Bayesian Inference for Multiband Image Segmentation via Model-Based Cluster Trees

Fionn Murtagh [*], Adrian E. Raftery, Jean-Luc Starck

Abstract

We consider the problem of multiband image clustering and segmentation. We propose a new methodology for doing this, called model-based cluster trees. This is grounded in model-based clustering, which bases inference on finite mixture models estimated by maximum likelihood using the EM algorithm, and automatically chooses the number of clusters by Bayesian model selection, approximated using BIC, the Bayesian Information Criterion. For segmentation, model-based clustering is based on a Markov spatial dependence model. In the Markov model case, the Bayesian model selection criterion takes account of spatial neighborhood information, and is termed PLIC, the Pseudolikelihood Information Criterion. We build a cluster tree by first segmenting an image band, then using the second band to cluster each of the level 1 clusters, and continuing if required for further bands. The tree is pruned automatically as part of the algorithm by using Bayesian model selection to choose the number of clusters at each stage. An efficient algorithm for implementing the methodology is proposed. An example is used to evaluate this new approach, and the advantages and disadvantages of alternative approaches to multiband segmentation and clustering are discussed.

[*] Corresponding author: F. Murtagh. F. Murtagh is with the Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England (e-mail fmurtagh@acm.org). A.E. Raftery is with the Department of Statistics, University of Washington, Box 354322, Seattle, WA 98915-4322 (e-mail raftery@stat.washington.edu). J.-L. Starck is with SEI-SAP/DAPNIA, CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France (e-mail jstarck@cea.fr).

Keywords: Bayesian model, Markov model, Potts, Ising, information criterion, quantization, clustering, segmentation, multiband, multispectral, hyperspectral, multichannel, information fusion

1 Introduction

Clustering and segmentation in an image analysis context have a long history [19]. Objectives include: quantization of data values for later use with a codebook in a compression context; targeting delivery to display devices supporting small, bounded pixel data value depth; as a preliminary to object and feature detection and analysis in images; and as a basis for other image processing operations such as image registration and archiving.

We will use the terms clustering or quantization to refer to determining clusters among image grayscale or pixel values. In the case of multiband images, the grayscale pixel values are multidimensional. This simply implies that in multiband data clustering we are dealing with clustering in multidimensional space, i.e., we are dealing with a form of vector quantization. Multiband images include the case of color images, with bands associated with red, green and blue colors, or a large number of alternative color formating schemes. As opposed to clustering or quantization, the term segmentation is used when neighborhood or spatial influence information is incorporated into the modeling. Ideally we could impose as a *necessary* objective that all segments be spatially contiguous. In practice we take this as a *sufficient* objective. Multiband images are also referred to as multispectral or multichannel or hyperspectral images.

In this paper, we propose a new method for multiband image clustering, called *model-based cluster trees.* This combines maximum likelihood estimation of finite mixture models with Bayesian model selection. For segmentation, a Markov neighborhood dependency model is used to include adjacency or local influence. The model-based clustering tree algorithm operates recursively on the image bands. First it clusters or segments the pixels on the basis of the first band. Then, using the second selected band, it clusters each of the clusters found in the first stage. Bayesian model selection is used at each stage to determine the number of clusters or segments, so that the data are used to decide adaptively the extent to which the tree is pruned.

The resulting method allows the number of quantization levels or numbers of segments to be chosen on the basis of the data. If the number of quantization levels is predetermined (see e.g. [23]) the method can easily handle this as a special case. Given that image bands are processed in sequence, it is helpful if the image bands have some inherent order. In chromaticity/luminosity color space, such an order can make use of the fact that chromaticities convey far less perceptual information than does the luminosity (see, e.g., [32]). Such an order can be readily accommodated in our approach. In more general cases, we impose an order on image bands which will be helpful for interpretation or further processing of the clustered or segmented output.

We can readily accommodate noise in our image data. This is implied by image features taken as realizations of distributional models. Explicit noise components are incorporated into our modeling as discussed in earlier work of ours [4]. Our MR software package [17] provides multiband image noise filtering, together with compression, functionality. See also chapter 6, "Multichannel data", in [30].

We can accommodate a very small number of classes (clusters or segments) for the pixels, or a large number. A small number of classes may be needed as a preliminary to a data interpretation, or high-level vision stage of the analysis. A large number of classes may be needed when high fidelity to the original image is required.

A major motivation for a cluster tree results from use of model-based clustering in cases like multiband segmentation in Earth observation [22]. Notwithstanding the Occam razor parsimony principle of a small number of clusters, it may be found that a larger number of clusters does greater justice to the data. Then, however, it may be necessary to further analyze the clusters found. A cluster tree approach is an appropriate way to do this.

The simple tree structure given by a quadtree can be valuable, in particular for permitting Markov modeling both spatially and in scale [7]. However two problems arise with such a simple tree structure: firstly, there is a sharp discontinuity at the boundaries between quadtree cells; and secondly the quadtree is quite a crude data-driven structure.

A further motivation for our cluster tree approach is that model-based Gaussian fitting of arbitrary multiband data is often unstable and algorithmically non-robust. The reason for this is singularity brought about by the following: (i) individual clusters or segments that are of small cardinality; (ii) correlation, possibly local, between bands; and (iii) relatively "flat" background that is not covered by the detector, in particular in medical imaging. Some of these issues are discussed by us in [22].

In Section 2 we describe the model-based cluster trees methodology. In Sections 3 and 4 we discuss aspects of algorithm design and properties. In Section 5 we will exemplify where the model-based tree approach is particularly important, and show how this algorithm performs exceedingly well in practice.

2 Model-Based Cluster Trees

Our basic framework is that of *model-based clustering*, as described, for example, by Fraley and Raftery [11, 12]. In this methodology, a finite mixture of normal distributions is fit to the data by maximum likelihood estimation using the EM algorithm, the number of groups is chosen using Bayesian model selection, and if hard clustering is desired, each pixel is assigned to its most likely group *a posteriori*. Model-based cluster *trees* produces a clustering of multivariate data by clustering on each band or dimension recursively.

We now briefly outline finite mixture modeling, Bayesian model selection, and model-based cluster trees.

2.1 Univariate Finite Gaussian Mixture Models

In the univariate finite Gaussian mixture model, one-dimensional observations x_i are assumed to be drawn from G groups, each of which is Gaussian distributed. The g-th group has mean μ_g and variance σ_g^2. Given observations $x = (x_1, \ldots, x_n)$, let γ be an unobserved $n \times G$ cluster assignment matrix, where $\gamma_{ig} = 1$ if x_i comes from the g-th group, and $\gamma_{ig} = 0$ otherwise. Our goals are to determine the number of clusters G, to determine the cluster assignment of each pixel, and to estimate the parameters μ_g and σ_g of each cluster.

The probability density for this model is

$$f(x_i|\theta, \lambda) = \sum_{g=1}^{G} \lambda_g f_g(x_i|\theta_g), \tag{1}$$

where $\theta_g = (\mu_g, \sigma_g^2)^T$, $f_g(\cdot|\theta_g)$ is a Gaussian density with mean μ_g and variance σ_g^2, $\theta = (\theta_1, \ldots, \theta_G)$, and $\lambda = (\lambda_1, \ldots, \lambda_G)$ is a vector of mixture probabilities such that $\lambda_g \geq 0$ $(g = 1, \ldots, G)$ and $\sum_{g=1}^{G} \lambda_g = 1$.

We estimate the parameters by maximum likelihood using the EM (expectation-maximization) algorithm [9, 16]. For its application to model-based clustering, see [15, 6, 8]. This is a procedure for iteratively maximizing likelihoods in situations where there are unobserved quantities and estimation would be simple if these were known. In the clustering case, the unobserved quantities are the cluster assignments given by the matrix γ.

