Bayesian Model Selection for Spatial Clustering in 3D Surveys

F. Murtagh[a], C. Donalek[b,c], G. Longo[c,b], and R. Tagliaferri[d,e]

[a] School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland
[b] INAF, Sezione di Napoli, via Moiariello 16, 80131 Napoli, Italy
[c] Department of Physical Sciences, University Federico II, Napoli, Italy
[d] Department of Mathematics and Applications, University of Salerno, Baronissi, Italy
[e] INFM Sezione di Salerno, via S. Allende, Baronissi, Italy

ABSTRACT

In this preliminary work on galaxy clustering, we study clusters of arbitrary shape using a 3D galaxy catalog (celestial coordinates and photometric redshifts) derived from Sloan Digital Sky Survey Early Release Data. Spatial influence is modeled using a Markov random field. Comparative model assessment is carried out using an approximation to Bayes factors, viz. the posterior odds of a hypothesis of a given number of clusters versus another. We conclude with a discussion of promising future research directions.

Keywords: astronomical surveys, galaxy cluster catalogs, Markov random fields, Gibbs distribution, cluster analysis, Bayes factors, model selection

1. INTRODUCTION

Large astronomical surveys, both photometric and spectroscopic, are providing the scientific community with an unprecedented amount of new high quality and high accuracy data, thus providing those statistically significant samples of objects which are needed to answer many still open astrophysical and cosmological questions. In this paper we address the problem of the construction of a large and well-defined (in terms of contamination and completeness) catalog of galaxy clusters. The need for such a catalog has been emphasized many times in the recent literature (cf. Refregier[1]) and much work continues to be carried out by different teams which are using different approaches to the problem.

First of all, we have to stress that the search for galaxy clusters is performed using either one of two types of dataset, namely 2D or 3D catalogs. The first type of catalog (e.g., the POSS galaxy catalogs[2]) provides projected coordinates (i.e., celestial coordinates: RA and Dec) for each galaxy, together with a possibly large set of auxiliary morphological and photometric parameters. These catalogs usually contain very large numbers of objects but, in order to extract putative galaxy clusters, they require extensive preprocessing aimed at increasing the contrast between "true" overdensities and the presumably homogeneous background and foreground distribution of field galaxies. In 3D catalogs the above information is complemented by a redshift, i.e. an estimate of the distance of the galaxies, thus allowing several orders of magnitude increase in the contrast between background/foreground galaxies and physical aggregates. The large demand for observing time needed to acquire the spectroscopic redshifts leads, however, to obtaining such data only for a small subsample of galaxies selected according to some magnitude or diameter criterion (cf. the Sloan Digital Sky Survey, SDSS[3]), thus biasing the final catalogs toward the brightest and the largest objects. In the case of large specifically-tailored multiband surveys it is possible to complement the redshift data using lower accuracy photometric redshifts (cf. Tagliaferri et al.[4] and references therein).

In this paper we make use of a catalog of galaxies extracted from the SDSS-Early Data Release.

Author contact information: e-mail f.murtagh@qub.ac.uk

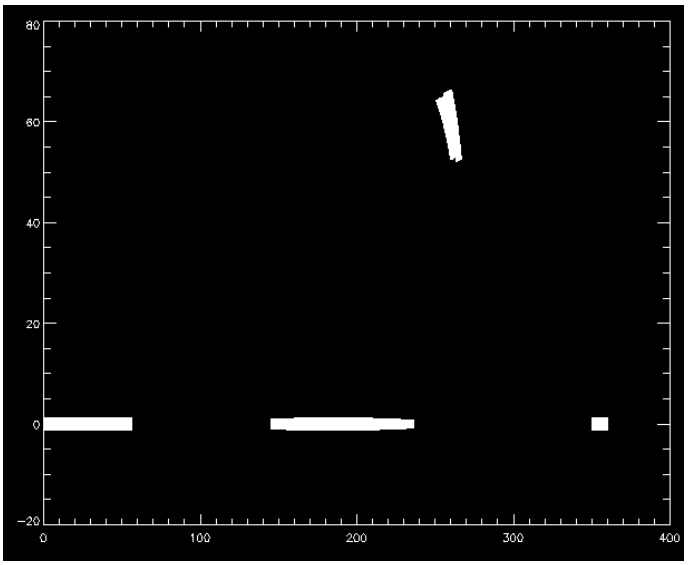

Figure 1. 345109 galaxies in right ascension and declination. In this work we used the contiguous rectangular area, or "bar", on the left (low RA).

