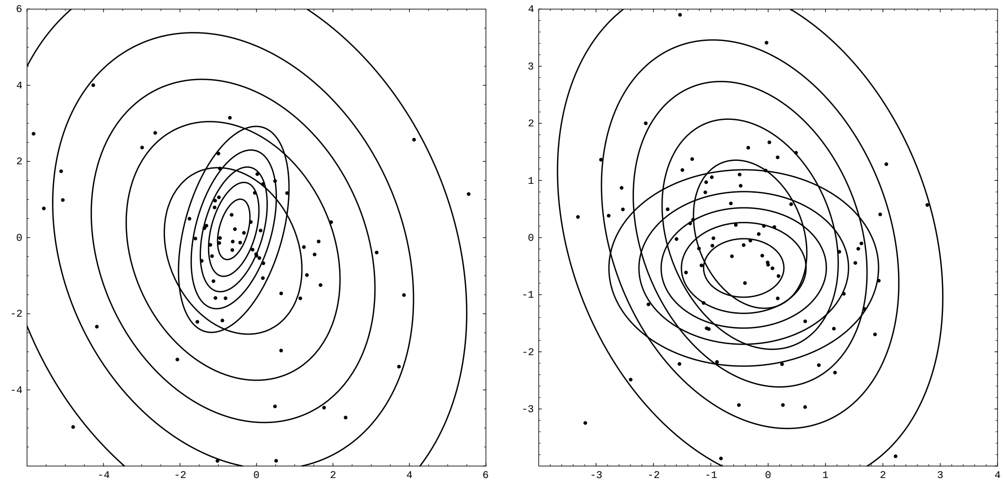


(d)

Figure 1: Contour plots of Akaike Information Criterion and Log-Bayes Factors for two vs one normal components: (a) AIC as a function of difference in means (horizontal) and population size (vertical), (b) LBF as in (a), (c) AIC as a function of difference in scale, (d) LBF as in (c). In all four cases, the proportions of the two components are equal and black represents negative values, ie a preference for a single component model.



(a) (b)



(c) (d)

Figure 2: Contour plot of estimated Gaussian components from a pair of 2-d normals, (a) having means $\boldsymbol{\mu}_1^T = (0,0), \boldsymbol{\mu}_2^T = (4,0)$ and covariances $\boldsymbol{\Sigma}_i = \boldsymbol{I}, i = 1,2$; (b) means at $\boldsymbol{\mu}_1^T = (0,0), \boldsymbol{\mu}_2^T = (2,0)$; (c) means at $\boldsymbol{\mu}_i^T = (0,0), i = 1,2$ and $\boldsymbol{\Sigma}_1 = 16\boldsymbol{\Sigma}_2$; (d) $\boldsymbol{\Sigma}_1 = 4\boldsymbol{\Sigma}_2$. Estimates obtained from 200 iterations of the Gibbs sampler. The data points - 40 from each component - are superimposed on the Gaussian plots.

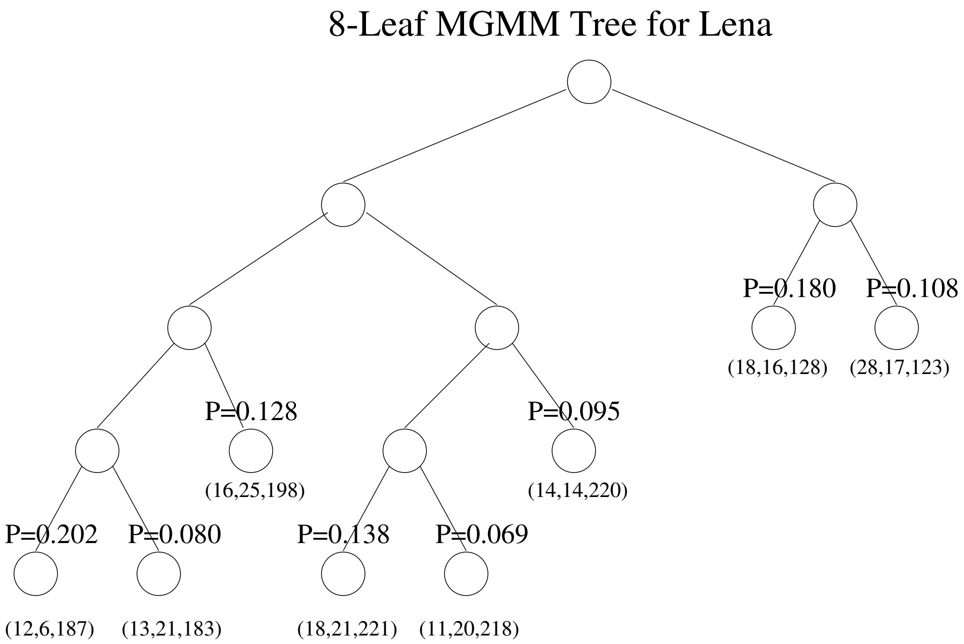


Figure 3: Schematic of the 8 leaf MGMM representation of the 256×256 'Lena' image. Population probabilities and mean vectors, in the order (x, y, z), shown for each leaf.

