Dirichlet Process Mixture Models for Autonomous Habitat Classification

Daniel M. Steinberg, Oscar Pizarro, Stefan B. Williams & Michael V. Jakuba
Australian Centre for Field Robotics (ACFR)
University of Sydney,
NSW 2006, Australia
Email: d.steinberg, o.pizarro, stefanw, m.jakuba@acfr.usyd.edu.au

***Abstract*—There is a need for truly unsupervised approaches to understanding acquired data in autonomous exploratory missions with minimal, or zero, bandwidth communication. This paper presents results of using a Bayesian non-parametric Dirichlet Process mixture model – the Infinite Gaussian Mixture Model (IGMM) – for the classification of benthic habitats. The IGMM is trained completely autonomously, without being given labelled data, or knowing the number of habitats present. It is able to infer the number of habitats present in the training data, and is also able to infer the presence of habitats in the test data that were not present in the training data. This is a powerful model for entirely autonomous labelling of benthic datasets, and will be used as the basis of completely autonomous approaches to understanding data in the future.**

I. Introduction

Autonomous underwater vehicles (AUVs) are an effective tool for gathering large scientific datasets in dangerous and inaccessible environments. However, this autonomy is often limited to performing fixed, pre-planned surveys [1], [2]. When there is little a priori information, as in the case of exploratory missions, the resulting data products may be of little scientific value, especially if the processes of interest are not fully captured. To overcome this limitation these autonomous agents should *adapt* their mission plans to suit scientific goals.

Benthic ecologists are often interested in characterising transition zones between distinct habitats. In the case of exploratory missions, an AUV may cover several transitions between an unknown number of habitats. In order to use survey time effectively, it would be highly desirable that the AUV were capable of recognising distinct habitats (without prior training data) and perform additional sampling in transition zones.

Adaptive sampling requires real-time processing and interpretation of data that is being gathered by the AUV; so subsequent actions can be taken that are likely to increase the value of the resulting data products. A method of achieving this is for the AUV to represent its environment as a probabilistic model that can be used to infer data values at unobserved locations. A higher-level decision process can then weigh the potential benefits and costs of a set of possible sampling actions using this model.

We believe there are two aims of autonomous habitat modelling for adaptive sampling. The first is to classify a habitat given processed sensor data of the environment. The second is to infer habitat classes at unobserved locations. More formally, we have three variables,

- $\mathbf{x}$ is a d-dimensional spatial location, typically in Euclidean space $\mathbf{x} = \{x, y, z\}$.
- $\mathbf{y}$ is an observation of the seafloor or substrate from processed sensor data (optical or acoustic). For our purposes $\mathbf{y} \in \mathbb{R}$.
- $\mathbf{z}$ are the discrete, unobservable, classes of habitat that generate $\mathbf{y}$.

The aim of classification is to find a model that allows us to find the posterior density of habitat labels given the sensor observations, $p(\mathbf{z}|\mathbf{y})$. The aim of inference is to then find the posterior density of habitat labels given observed and *unobserved* locations, $p(\mathbf{z}|\mathbf{x})$. So, to generalise, the overall aim of autonomous habitat modelling is then to estimate the posterior density of the habitat labels given locations and noisy sensor observations, $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$.

Various probabilistic and non-probabilistic models exist that fulfil the aims of classification and inference. Some examples of common *supervised* classification models are logistic and probit regression, multinomial logit regression, and Support Vector Machines [3]. These approaches to classification may not be suitable for detecting and classifying habitats in any given environment. There may be insufficient, or a complete lack of labeled habitat data from which to train such classifiers. Many *unsupervised* machine learning techniques for classification rely upon assumptions about the data, such as the number of classes present, e.g. k-means, or what defines a class. This may also be unknown a priori. Conversely, *non-parametric* approaches make few assumptions about the form of the distribution the data is drawn from. Models created using non-parametric techniques are largely inferred from structures that are present in the data itself. Consequently this approach has the potential to lend itself to truly autonomous applications.

In this paper we present a Bayesian non-parametric model that autonomously learns $p(\mathbf{z}|\mathbf{y})$, without being given the class labels, $\mathbf{z}$, or knowing how many classes are present. Sections II and III present the model and the training and classification algorithms. Section IV presents the datasets and the results of the autonomous habitat labelling. Finally Sections V and VI present a discussion and concluding remarks.

978-1-4244-5222-4/10/$26.00 ©2010 CROWN

II. Dirichlet Process Mixture Models

The general form of the non-parametric model used in this work is the Dirichlet Process Mixture Model (DPMM) [4], [5]. It has the appealing property that the number of mixtures, or clusters that are present in a dataset, does not need to be known *a priori*. The number of mixtures is dependant on an aggregation parameter, α. This parameter is inferred from the data, and generally only allows the number of mixtures to increase as more data is observed. In the case considered here, a mixture component should correspond to a class of habitat.