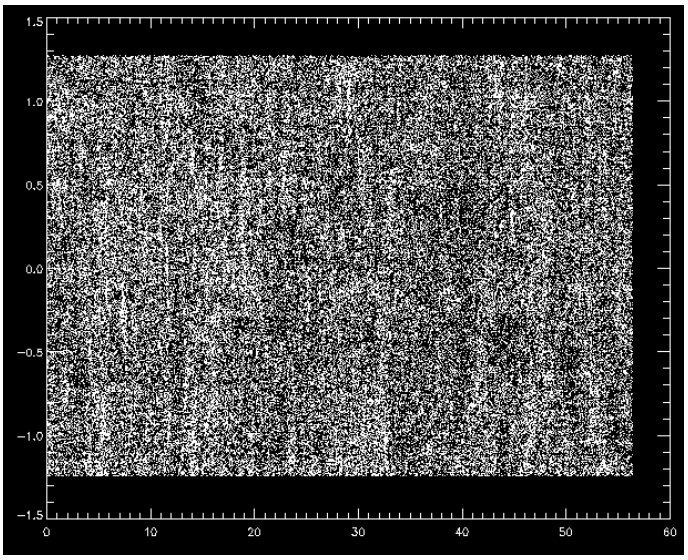

Figure 2: 118279 galaxies in RA and Dec contained in the rectangular region less than 60 degrees in RA in Figure 1.

2. DATA SELECTION AND INITIAL PROCESSING

The data were extracted from the SDSS-EDR and photometric redshifts were estimated as described in Tagliaferri et al..[4] The final catalog provides RA, Dec, and photometric redshift for all galaxies detected in all five bands of the SDSS and with estimated photometric redshift in the range 0.005–0.5 (with an average error in z of 0.021 and an estimated contamination of 0.004).

To explore clustering, we used the contiguous rectangular area on the left (low RA) in Figure 1. See Figure 2 for this subset of the data.

It is common to project positional data into a regular grid as a starting point for clustering.[5,6] The 118279

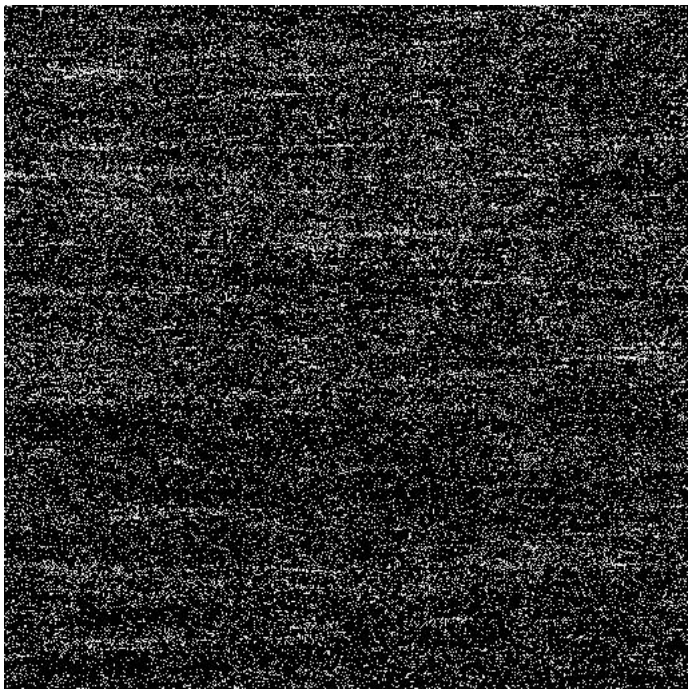

Figure 3. The lowest redshift hyperrectangle of the 118279 galaxies selected from Figure 1. Note: Relative to Figure 2, this and subsequent figures are displayed with a 90 degree clockwise rotation.

galaxy positions, and redshifts, were mapped into a data volume of dimensions $64 \times 64 \times 4$. Figures 3, 4, 5 and 6 show the four redshift planes (or more accurately, hyperrectangles) resulting from this coding.