The EM algorithm iterates between the E step and the M step. In the E step, the conditional expectation, $\hat{\gamma}$, of γ given the data and the current estimates of θ and λ is computed, so that $\hat{\gamma}_{ig}$ is the conditional probability that x_i belongs to the g-th group. In the M step, conditional maximum likelihood estimators of θ and λ given the current $\hat{\gamma}$ are computed.

The E step and the M step are both simple, so that the EM algorithm as a whole is also simple. By contrast, direct maximization of the likelihood for the mixture model is complex in general. Although the EM algorithm has some limitations (e.g. it is not guaranteed to converge to a global rather than a local maximum of the likelihood), it is generally efficient and effective for Gaussian clustering problems.

This procedure is especially efficient for clustering image pixels using single color bands or grayscale images. In general the EM algorithm requires $O(n)$ time, where n is the number of pixels. However, typically pixels can have one of only a limited number, ℓ, of intensities in each band, such as 256. If we first summarize the data by the counts of the numbers of pixels with each intensity level, the EM algorithm becomes an $O(\ell)$ algorithm rather than an $O(n)$ one. Since n, the number of pixels, is often very large, and ℓ is typically 256, this is a major speed-up and provides a reason for clustering one band at a time if this can be done without degrading performance too much.

2.2 Choosing the Number of Clusters via Bayesian Model Selection

We use Bayesian model selection to choose the number of clusters. For review of Bayesian model selection, see Kass and Raftery [14] and Raftery [24]. Pioneering work in this area was due to H. Jeffreys, I.J. Good and (according to the latter) A. Turing.

We consider a range of candidate numbers of clusters, $G = G_{\min}, \ldots, G_{\max}$. Each possible number of clusters, G, implies a different statistical model for the data, M_G. The model M_G has a vector of unknown parameters, ψ_G, consisting of the G means, the G variances, and the $(G-1)$ independently estimated mixture probabilities: $(3G-1)$ parameters in all. Our prior model probabilities are $p(M_G)$ for $G = G_{\min}, \ldots, G_{\max}$, where $\sum_{G=G_{\min}}^{G_{\max}} p(M_G) = 1$. Often each number of clusters considered is taken to be equally likely *a priori*, so that $p(M_G) = 1/(G_{\max} - G_{\min} + 1)$ for each G. The model parameters ψ_G also have prior distributions $p(\psi_G|M_G)$, which are typically rather diffuse and do not affect the final conclusions unduly. The data produce posterior model probabilities, $p(M_G|x)$, where again $\sum_{G=G_{\min}}^{G_{\max}} p(M_G|x) = 1$.

By Bayes' theorem,

$$p(M_G|x) = \frac{p(x|M_G)p(M_G)}{\sum_{H=G_{\min}}^{G_{\max}} p(x|M_H)p(M_H)}, \quad G = G_{\min}, \ldots, G_{\max}. \tag{2}$$

In (2), $p(x|M_G)$ is the *integrated likelihood* of model M_G, which requires integration over the model's parameter space, as follows:

$$p(x|M_G) = \int p(x|\psi_G, M_G)p(\psi_G|M_G)d\psi_G, \tag{3}$$

by the law of total probability.

The integral (3) is intractable analytically and is not easy to evaluate. However, twice the logarithm of the integrated likelihood can be approximated by the Bayesian Information Criterion, or BIC:

$$\begin{aligned} 2\log p(x|M_G) &\sim 2\log p(x|\hat{\psi}_G, M_G) - (3G-1)\log n \\ &= \mathrm{BIC} \end{aligned} \tag{4}$$

(See [26, 14, 24].) In (4),

$$p(x|\hat{\psi}_G, M_G) = \prod_{i=1}^{n} \sum_{g=1}^{G} \hat{\lambda}_g f_g(x_i|\hat{\theta}_g)$$

is the maximized likelihood. In words, BIC = 2(log maximized likelihood) + (log n)(number of parameters). The BIC measures the balance between the improvement in the likelihood and the number of model parameters needed to achieve that likelihood. While the absolute value of the BIC is not informative, differences between the BIC values for two competing models provide estimates of the evidence in the data for one model against another. The use of the BIC in choosing clusters in a mixture or clustering model is discussed by Roeder and Wasserman [25] and Dasgupta and Raftery [8]. Applications of ours using the same Bayesian decision principles with different imaging problems can be found in Campbell et al. [4, 5] (machine vision), Mukherjee et al. [18] (data mining), and in Murtagh et al. [22] (remote sensing). An alternative derivation of BIC as a minimum description length (MDL) criterion is described by Hansen [13].

2.3 Model-Based Cluster Trees Algorithm

The algorithm can be summarized as follows where at each level of the clustering tree we make use of the BIC in order to allow for objective choice of number of model components:

1. For the first image band, use BIC to choose the number of clusters (we evaluate BIC for $G_{\min} = 1, \ldots, G_{\max} = 40$). Use the EM algorithm to estimate the parameters of the mixture model, and assign each pixel to the group to which it is most likely to belong *a posteriori*.

   (In section 2.6 below we will enhance this clustering step to a segmentation step.)

2. For each cluster identified in step 1, carry out a separate model-based cluster analysis, this time using only the pixel intensities in the second image band. Each cluster identified in step 1 is then itself subdivided into several clusters.

3. On demand, for each (sub)cluster identified in step 2, subdivide it further using the same procedure as in step 2, but this time using only the pixel intensities in the third image band.

The univariate Gaussian mixture model fitting was carried out using an algorithm initially developed by Stanford [28]. As an indication of algorithm performance on a Sun SparcStation 10, dependence on the number of clusters, G, is approximately linear. (For a 768×512 image, $G = 3$ required 20 seconds, and $G = 39$ required 131 seconds.) Dependence on image dimensionality was found to be sub-linear. (Dimensions 300×300: 11 seconds. 900×900: 87 seconds.)

2.4 Spatial Segmentation and a Modified Bayes Information Criterion

Model fitting to the marginal density pays no attention to two-dimensional image spatial information. We can take such information into account using a hidden Markov model (HMM). Background on the approach pursued here can be found in Stanford [28] and Stanford and Raftery [29].

We consider an unknown, true pixel state, for pixel i, as $X_i \in \{1, 2, \ldots K\}$ for K states. The observed image pixel is Y_i. This can be taken either as a scalar, or instead as a vector for color or multiband images. Consider an indicator function, $I(X_i, X_j) = 1$ if $X_i = X_j$ and otherwise $= 0$.

We now use a Markov random field to define spatial structure on X. We take $p(X)$ as being proportional to $\exp(\phi \sum_{i,j} I(X_i, X_j))$. This is a Potts or Ising model. ϕ is a spatial homogeneity parameter, a small value implying randomness, and a large value implying uniformity. A negative value of ϕ implies dissimilarity between neighboring pixels, and is not of interest here. Our model is a hidden Markov model because the variables X are only known through the observed Y.

Let $N(X_i)$ be the neighborhood of X_i, e.g. 3×3 pixels. Let $U(N(X_i), k)$ be the number of neighborhood pixels with state k.

From $p(X)$ we have the conditional distribution:

$$p(X_i = j \mid N(X_i), \phi) = \frac{\exp(\phi U(N(X_i)), j)}{\sum_k \exp(\phi U(N(X_i)), k)} \tag{5}$$

Having looked at the latent space, we now return to the observed data. We assume the following conditional density model connecting the observed and hidden variables: $f(Y_i \mid X_i = j)$ is Gaussian with mean μ_j and standard deviation σ_j. In the multiband case, where y is a vector, the mean vector is used, and the variance-covariance matrix. The Y_i are conditionally independent given the X_i or, alternatively expressed, dependence among the Y_i only occurs via dependence among the X_i. Call θ_k the set of parameters, (μ, σ^2) for state k. We have $f(Y \mid X) = \Pi_i f(Y_i \mid X_i) = \Pi_i f(Y_i \mid \theta_{X_i})$.