One way to view a DPMM is as a finite mixture model that has been taken to the limit of having an *infinite* number of mixture components. For instance, a Bayesian formulation of a finite mixture model with associated priors [6],

$$\begin{aligned} y_i|z_i, \boldsymbol{\theta} &\sim F(\theta_{j=z_i}) \\ z_i|\boldsymbol{\pi} &\sim Discrete(\pi_1, \ldots, \pi_K) \\ \theta_j &\sim G_0 \\ \boldsymbol{\pi} &\sim Dir(\alpha/K, \ldots, \alpha/K). \end{aligned}$$

Here $\boldsymbol{\theta} = \{\theta_j\}$ is a vector of the distribution parameters for the mixture component distributions $F(\cdot)$, and G_0 is the prior distribution over the parameters for each mixture, θ_j. It is assumed that $y_1, \ldots, y_n$ are exchangeable and i.i.d. Now, if we take $K \to \infty$, and integrate over the mixture weights, $\boldsymbol{\pi}$, we get a DPMM [6], [7],

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G &\sim \mathcal{DP}(G_0, \alpha). \end{aligned}$$

G is a (posterior) mixing distribution over the parameters θ_i. The mixing distribution is now over the θ_i rather than the weights, $\boldsymbol{\pi}$. Although there is a θ_i associated with each observation, y_i, the distributions drawn from the Dirichlet Process, $\mathcal{DP}(\cdot)$, are discrete, thus 'countably infinite', [8], [6]. There is also a non-zero probability associated with samples of θ_i being the same [9], [7]. This results in mixture models having an infinite number of possible mixture components, but with only a finite amount of these actually having one and more observations belonging to them. The number of θ_i that are the same is determined by the aggregation parameter, α.

III. The Infinite Gaussian Mixture Model

The form of DPMM used in this work is the *univariate* Infinite Gaussian Mixture Model (IGMM) [10]. Following the preceding discussion, it is simply a Bayesian formulation of a finite Gaussian mixture model, which has been generalised to have an infinite number of mixture components,

$$p(y_i|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_j^{\infty} \pi_j \mathcal{N}(y_i|\theta_j)\,. \tag{1}$$

where $\theta_j = \{\mu_j, s_j^{-1}\}$, and s_j is the precision, or inverse variance. This is sampled according to

$$\theta_j \sim p(\Theta)\,,$$

which is a Normal-Gamma shaped prior distribution with hyperparameters $\Theta = \{\beta, w, \lambda, r\}$. Specifically,

$$\mu_j \sim \mathcal{N}\big(\lambda, r^{-1}\big) \quad \text{and} \quad s_j \sim \mathcal{G}\big(\beta, w^{-1}\big)\,.$$

These hyperparameters also have distributions from which they are drawn, most are conjugate except for β, which requires Adaptive Rejection Sampling [11]. We can also sample the mixture weights according to a Dirichlet distribution [12],

$$\boldsymbol{\pi} \sim Dir\,(n_1 + \alpha/K, \ldots, n_K + \alpha/K)\,,$$

where K is the number of components with one or more observations assigned to them. However this is not explicitly in the formulation of the IGMM [10]. The hierarchy of this model is illustrated in Figure 1.

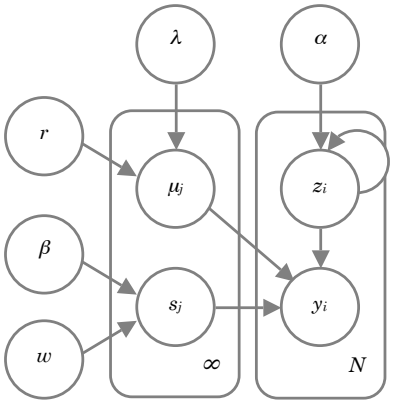


Fig. 1: Graphical model of the IGMM (adapted from [10] & [12]). The hyperparameters are α and $\Theta = \{\beta, w, \lambda, r\}$. Each mixture component has its own parameters, $\theta_j = \{\mu_j, s_j^{-1}\}$, drawn from the hyperparameter space.