We embedded the positional data into a set of image planes, by convolving with a Gaussian of FWHM = 15.0 (equivalent to a variance of 40.5758). Figure 7 shows the result of doing this, in the case of Figure 3.

3. SPATIAL CLUSTERING

Background on the approach pursued here can be found in Stanford[7] and Stanford and Raftery.[8]

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. In this paper, Y_i is a point in a 4-dimensional redshift space. 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 implies randomness, and a large value implies 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:

$$p(X_i = j \mid N(X_i), \phi) = \frac{\exp(\phi U(N(X_i)), j)}{\sum_k \exp(\phi U(N(X_i)), j)} \tag{1}$$

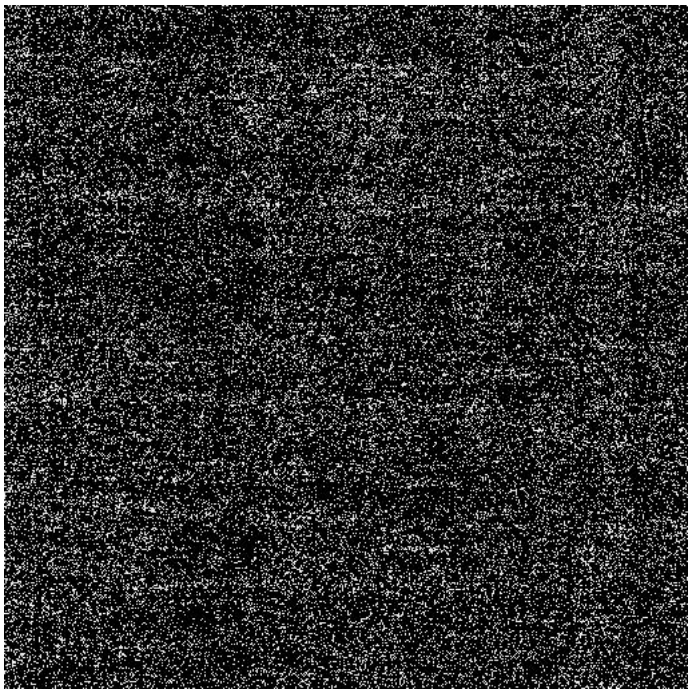

Figure 4: The second redshift hyperrectangle of the 118279 galaxies.

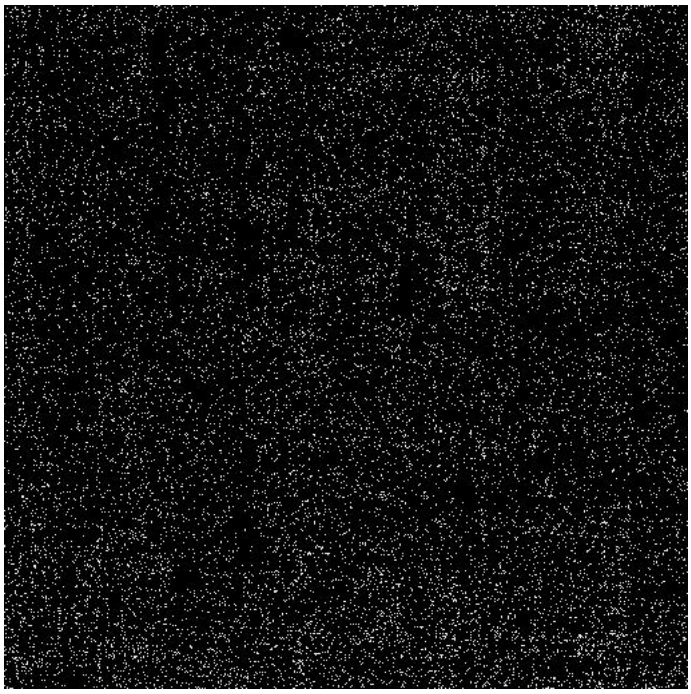

Figure 5: The third redshift hyperrectangle of the 118279 galaxies.