Our solution algorithm is as follows. It is based on Besag's [3] iterated conditional modes (ICM) algorithm, which reconstructs an image based on local properties modeled as a Markov random field. This iterative algorithm requires an initial estimate of X, $\hat{X}$, and proceeds to estimate the parameters of $p(Y_i \mid X_i)$, as well as ϕ and X. To initialize X, we note that in taking $p(Y_i \mid X_i)$ as Gaussian, then the marginal density of Y is a finite mixture of Gaussians. In the multidimensional case, we either use a marginal density model on some selected band, or alternatively use the marginal density model of the eigen or principal component image. The EM-based modeling of the marginal density discussed in section 2.1 then applies.

Segmentation Algorithm:

Step 0: Initialize $\hat{X}$ using a marginal segmentation.

Step 1: Update $\hat{\theta} = \text{ argmax } f(Y \mid \hat{X})$ based on maximum likelihood estimates of μ_j and θ_j for each class, j.

Step 2: Update ϕ using the maximum pseudo-likelihood: $\hat{\phi} = \text{argmin}_{\phi}(-\log \text{PL}(\hat{X} \mid \phi))$. The pseudo-likelihood is given by $\text{PL}(\hat{X} \mid \phi) = \Pi_i p(\hat{X}_i \mid N(\hat{X}_i, \phi))$.

Step 3: Update $\hat{X}$: for each pixel i, $\hat{X}_i = \text{argmax}_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i, \hat{\phi}))$.

Implementation details: In step 2, we initialize $\hat{\phi}$ to 1.4 (which was found to work well with the golden ratio search algorithm used, with overall search limits of -2 and 15) and we constrain $\hat{\phi}$ to be greater than zero. In all calculations, we exclude boundary pixels from consideration. Step 1 is one step of Besag's ICM algorithm.

2.5 An Information Criterion with Spatial Interaction, PLIC

We now turn attention to model selection. This is developed not for the homogeneity parameter, ϕ, nor for the neighborhood, but rather for the number of segments, K (see [29]).

In the spatial (Markov) case, the likelihood (first term) in the BIC, equation (3), is problematic for computational reasons.

The posterior distribution of X conditional on Y is: $f(X \mid Y) = f(Y \mid X)f(X)/f(Y) \propto f(Y \mid X)f(X)$. Since there is conditional independence between Y and X, we have that $f(Y \mid X) = \Pi_i f(Y_i \mid X_i)$ which, it has already been noted, is taken as Gaussian.

The density of x, $f(X)$, is related to all possible states, which is combinatorially explosive. Therefore the pseudo-likelihood, PL(X), is taken as a proxy for $f(X)$. The pseudo-likelihood, introduced in Besag [2], restricts where the integrated likelihood is defined. We have

$$\mathrm{PL}(X, \phi) = \prod_i p(X_k \mid N(X_i), \phi) = \prod_i \frac{\exp(\phi U(N(X_i)), X_i)}{\sum_k \exp(\phi U(N(X_i)), k)} \tag{6}$$

The likelihood is made conditional on the neighborhood of pixel i. Previously we had

$$L(Y_i \mid X_i) = \sum_j f(Y_i \mid X_i = j) p(X_i = j) \tag{7}$$

for state or label j.

Instead, denoting X_{-i} the neighborhood of X_i not including pixel i, and with $\hat{X}$ denoting an estimate of X, we use:

$$L(Y_i \mid N(\hat{X}_{-i})) = \sum_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i)) \tag{8}$$

As already noted, the first part of the right hand side term requires evaluation of a Gaussian; and the second part uses the conditional distribution defined for $p(X)$ in equation 5.

From the product of pseudo-likelihoods for all pixels, we arrive at a modified BIC, modifying equation (3). This modified criterion is termed the pseudo-likelihood information criterion, PLIC [28, 29].

2.6 Model-Based Segmentation/Clustering Trees Algorithm

The algorithm can be summarized as follows where initially we use a segmentation, with use of the PLIC in order to allow for objective choice of number of model components:

1. For the first image band, use PLIC to choose the number of segments (we evaluate PLIC potentially for $G_{\min} = 1, \ldots, G_{\max} = 40$). Use the ICM algorithm to estimate the parameters of the mixture model, and assign each pixel to the segment to which it is most likely to belong *a posteriori*.

2. For each segment identified in step 1, carry out a separate model-based cluster analysis, this time using only the pixel intensities in the second image band. Given that pixel intensities are used, the appropriate assessment criterion here is the BIC. Each cluster identified in step 1 is then itself subdivided into several clusters.

3. On demand, for each (sub)cluster identified in step 2, subdivide it further using the same procedure as in step 2, but this time using only the pixel intensities in the third image band.

The segmentation model fitting was carried out using an algorithm initially developed by Stanford [28]. We can quite straightforwardly carry out this segmentation on a number of bands simultaneously in step 1. In this case we are fitting our Gaussian mixture and Markov model to the pixel vectors in a multidimensional pixel space.

As an indication of algorithm performance, indicative timings on a Sun SparcStation 10 relative to the number of clusters, G, are as follows. For a 256×256 image, $G = 2$ and $G = 20$ required, respectively, 74 and 412 seconds. Indicative timings relative to image dimensionality were as follows. For the $G = 2$ case above we have 74 seconds for a 256×256 image, and we have 317 seconds for a 768 × 512 image. Finally, indicative timings as a function of number of bands are as follows. For $G = 2$, from the foregoing result we have 317 seconds for a $768 \times 512 \times 1$ image, and we find 535 seconds for a $768 \times 512 \times 3$ image.

3 Some Algorithm Properties

3.1 Band Ordering

In this section, we follow closely Tate [31] who considers the band ordering problem for compression of multispectral images.

We consider the problem of clustering on one band coordinate, assuming that this band presents good clustering properties, followed by clustering on a second band based on the first band clustering, and so on. If c_{1i} is the ith cluster from the first band, then we seek clusters c_{2j} such that $c_{1i} = \cup_{j \in J_i} c_{2j}$, i.e., J_i is a partition of cluster c_{1i}. Similarly we proceed to a third band, based on available results for bands 1 and 2.

The order in which we consider the bands is evidently important. Let us define goodness of clustering as the tightness of the clusters, i.e. the minimum sum of variances of clusters for all bands. This is in line with the compression objective of [31] and can be justified on minimal entropy grounds also.

Clearly a result of this definition is that when we find no subclustering at bands 2 and 3, the corresponding subcluster variances are equal to 0, and hence the contribution to the overall sum of variances is thereby minimized.

Finding the optimal band ordering for clustering can be tackled by exhaustively checking all orderings of bands. But this can be computationally demanding.

Define graph G = (V, E) such that an edge E_{ij} has weight w_{ij} representing the added clustering quality attainable by clustering band i before band j. By design, band i is partitioned, and the clusters of the i-partition are each partitioned based on band j information. Define w_{ij} as the *improvement* in the overall sum of posteriors (or sum of intra-cluster variances: the criterion used to quantify clustering quality is not important here) by taking band i before band j.

The problem of finding an optimal band order in the general case of many bands is equivalent to finding an optimal traveling salesman path in the graph, G. This Hamiltonian path problem in NP-hard and the corresponding decision problem of knowing whether we have or do not have a Hamiltonian path in G is NP-complete.

3.2 Model Component Labels

For a Gaussian mixture model fit to a single grayscale image, i.e., fitting a mixture model to the image's marginal density, we can impose the following label monotonicity rule:
For clusters c_i of means $m(c_i)$, and labels $l(c_i)$ we require: $l(c_i) < l(c_j) \iff m(c_i) < m(c_j)$ for all i and j.