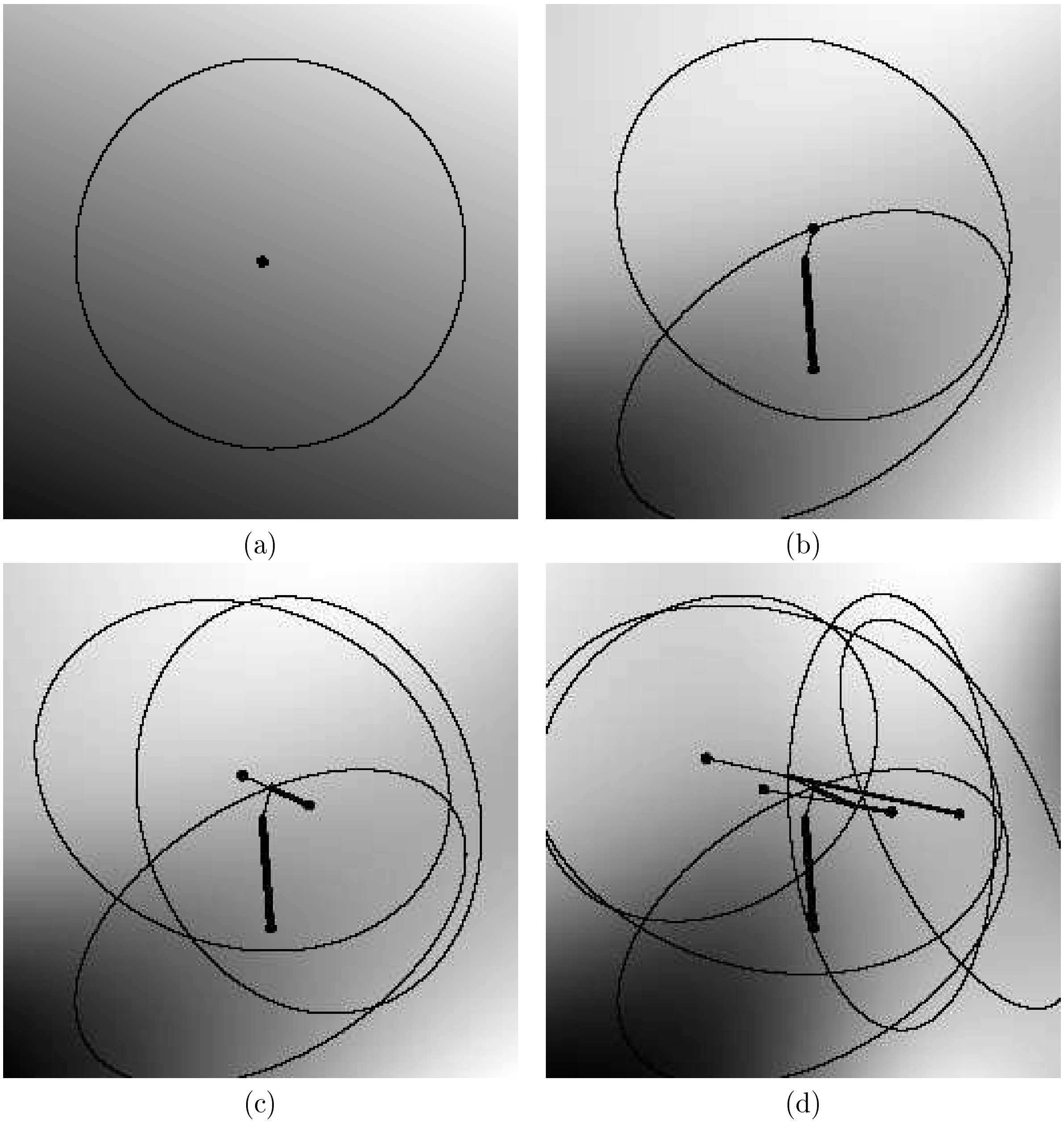
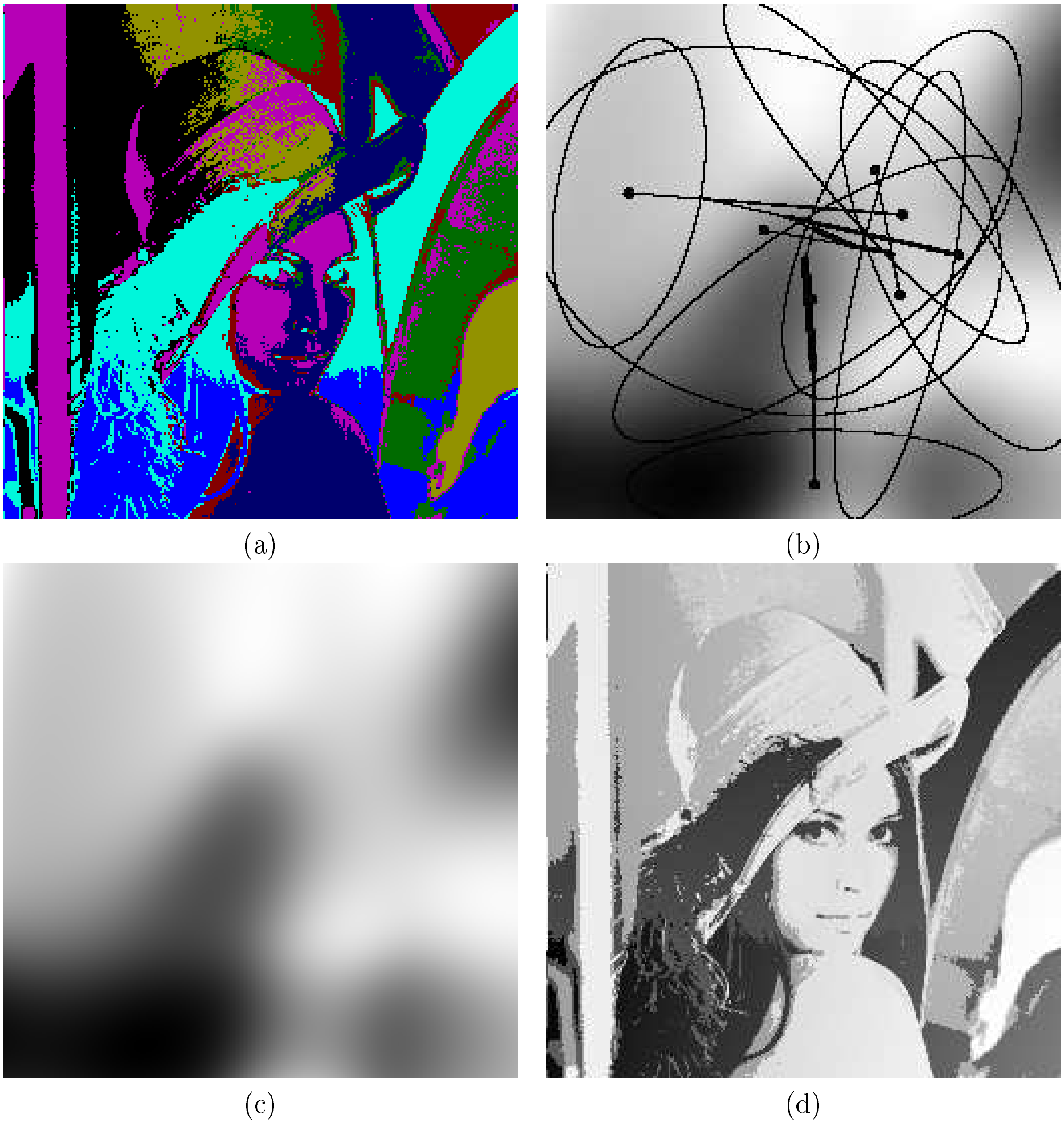
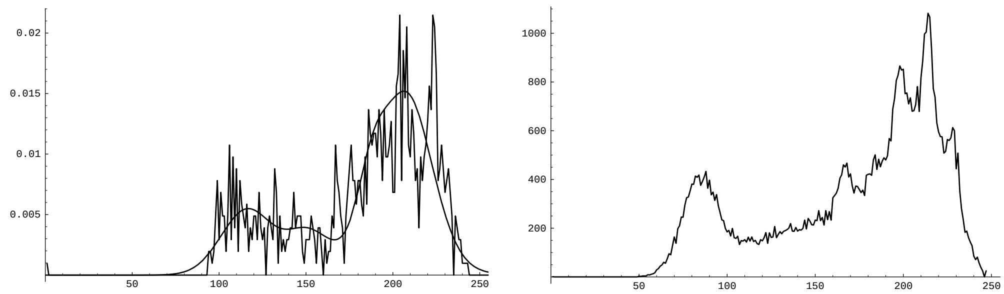