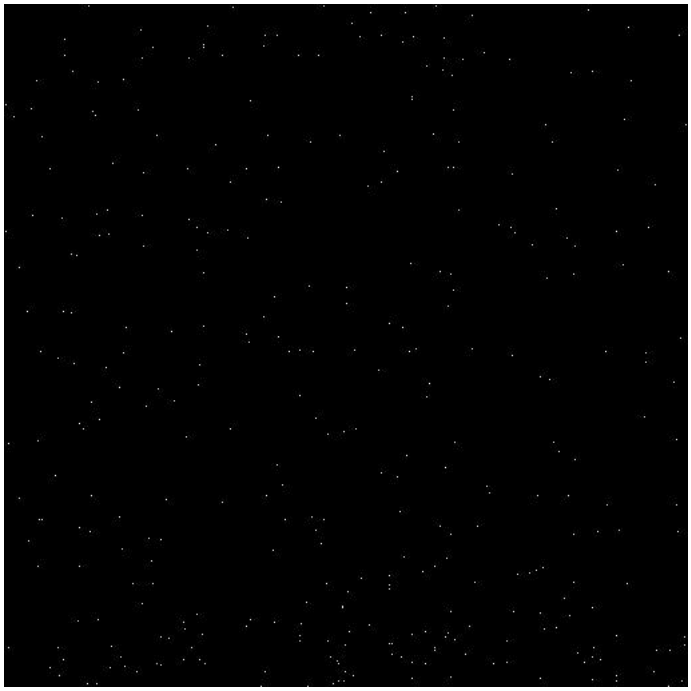

Figure 6: The highest redshift hyperrectangle of the 118279 galaxies.

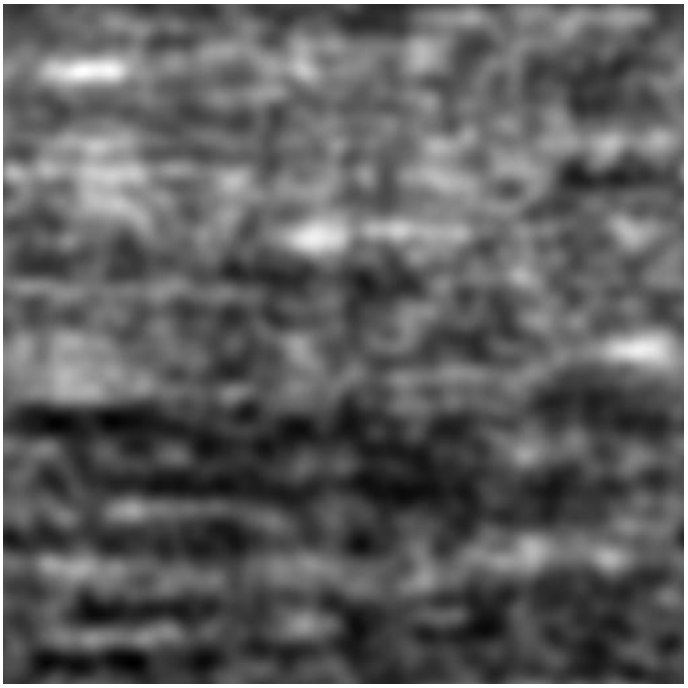

Figure 7: Smoothed version of Figure 3, the lowest redshift hyperrectangle of the 118279 galaxies.

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 vector μ_j and variance and covariance matrix V_j. 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, V) 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[9] 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. Therefore to initialize X we use a Gaussian mixture model fit to the marginal density of a selected band (the fourth or last one was used here). For this, we use the EM (expectation-maximization) iterative algorithm, which is a simple version of the full segmentation algorithm.

Segmentation Algorithm

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

Step 1: Update $\hat{\theta} =$ argmax $f(Y \mid \hat{X})$ based on maximum likelihood estimates of $\theta_j = \{\mu_j, V_j\}$ 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, if $\hat{\phi}$ goes negative, then we reset it to zero. In all calculations, we exclude boundary pixels from consideration. Step 1 is one step of Besag's ICM algorithm.

4. MODEL SELECTION

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 Stanford and Raftery[8]).