Next consider the level 2 analysis of cluster c_i above. For all pixels in band 2, which carry label $l(c_i)$ at level 1, we form the marginal density, fit a mixture model to this, and determine level 2 labels in accordance with the level 2 result: $l(c_i^{(2)}) < l(c_j^{(2)}) \iff m(c_i^{(2)}) < m(c_j^{(2)})$

Clearly, considering both the set of level 1 clusters, and the embedded level 2 clusters, it is not difficult to impose a clustering labeling which varies monotonically with cluster means. We are, after all, dealing with scalar means, m, throughout this processing.

In the case of segmentation, where neighborhood influence is handled using a Markov model, the processing of a single grayscale image reverts to an identical perspective on model component labels as was seen above. Scalar segment means imply monotonicity of cluster labels, even in the tree-based or multi-level situation.

The tree-based approach thus offers a practical advantage over the the multidimensional segmentation alternative. In the multidimensional segmentation case, i.e. segmentation of multiply valued pixels, there is no immediate monotonicity property for the segment labels. Therefore in this case the labeling which is established is arbitrary. This case of multidimensional segmentation has the potential to make life difficult for us in regard to comparative assessment and evaluation of different result. In such a situation, we have little alternative but to construct a cross-tabulation of pixel assignments to all segments, taking pairs of analysis results into consideration, and then select the best-match segment associations between pairs of results.

4 Discussion of Alternative Approaches

We will consider the characteristics of three different case studies. The third one provides the theme of section 5 to follow.

The first case study [7] relates to 6 Hubble Space Telescope NICMOS (Near Infra-Red Camera and Multi-Object Spectrometer) infrared images (0.8–2.5 microns) of the M82 region. M82 will also be the focus of our third study below. M82 is the nearest starburst galaxy at a distance of 11 million light years from Earth. M82, cigar shaped, is bright (magnitude 6.9). In it, massive stars are forming and expiring at ten times the corresponding rate in our Galaxy.

The 6 images were exactly registered. They were of dimensions 256×256. The images were highly correlated, with correlation coefficients between the 6 bands ranging between 1.0 (near identity) to 0.83. Hardly surprisingly, therefore, a segmentation carried out in 6-dimensional space, with use of the PLIC criterion to evaluate the optimal number of segments, encountered cluster singularity problems. This points to one limitation of a Markov model-based mixture fit to multiband data: when the data lie on a less than full dimensionality manifold, or when the clusters have zero variance, then the approach does not work and instead further enhancements of the approach are needed.

The second case study [1] relates to AVHRR/2 and /3 imagery from the NOAA-14 satellite of the Atlantic Ocean in the region of the Canary Islands. Five bands were used, each of pixel dimensions 3313×2048. Again, correlations between some of these images were very strong (1.00, 0.99) and less so between others ($-0.40, -0.43$). A principal component image, or eigen-band, was used as a starting point for segmentation. This avoided the problem of singularity, which otherwise was problematic. Further work in [22] used data from the MODIS (Moderate Resolution Imaging Spectroradiometer) instrument, on the Terra (EOS AM-1) spacecraft commissioned as part of NASA's Earth Observing System.

The study in the next section, section 5, relates to images of the M82 galaxy, this time using the Palomar (California) Digitized Sky Survey providing optical images; and the 2MASS near-infrared (J-band or 1.25 micron band) survey, using telescopes at either Mt. Hopkins (Arizona) or the Cerro Tololo Inter-American Observatory (CTIO, Chile). The images were selected and obtained from [27]. Figures 1 and 2, respectively, show these images. These images are of dimensions 300×300, but are, respectively, of size 0.14166667 and 0.083310007 degrees square. Hence Figure 1 was rescaled to the size spanned by Figure 2, using cubic convolution interpolation (which provides a good approximation to the theoretically optimal sinc interpolation). Then extracting a 300×300 image from rescaled Figure 1 yielded Figure 3.

The multiband image set on which we will now continue our work comprised a first band as shown in Figure 3, and a second band as shown in Figure 2. We deliberately targeted a small multiband set, here a 2-band set, for expository convenience. The results we obtain here may be compared with the similar results obtained in [7]. The latter work of ours uses multiband data for the same region of the sky, but taken with a different detector.

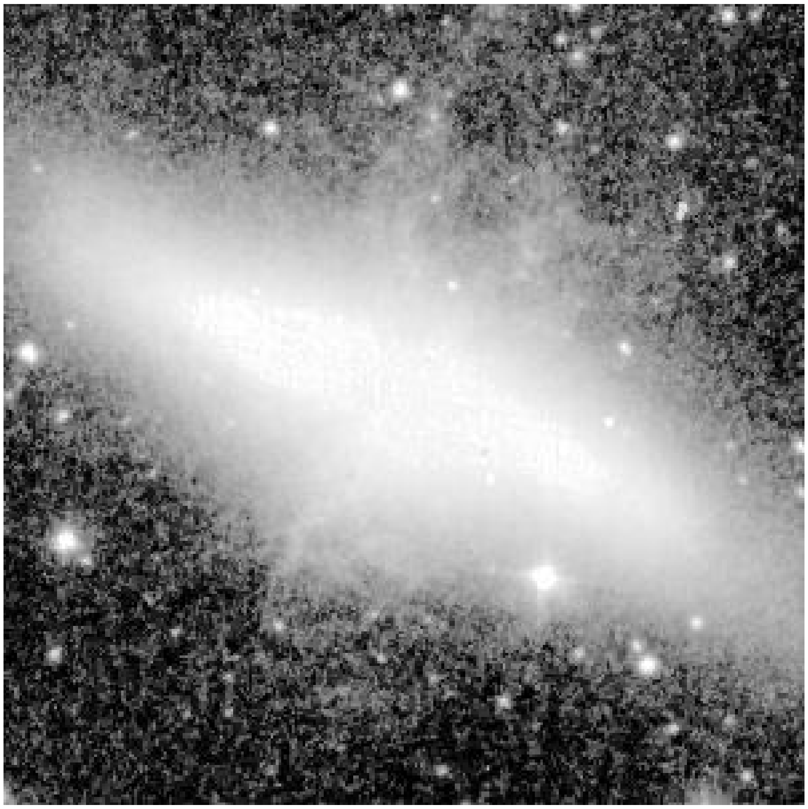

Figure 1: Digitized Sky Survey image of M82.

5 Appraisal

We took the band in Figure 3 as our point of departure. Segmentation of it was carried out for varying numbers of segments. This segmentation used a Markov model and ICM-based component mixture fit, as described in sections 2.4, 2.5, and 2.6 above. The PLIC criterion provides a basis for the selection of the best segmentation. This provides an assessment of a number of segments G_{i+1} against the alternative of G_i segments. The greater the value of PLIC for G_i then the more favorable is the evidence for this segmentation.