Training an IGMM is typically done using Gibbs sampling over the class or mixture labels, z_i, in the following manner,

$$p(z_i = j|\mathbf{z}_{-i}, \alpha, \theta_j) \propto \frac{n_{-i,j}}{n - 1 - \alpha} \mathcal{N}(y_i|\theta_j)\,, \tag{2}$$

$$p(z_i \neq j\ \forall j|\mathbf{z}_{-i}, \alpha, \Theta) \propto \frac{\alpha}{n - 1 - \alpha} \int \mathcal{N}(y_i|\theta_j)\, p(\theta_j|\Theta)\ \mathrm{d}\theta_j. \tag{3}$$

Here z_i indicates the assignments of observations of rugosity, y_i to an *existing* mixture component j. The subscript $-i$ means all observations except i, so $n_{-i,j}$ means the number of observations, not including i, that belong to mixture component j. The probability of an observation belonging to an existing Gaussian mixture component is given by (2). The probability of the observation belonging to a new component is given by (3). Unfortunately the integral in (3) is intractable, so the prior is sampled to generate an estimate of the probability of generating a new class [6], [10]. Once a new observation is classified, the model hyperparameters and parameters are updated. Specifically, Algorithm 8 from [6] is used – for more detail see [10] and for sampling α see [13].

The Gibbs sampler is run until apparent convergence, and since we are using the IGMM for classification, only *one* sample of the IGMM is used as the classification model. A few low weight mixtures may still remain in the chosen sample, so one hard assignment Expectation-Maximisation (EM) iteration is run over the observation labels. This makes the IGMM more reasonable for classification. The 'E' step performed for every observation label is:

$$z_i = \max \left\{ \frac{\pi_1 \mathcal{N}(y_i|\theta_1)}{Z}, \ldots, \frac{\pi_K \mathcal{N}(y_i|\theta_K)}{Z} \right\}, \qquad (4)$$

where Z is a normalising constant,

$$Z = \sum_{j}^{K} \pi_j \mathcal{N}(y_i|\theta_j)\,,$$

and the weights for this 'E' step (only) are calculated as $\pi_j = n_j/n$. The 'M' step consists of updating the model parameters and hyperparameters in the same manner as we do for Gibbs sampling.

Classification is then performed in a manner consistent with the concept of DPMMs, so uses (2) and (3) to assign new observations to labels, rather than using the traditional GMM 'responsability' assignment [3]. The parameters and hyperparameters of the model are also updated as new test observations are observed. This allows *test* observations to form new mixture components if there is low probability of them belonging to those already existing.

IV. Results

Two datasets obtained by stereo cameras on ACFR's Sirius AUV were used to test the unsupervised learning, and classification ability of the IGMM. The first dataset was obtained on the Great Barrier Reef, Queensland, Australia, and the second on Scott Reef in northern Australia. The observation input to the IGMM is a visually based benthic rugosity metric. All of the data is fully geo-referenced by a visual information-form SLAM filter [14].

The Great Barrier Reef dataset consisted of two sparse grids at different depths. The deeper grid, Figure 2a, was mainly over sand habitats, Figure 2b and Figure 2c, with a few rocky outcroppings present. The shallower grid, Figure 2d, was mainly over a habitat with a rocky substrate, Figure 2e, again with a few rocky outcroppings present, Figure 2f. There were also many images, in both grids, where a mixture of these classes were apparent.

The Scott Reef dataset is a dense, fully overlapping, grid as shown in Figure 3. This dive features clear transitions between dense coral cover, barren sand and an intermediate, partially populated substrate class, shown in Figure 3a and Figure 3b.

This section presents the results of training the IGMM over a subset of the habitats present in the datasets, and then using the rest of the data as test data for classification. The aims are to verify whether training can reasonably label habitat classes, and to test whether additional habitats can be recognised during classification.

A. The Rugosity Habitat Descriptor

The geo-referenced stereo imagery provided by the AUV can be used to generate fine-scale bathymetric reconstructions in the form of 3D triangular meshes [15]. From these meshes, it is possible to derive multi-scale terrain complexity measures of rugosity [16]. Rugosity, y_{rug}, is essentially the scalar ratio between the (draped) benthic surface area, and this area projected onto a plane of best fit. A flat surface has a rugosity of 1, while more complex structures have higher values. Rugosity calculated on a per-stereo pair basis was used as the sensor observation inputs to the IGMM, and proved to be very effective at discriminating between habitats. Each stereo pair represents an area of about one-by-one meter on the seafloor.

Histograms of rugosity for a single habitat tend to be distributed in a log-normal fashion, so we apply the transformation:

$$y = C \cdot \log(y_{rug} - 1), \qquad (5)$$

where C is an arbitrary scaling factor, so data for habitats have Gaussian shapes. Training data and complete dataset histograms of the rugosity descriptor for the Great Barrier Reef and Scott Reef datasets used are shown in Figure 4 and Figure 6.

B. Great Barrier Reef Results

The last 2000 rugosity observations in the Great Barrier Reef dataset were used as training data. This training data is made up of a few horizontal transects from the deeper grid in Figure 2a. Only the sand habitats shown in Figure 2b and Figure 2c, and mixtures of sand with some rocky substrate, were present. The rest of the dataset (23000 observations) made up the test data.