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, $\text{PL}(X)$, is taken as a proxy for $f(X)$. The pseudo-likelihood, introduced in Besag,[10] restricts where the integrated likelihood is defined. We have

$$\text{PL}(X, \phi) = \Pi_i p(X_k \mid N(X_i), \phi) = \Pi_i \frac{\exp(\phi U(N(X_i)), X_i)}{\sum_k \exp(\phi U(N(X_i)), k)} \tag{2}$$

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

$$L(Y_i \mid X_i) = \sum_j f(Y_i \mid X_i = j) p(X_i = j)$$

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:

$$L(Y_i \mid N(\hat{X}_{-i})) = \sum_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i)) \tag{3}$$

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 equations 1 and 2.

Let us consider a model as M_k, a function of the number of classes, k, and defined by the estimates of means, variances and covariances, and other properties related to homogeneity parameter ϕ and neighborhood N. By Bayes theorem, the posterior probabilities are:

$$p(M_k \mid Y) = \frac{p(Y \mid M_k) p(M_k)}{\sum_{h=\mathrm{Kmin}}^{\mathrm{Kmax}} p(Y \mid M_h) p(M_h)}$$

Since each number of clusters is considered equi-likely, the prior $p(M_k)$ is constant. The model prior, $p(M_h)$, can be ignored as not effecting the final conclusions unduly. The term $p(Y \mid M_k)$ is the integrated likelihood of model M_k. The Schwarz criterion,[11–13] or Bayes information criterion, approximates the integrated likelihood as $p(Y \mid \hat{\theta}, M_k)$, for which we use equation 3.

Given our use of pseudo-likelihoods for all pixels, this criterion is termed the pseudo-likelihood information criterion, PLIC.[7, 8]

Our model seletion algorithm in practice entails looking at values of $k = 1, 2, \ldots, 20$, and finding the first local maximum.

5. RESULTS

The clustering arrived at (i) takes account of spatial influence, using the Markov model, and (ii) makes use of a Gaussian model, of varying means, variances and covariances, for the classes. In particular, we note that the clustering itself is two-dimensional. The 2D map of labels or cluster numbers is derived from the iterative optimization procedure used.

The approximate Bayes factor provided by the PLIC (pseudo-likelihood information criterion) is shown for increasing numbers of clusters in Figure 8. Already a 2-cluster solution is very satisfactory. Values of PLIC climb to a relatively stable plateau in the region of an 18-cluster solution. Figures 9 and 10 show, respectively, the 2-cluster and 18-cluster solutions. Further clarification of Figure 10 is provided by Figure 11. A number of interesting alignments and patterns are revealed in this figure.

Figure 11 is highly influenced by one band in the data used which, we recall, consists of four redshift hyperrectangle bands. The influential hyperrectangle is the last one, shown in Figure 6. This is clearly shown when we superimpose Figure 11 and (rotated) Figure 6: see Figure 12. This explains our result very well. However an interesting question is then raised as to why the other redshift hyperrectangles, displayed in Figures 3, 4 and 5, played such a subsidiary role. The answer to this lies in the fact that the last redshift hyperrectangle (Figure 6) was used to initialize the clustering. Relative to it, the other redshift hyperrectangles provide relatively little information of value for the clustering.

6. CONCLUSIONS

From the astronomical point of view, it has to be stressed that the largest available surveys do not provide estimates of the redshifts (neither spectroscopic nor photometric) and therefore it is crucial to investigate the performances of our method (as well as of any other method) on 2D data based on RA and Dec only. We therefore plan to evaluate the performances of our method on 2D data by comparing the results with those obtained from the 3D data. The higher accuracy of the higher contrast 3D catalog will allow us to estimate on objective grounds the completeness and the fraction of spurious clusters detected in the 2D data. The method will therefore be applied to available public surveys such as the Digital Palomar Sky Survey and the whole SDSS.