From Figure 4 we observe the following. The best number of segments

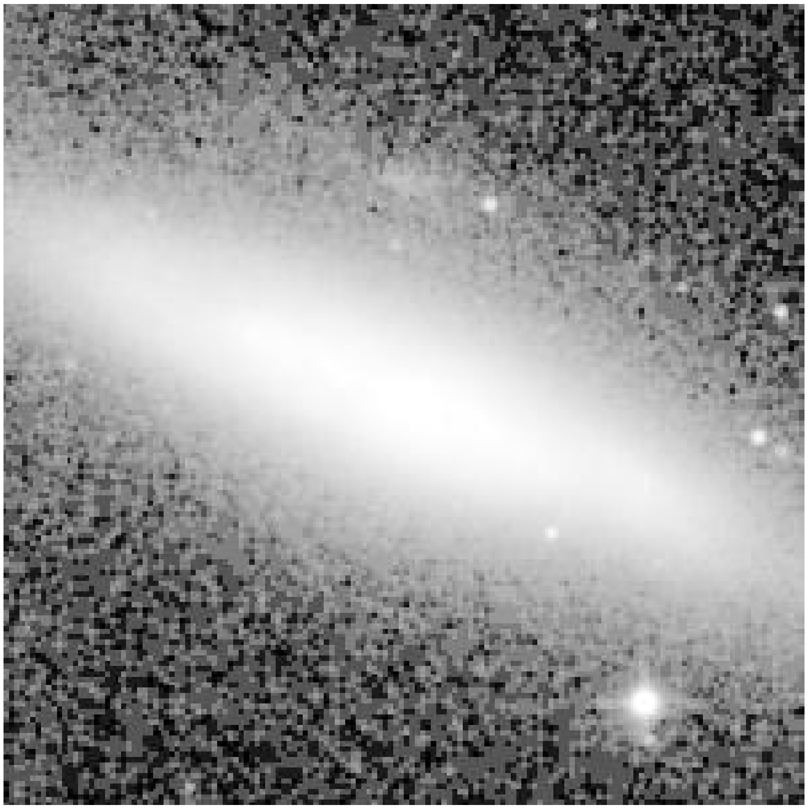

Figure 2: 2MASS Survey image of M82.

is 1, which is not acceptable, so we exclude this solution. The next best solution is given by 2 segments. This we will use now, in order to proceed to the next level of a tree- or level-based clustering. We also note that for a number of segments greater than 2, the 8-segment solution is best. We will compare this later to the cluster tree solution. The 2-segment solution is shown in Figure 5, and the 8-segment solution is shown in Figure 6.

We now proceed to a level 2 clustering. A Markov spatial model at level 2 is impossible: consider the fact that contiguous zones are distributed throughout the image, with irregular neighborhoods. Therefore level 2 and later levels, if required, make use of quantization or clustering, as reviewed in sections 2.1, 2.2 and 2.3.

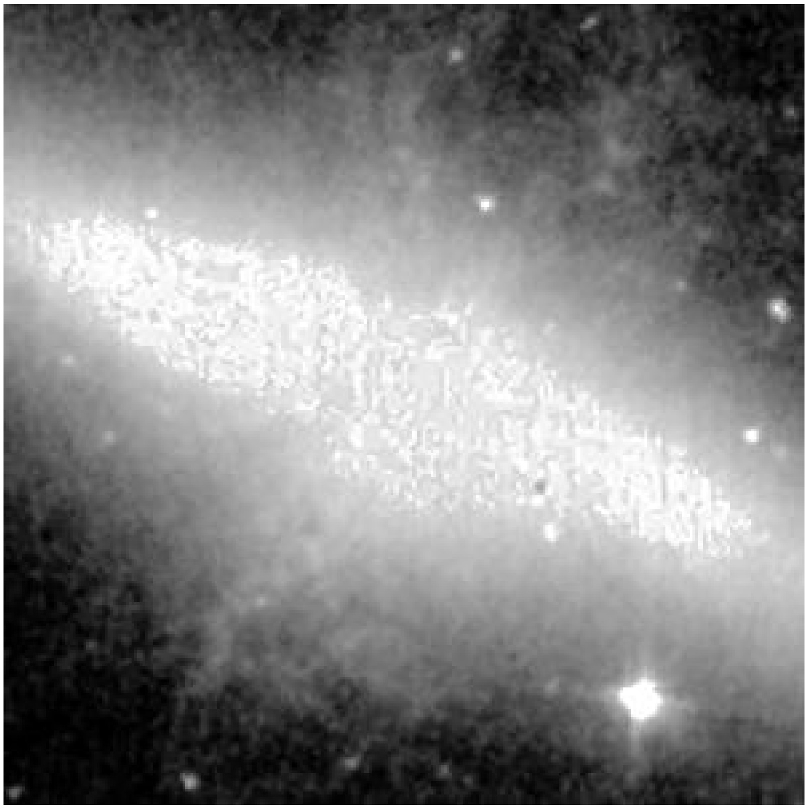

Figure 3: Digitized Sky Survey, rescaled and extracted, image of M82.

Figures 7 and 8 show the BIC values, which motivate in both cases a 3-cluster result. In both cases, this is essentially when the plateau is reached. In the case of BIC, we can usually continue indefinitely with a greater number of components to obtain a better fit. BIC values usually stabilize on an approximate curve plateau as can be observed in Figures 7 and 8.

To facilitate comparison with the direct 8-segment result in Figure 6, we show the results of Figures 9 and 10 in one Figure: see Figure 11. As always histogram-equalization is used in these figures to show fine detail.

In the upper central regions, Figure 11 seems to perform better in demarcating plume-like structures. Similarly towards the bottom, in left-of-center regions, the faint plume structures are fairly well characterized by the clusters

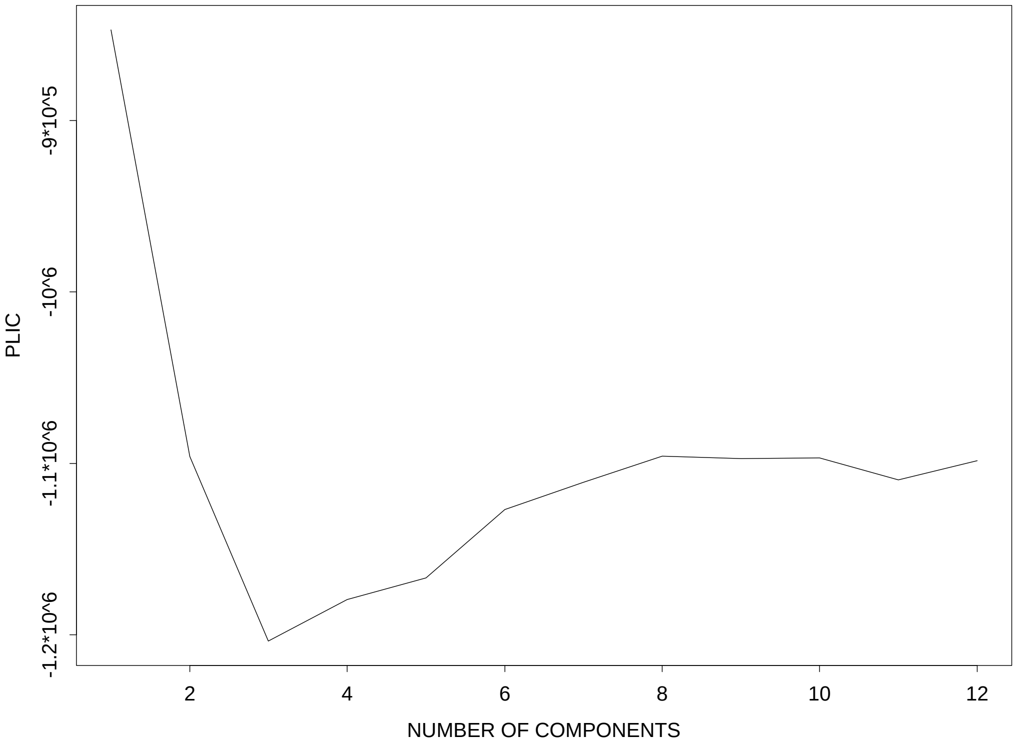

Figure 4: PLIC values found for varying numbers of segments, arising from segmentation of Figure 3.

or segments found.

In this work, we use Figure 6 simply to demonstrate that the tree cluster model, with final result in Figure 11, performs very satisfactorily. The most important argument in favor of Figure 11 is that it has allowed us to perform information fusion, between the Digitized Sky Survey and 2MASS Survey image data, in obtaining our clustering result. On the other hand, Figure 6 uses the Digitized Sky Survey data only.