Figure 4: (a)-(d) First four levels of the MGMM tree for 256×256 'Lena', superposed on the corresponding least squares approximations. Line thickness indicates tree level and the ellipses show the covariance of the Gaussian component at each node. Note, however, that all leaves appear at the same level in this representation.



(a) (b) (c) (d)

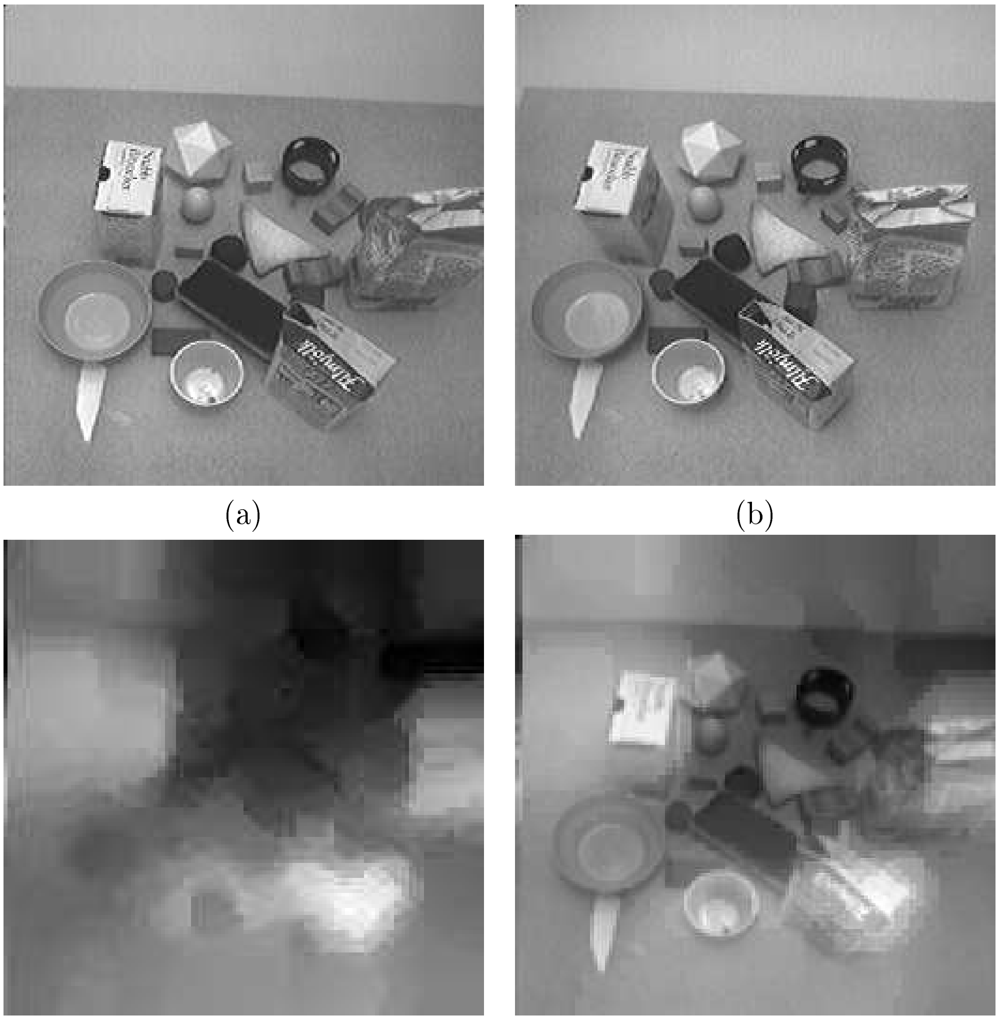
Figure 5: (a) Classification of Lena image from MGMM representation with 8 leaves. (b) MGMM tree corresponding to (a), as in previous figure; (c) Reconstruction using only spatial co-ordinates. (d) Reconstruction using grey level and spatial co-ordinates.



(a) (b)

Figure 6: (a) Comparison of histogram with MGMM approximation to probability density for level 3 of a gray level pyramid.(b) Histogram from original image.

Figure 7: (a) Frame 15 of Miss America sequence and (b) reconstruction from MGMM tree using 66 classes.(c) Original frame 16 and (d) reconstruction based on moved MGMM from frame 16.



(a) (b) (c) (d)

Figure 8: (a) Left image of stereo pair, (b) right image. (c) MGMM approximation of disparity image from (a) and (b). (d) MGMM approximation superposed on left image.