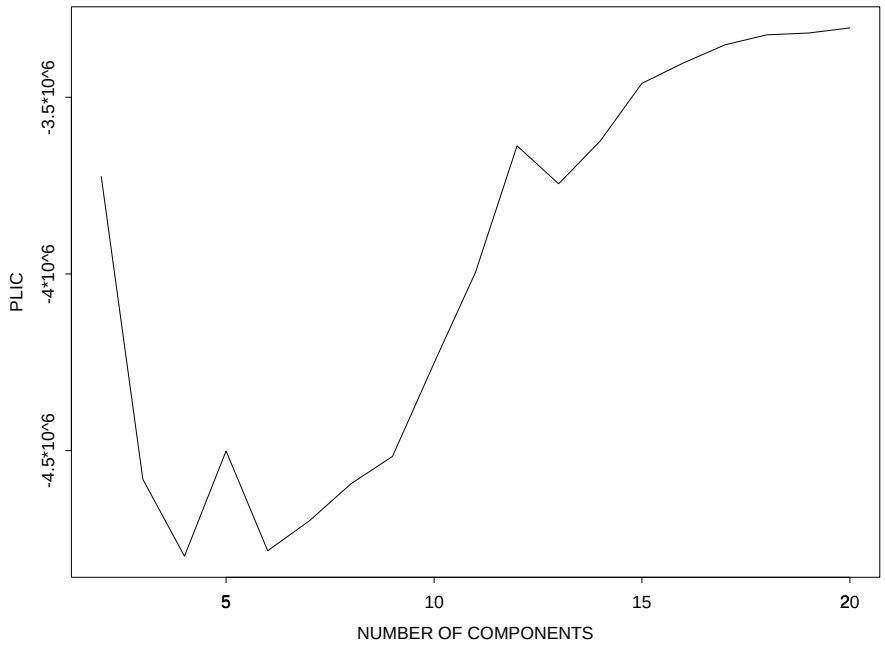

Figure 8. Pseudo-likelihood information criterion for 2 to 20 components. The general rule in reading this is to favor the first relative maximum. See text for discussion of why we look further at the 2-component and the 18-component results.

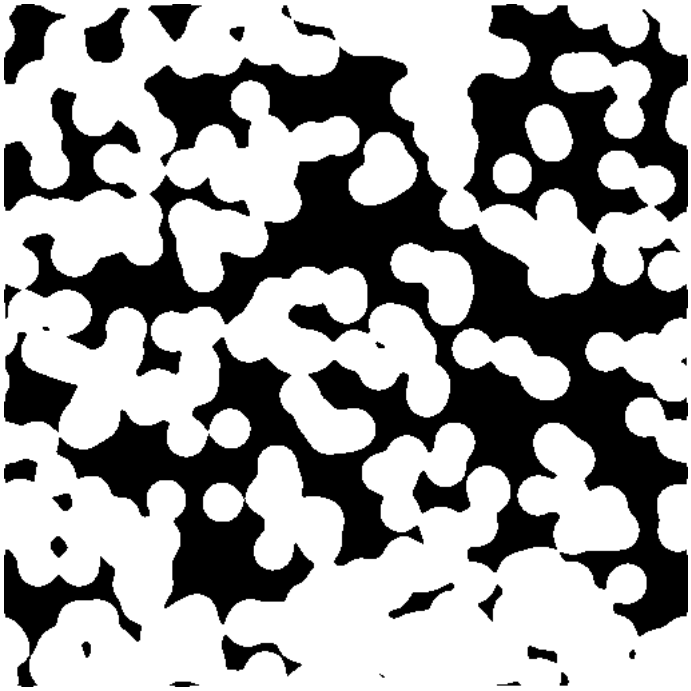

Figure 9: Two-segment solution.