The cluster tree used in Figure 11 has a two-way split at level 1, followed by a three-way split at both nodes at level 2. Node splitting was determined by the PLIC criterion at level 1, which corresponds to segmentation, and

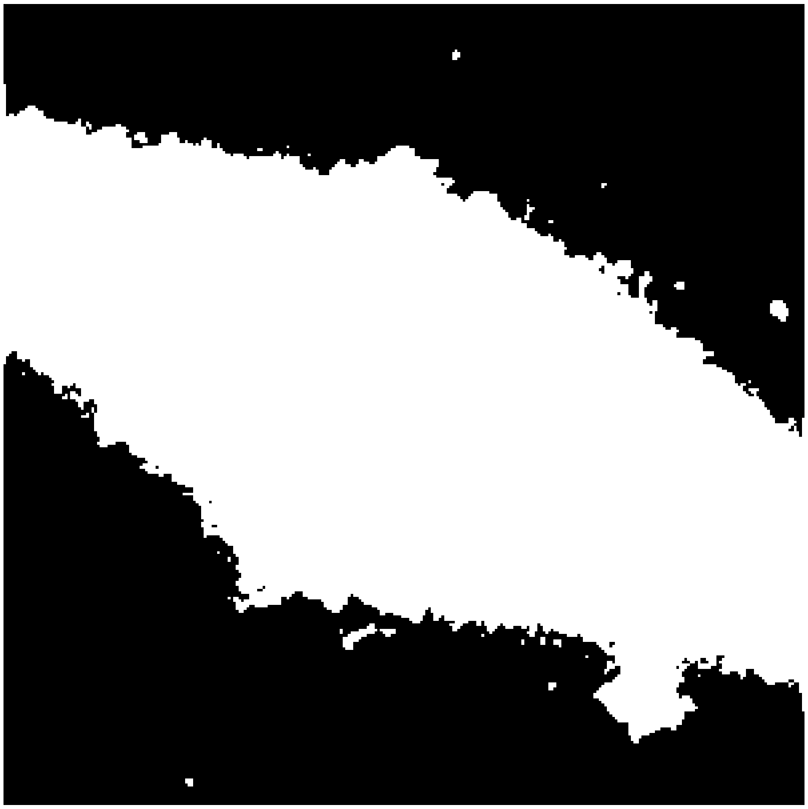

Figure 5: Result of segmenting Figure 3 into 2 segments. Cf. PLIC value corresponding to the 2-component solution in Figure 4.

twice by the BIC criterion at level 2, which correspond to grayscale quantization.

The cluster tree analysis approach is justified when image bands are sufficiently different. If, on the other hand, they are not different (as expressed, for example, by high-valued correlation coefficients), then a direct multiband segmentation is appropriate, or a principal component analysis. In all cases, of course, the pixel data must be very carefully registered before any processing.

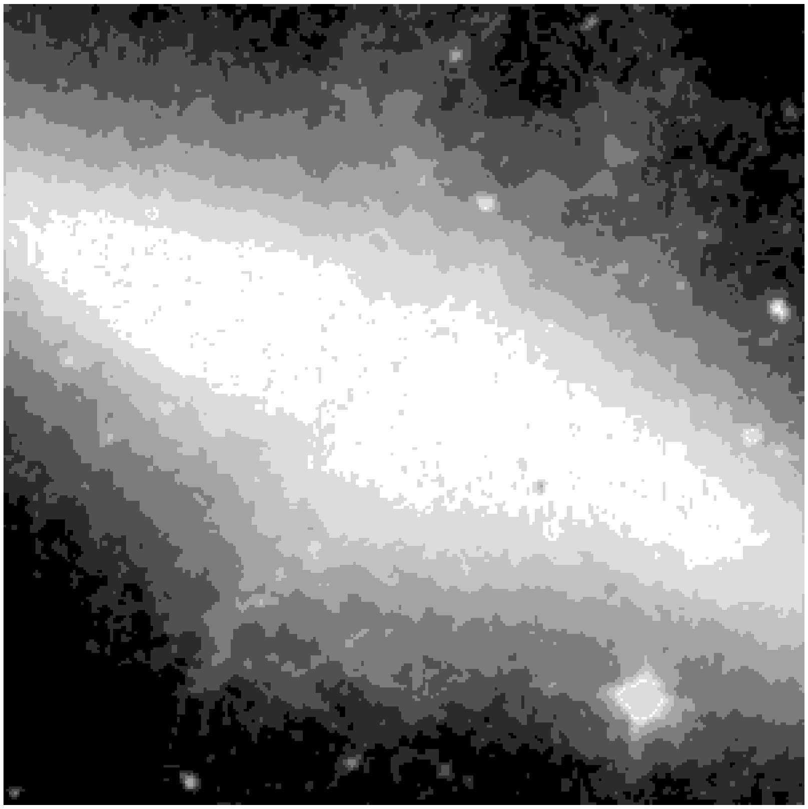

Figure 6: Result of segmenting Figure 3 into 8 segments. Cf. PLIC value corresponding to the 8-component solution in Figure 4.

6 Discussion

We have shown how a Bayesian modeling approach, model-based cluster trees, can lead to excellent results in the area of multiband image clustering. Formal underpinnings for such an algorithm facilitate choice of system parameters (e.g. number of clusters) which in a general setting would be set arbitrarily.

This approach allows us to carry out information fusion from multiband image data in a fully integrated way. We have discussed where and when this approach is particularly appropriate.

The fitting of a tree of Gaussian components may be of benefit in the

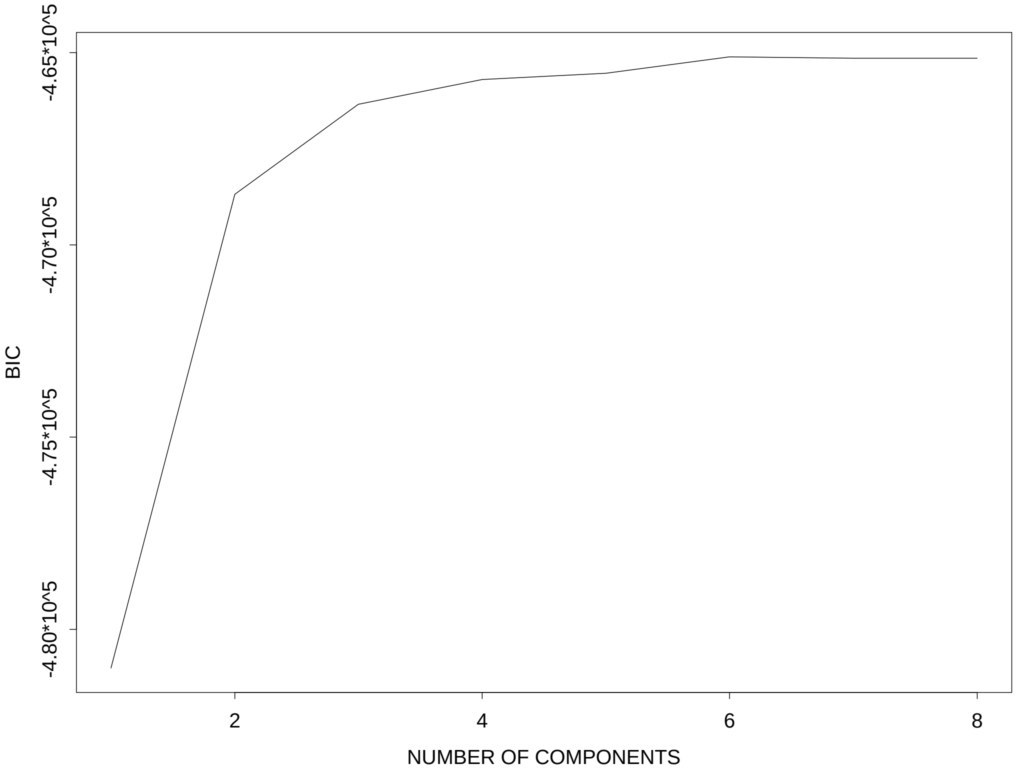

Figure 7: BIC values for class 1 (regions outside central area) in Figure 5.

exploration of general parameter spaces, since we are not overly dependent on an analysis function or kernel (here: Gaussian) of given morphology. The partial order resulting from the tree of Gaussian components can (at least in principle) accommodate arbitrary alignments or curves in multidimensional clusters. Djorgovski et al. [10] describe a burgeoning need for solving such problems in the general area of data mining. In section 1 we mentioned large-scale multiband image segmentation as a further domain of application. We can easily envisage application to other applications also, such as signal quantization [20] and image edge detection and processing [21].