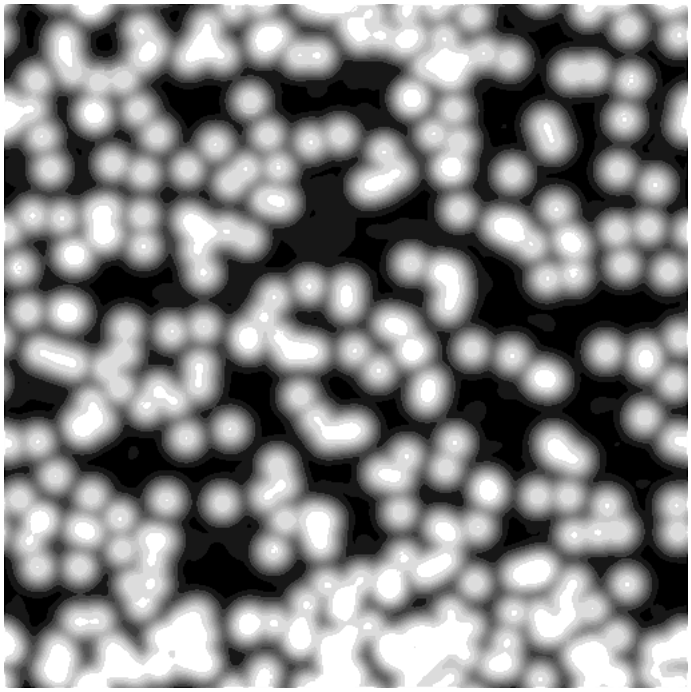

Figure 10: Eighteen-segment solution.

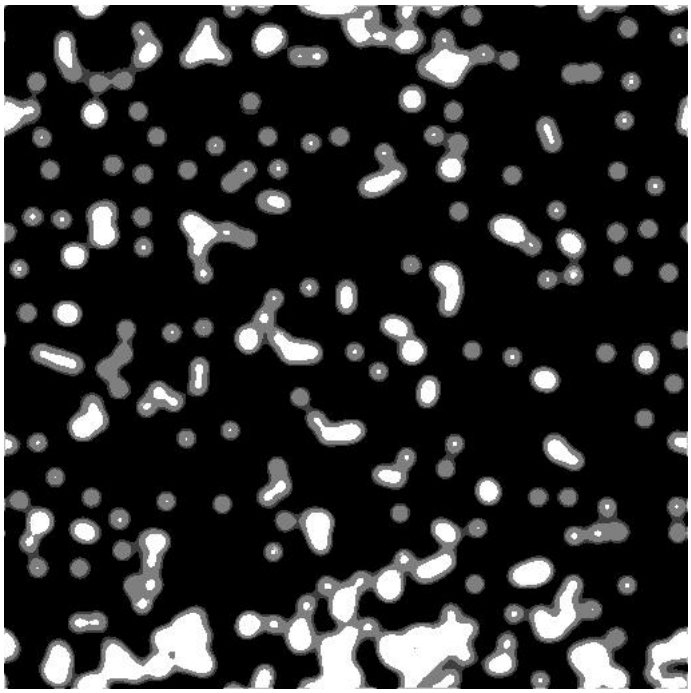

Figure 11. Eighteen-segment solution showing only clusters with labels greater than sequence number 14. These clusters clearly correspond to high level mean vectors. We recall that the clustering is carried out in multidimensional space, i.e. the 4-dimensional space defined by the redshift bands used.

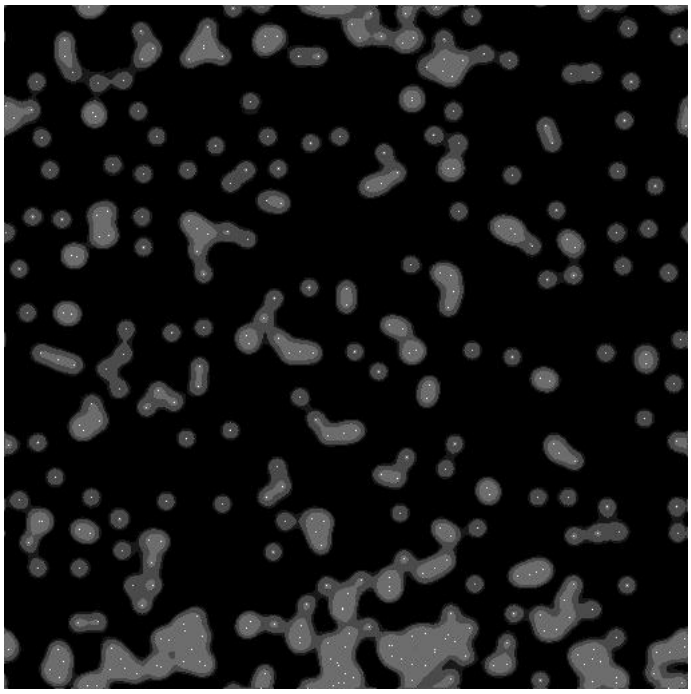

Figure 12. A superimposed representation of the clusters shown in Fig. 11, vis à vis the (suitably rotated) highest redshift hyperrectangle shown in Fig. 6. The former are the extended regions, with four levels of grey. The latter are the points. See text for discussion.