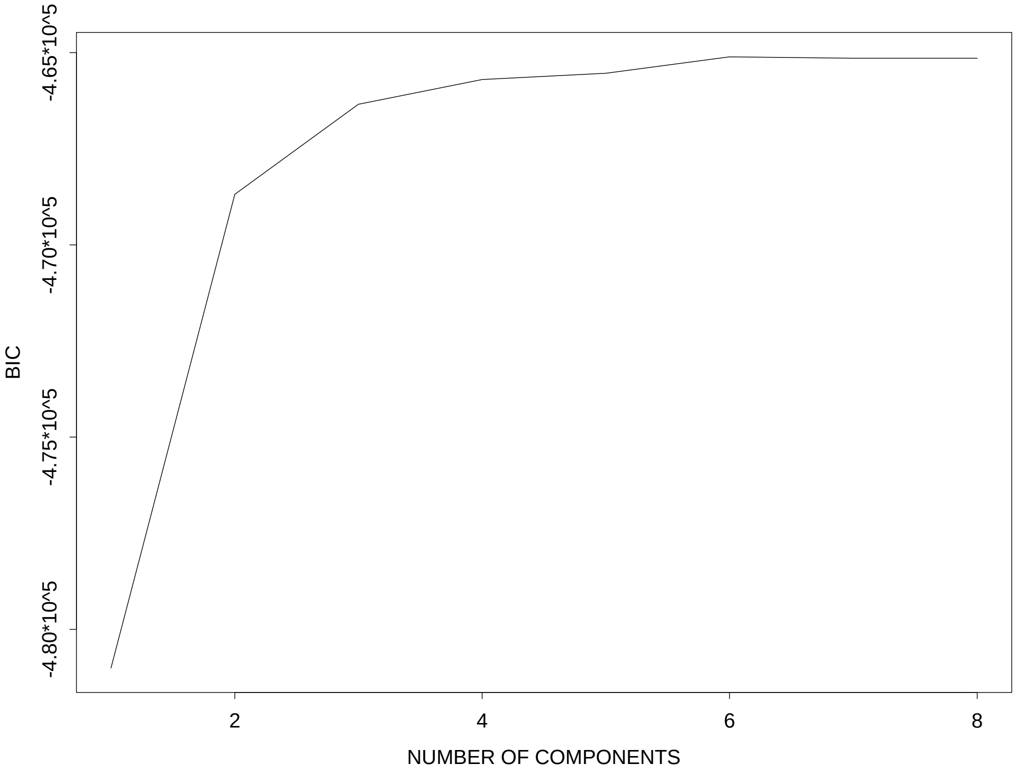

Figure 8: BIC values for class 2 (central area) in Figure 5.

Acknowledgments

Raftery's research was supported by NIH Grant 1R01CA094212-01 and ONR Grant N00014-01-10745. Murtagh acknowledges helpful discussions within the IDHA and MDA projects supported by the French Ministry of Education's "Data Masses" program, and also the European project iAstro, COST Action 283.

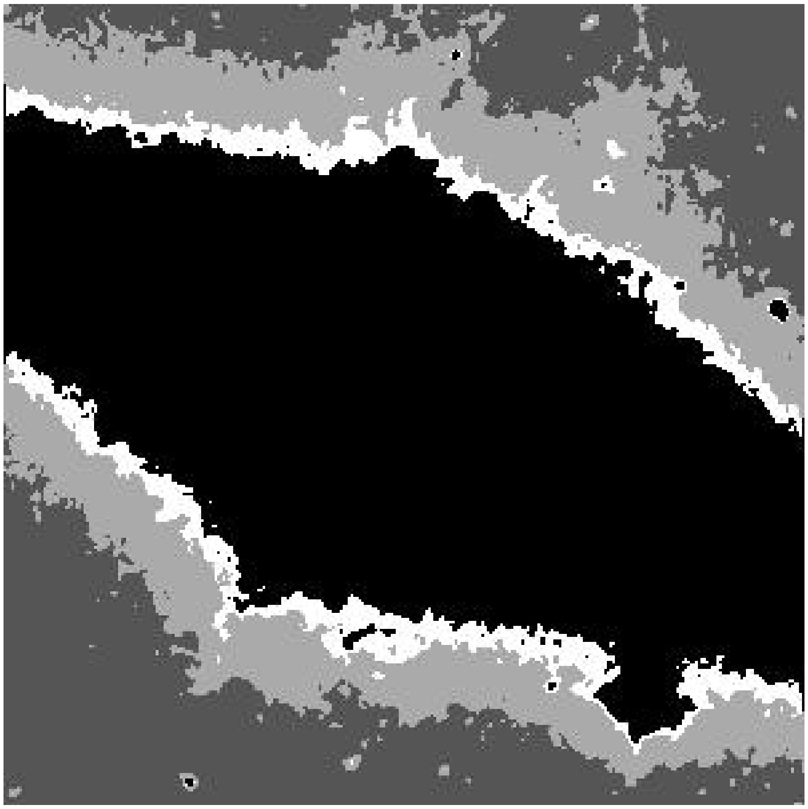

Figure 9: Class 1 from Figure 5 further subdivided into 3 clusters or quantization levels.

References

[1] D. Barreto, F. Murtagh and J. Marcello, “Bayesian segmentation and clustering for determining cloud mask images”, in: Opto-Ireland 2002: Optical Metrology, Imaging, and Machine Vision. Eds. A. Shearer, F.D. Murtagh, J. Mahon, and P.F. Whelan, SPIE Proceedings, Vol. 4877, SPIE, Bellingham (2003) pp. 144-155.

[2] J. Besag, “Statistical analysis of non-lattice data”, Statistician, 24 (1975) 179–195.

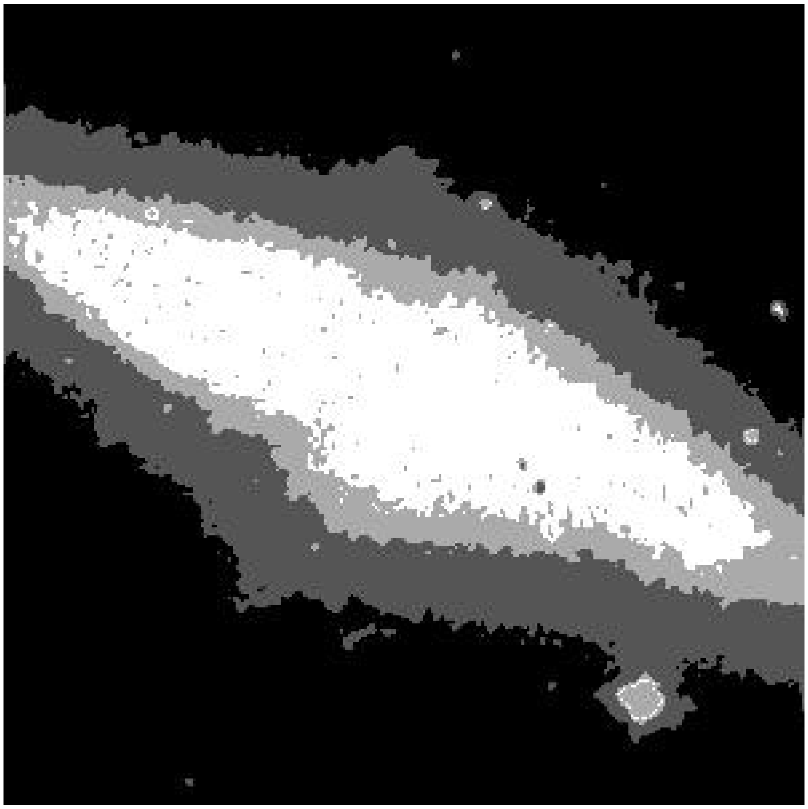

Figure 10: Class 2 from Figure 5 further subdivided into 3 clusters or quantization levels.