From the point of view of the methods used, two major lines of investigation are currently envisaged.

Firstly, we will seek to use a wavelet transform to expedite the cluster-finding. In Gao et al.[14] a coarse and low-resolution segmentation is found, and this is further refined based on a multiresolution transform of the data. It is not clear however that the work of these authors leaves scope for general application. Our desire, at all events, is to be able to avail of a Bayes factor model selection methodology.

A second line of enquiry relates to the need to have more general cluster shapes than Gaussian with arbitrary variances and covariances. In the discussion following Djorgovski,[15] presented at the conference on *Statistical Challenges in Modern Astronomy III*, Raftery referred to the usefulness of "sausage" shaped clusters being approximated by a (parsimonious) set of Gaussians. Djorgovsi et al. have pointed to the burgeoning need for this kind of cluster analysis capability in catalog and survey data mining. Raftery has pointed to the methodology which is capable of solving this problem.

If an optimally parsimonious set of Gaussians is to be fit to complicated cluster shapes in 3D and higher dimension catalogs, then we are dealing with a cluster tree model. A tree of Gaussians, optionally taking into account a Markov spatial model, is a powerful modeling tool, especially when complemented by hypothesis testing based on the Bayes factors. Initial work on Bayesian cluster trees can be found in Murtagh et al.[16]

Acknowledgments

This work is based on discussions initiated in the iAstro project (COST Action 283, Computational and Information Infrastructure in the Astronomical DataGrid, www.iAstro.org).

REFERENCES

1. A. Refregier, I. Valtchanov, and M. Pierre, "Cosmology with galaxy clusters in the XMM large scale structure survey," *Astronomy and Astrophysics* **390**, pp. 1–12, 2002.

2. R. Gal, R. de Carvalho, S. Odewhan, S. Djorgovski, R. Brunner, and V. Margoniner, "The northern sky optical cluster survey," in *Mining the Sky*, pp. 160–167, Springer-Verlag, 2002.
3. S. et al., "Sloan digital sky survey: Early data release," *Astrophysical Journal* **123**, pp. 485–548, 2002.
4. R. Tagliaferri, G. Longo, S. Andreon, S. Capozziello, C. Donalek, and G. Giordano, "Neural networks and photometric redshifts," *Astronomy and Astrophysics* , 2002. astro-ph/0203445, submitted.
5. F. Murtagh, *Multidimensional Clustering Algorithms*, Physica-Verlag, 1985.
6. J. S. F. Murtagh and M. Berry, "Overcoming the curse of dimensionality in clustering by means of the wavelet transform," *Computer Journal* **43**, pp. 107–120, 2000.
7. D. Stanford, *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Patterns.* PhD thesis, Department of Statistics, University of Washington, 1999.
8. D. Stanford and A. Raftery, "Determining the number of colors or gray levels in an image using approximate Bayes factors: the pseudolikelihood information criterion (PLIC)," tech. rep., Department of Statistics, University of Washington, 2001.
9. J. Besag, "Statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B* **48**, pp. 259–302, 1986.
10. J. Besag, "Statistical analysis of non-lattice data," *Statistician* **24**, pp. 179–195, 1975.
11. G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics* **6**, pp. 461–464, 1978.
12. R. Kass and A. Raftery, "Bayes factors," *Journal of the American Statistical Association* **90**, pp. 773–795, 1995.
13. H. M.H and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association* **96**, pp. 746–774, 2001.
14. J. Gao, J. Zhang, and M. Fleming, "Novel technique for multiresolution color image segmentation," *Optical Engineering* **41**, pp. 608–614, 2002.
15. S. 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 *SPIE Proceedings Vol. 4472*, J. Starck and F. Murtagh, eds., pp. 43–52, SPIE, 2001.
16. F. Murtagh, A. Raftery, and J. Starck, "Bayesian inference for multiband image segmentation via model-based cluster trees," tech. rep., Computer Science, Queen's University Belfast, 2002.