[3] J. Besag, "Statistical analysis of dirty pictures", Journal of the Royal Statistical Society, Series B, 48 (1986) 259–302.

[4] J.G. Campbell, C. Fraley, F. Murtagh and A.E. Raftery, "Linear flaw detection in woven textiles using model-based clustering", Pattern Recognition Letters, 18 (1997) 1539–1548.

[5] J.G. Campbell, C. Fraley, D. Stanford, F. Murtagh and A.E. Raftery, "Model-based methods for textile fault detection", International Journal of Imaging Science and Technology, 10 (1999) 339–346.

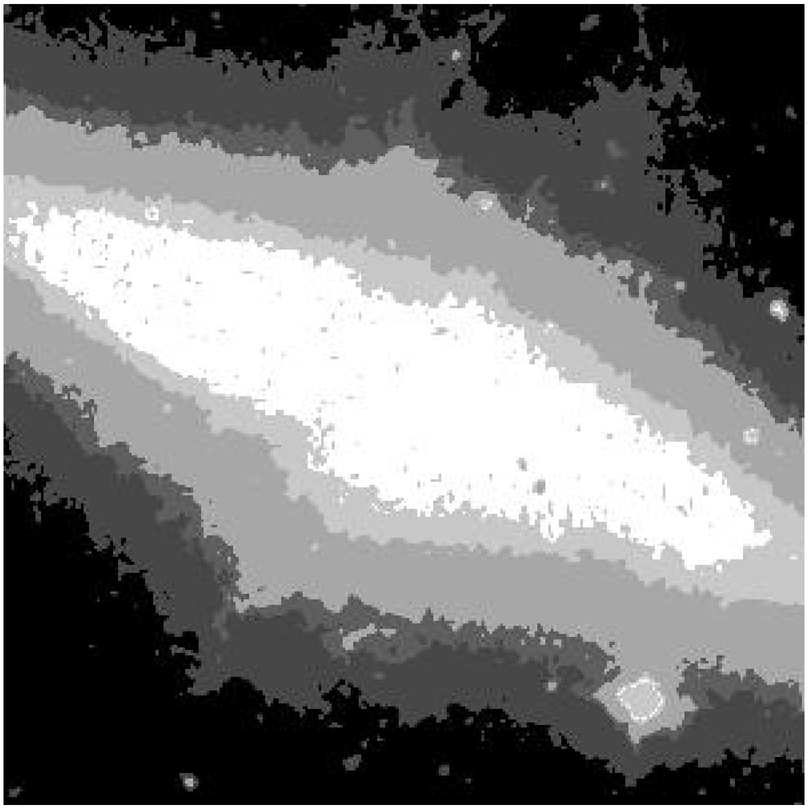

Figure 11: Figures 9 and 10 are shown together in this figure. Here we see the full level 2 clustering result, based on the level 1 segmentation result of Figure 5.

[6] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models", Pattern Recognition, 28 (1995) 781–793.

[7] C. Collet and F. Murtagh, "Multiband segmentation based on a hierarchical Markov model", Pattern Recognition, 37 (2004) 2337–2347.

[8] A. Dasgupta and A.E. Raftery, "Detecting features in spatial point processes with clutter via model-based clustering", Journal of the American Statistical Association, 93 (1998) 294–302.

[9] 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, Series, B, 39 (1977) 1–22.

[10] S.G. Djorgovski, A. Mahabal, R. Brunner, R. Williams, R. Granat, D. Curkendall, J. Jacob and P. Stolorz, "Exploration of parameter spaces in a virtual observatory", in J.L. Starck and F. Murtagh, Eds., SPIE Proceedings Vol. 4472, SPIE, Bellingham (2001) 43–52.

[11] C. Fraley and A.E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis", The Computer Journal, 41 (1998) 578–588.

[12] C. Fraley and A.E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," Journal of the American Statistical Association, 97 (2002), 611–631.

[13] M.H. Hansen and Bin Yu, "Model selection and the principle of minimum description length", Journal of the American Statistical Association, 96 (2001) 746–774.

[14] R.E. Kass and A.E. Raftery, "Bayes factors", Journal of the American Statistical Association, 90 (1995) 773–795.

[15] G. McLachlan and K. Basford, Mixture Models (1988) Marcel Dekker, Basel.

[16] G. McLachlan and T. Krishnan, The EM Algorithm and Extensions (1997) Wiley, New York.

[17] MR Multiresolution Analysis Software Environment, Volume 3 – MR/3 Multichannel Data, www.multiresolution.com (2001).

[18] S. Mukherjee, E.D. Feigelson, G.J. Babu, F. Murtagh, C. Fraley and A. Raftery, "Three types of gamma-ray bursts", The Astrophysical Journal, 508 (1998) 314–327.

[19] F. Murtagh, "A survey of algorithms for contiguity-constrained clustering and related problems", The Computer Journal, 28 (1985) 82–88.

[20] F. Murtagh and J.L. Starck, "Quantization from Bayes factors with application to multilevel thresholding", Pattern Recognition Letters, 24 (2003) 2001–2007.

[21] F. Murtagh and J.L. Starck, "Bayes factors for edge detection from wavelet product spaces", Optical Engineering, 42 (2003) 1375–1382.

[22] F. Murtagh, D. Barreto and J. Marcello, "Decision boundaries using Bayes factors: the case of cloud masks", IEEE Transactions on Geoscience and Remote Sensing, 41 (2003) 2952–2958.

[23] Soo-Chang Pei, Ching-Min Cheng and Lung-Feng Ho, "Limited color display for compressed image and video", IEEE Transactions on Circuits and Systems for Video Technology, 10 (2000) 913–922.

[24] A.E. Raftery, "Bayesian model selection in social research (with discussion by A. Gelman, D.B. Rubin and R.M. Hauser)". In Sociological Methodology 1995, Ed. Peter V. Marsden, Blackwells, Oxford (1995) pp. 111-196.

[25] K. Roeder and L. Wasserman, "Practical Bayesian density estimation using mixtures of normals", Journal of the American Statistical Association, 92 (1997) 894–902.

[26] G. Schwarz, "Estimating the dimension of a model", Annals of Statistics, 6 (1978) 461–464.

[27] Skyview Image Server, NASA Goddard Space Flight Center, http://skyview.gsfc.nasa.gov

[28] D.C. Stanford, Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Patterns, Ph.D. Dissertation, Department of Statistics, University of Washington, 1999.

[29] D.C. Stanford and A.E. Raftery, "Determining the number of colors or gray levels in an image using approximate Bayes factors: the pseudo-likelihood information criterion (PLIC)", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), 1517–1520.

[30] J.L. Starck and F. Murtagh, Astronomical Image and Data Analysis, Springer-Verlag, 2002.

[31] S. R. Tate, "Band ordering in lossless compression of multispectral images". IEEE Transactions on Computers, 46 (1997) 477–483.

[32] A.B. Watson, G.Y. Yang, J.A. Solomon and J. Villasenor, "Visibility of wavelet quantization noise", IEEE Transactions on Image Processing, 6 (1997) 1164–1175.