After training, the barren sand habitat was labeled as the red Gaussian in Figure 4c, the sand habitat the blue, a mix of sand and rocky substrate was labelled green, with some data points labeled cyan. After the IGMM classification algorithm was run over the entire test set, the rocky substrate was given the cyan label, and rocky outcroppings were labelled purple, Figure 4d.

All of these habitats were present in the deeper grid, and the transitions from barren sand to rocky outcroppings, then to a mixture of barren sand with sparse rocky outcroppings were correctly identified, as shown in Figure 5a. The shallower grid was mainly rock substrate with a few outcroppings, with some areas where the rock substrate had accumulated sand. These areas were also correctly identified as shown in Figure 5b.

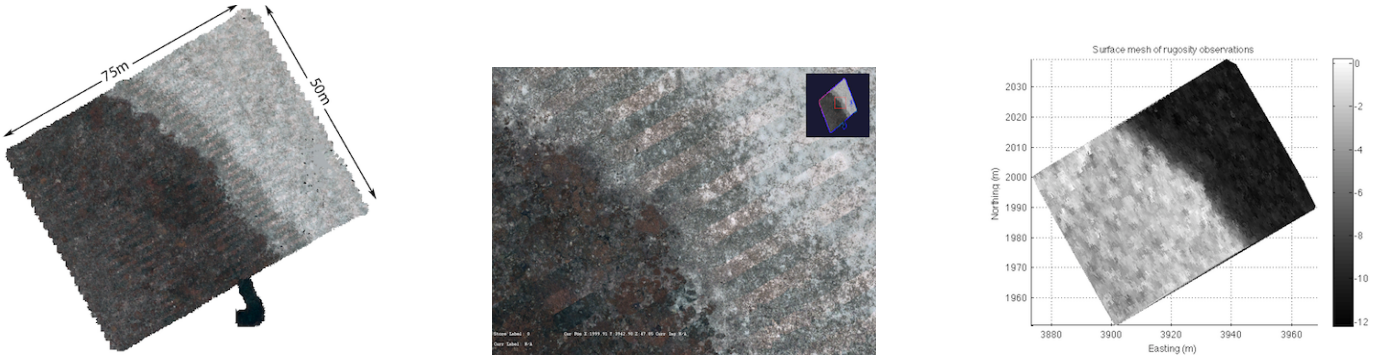
C. Scott Reef Results

For the Scott Dataset we used approximately 600 rugosity observations of the coral habitat as training data, as shown in Figure 7a. The remaining full transects of the dataset serve as the test data (9000 or so observations), Figure 7b. Training resulted in the coral (blue Gaussian) and a small amount of the intermediary mixture class (green Gaussian) being identified, Figure 6c.

(a) Deeper grid (b) Barren sand (c) Sand

(d) Shallower grid (e) Rock substrate (f) Rock outcropping

Fig. 2: Great Barrier Reef dataset consisting of a deep (a) and shallow (d) connected grids. Examples of the habitats surveyed in the deep grid are shown in (b) and (c), while (e) and (f) show sample habitats from the shallow grid.



(a) (b) (c)

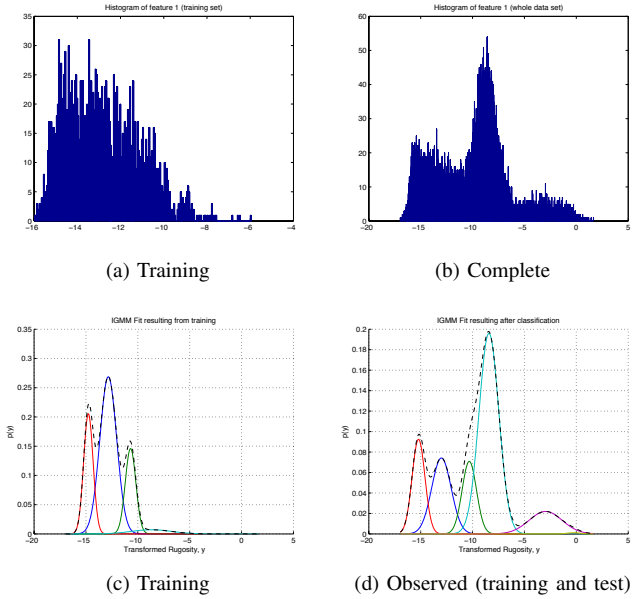
Fig. 3: Scott Reef dataset. (a) Image reconstruction of the dense survey, consisting of 50 parallel tracklines, each 75m long and spaced one meter apart. (b) Close up showing the three habitat types. (c) Surface mesh of the rugosity descriptor for the dense survey.

When the classifier was run over the test data, it was able to infer and classify observations from all habitat classes, Figure 6d. However, we can see that the IGMM has inferred the sand habitat consists of two labels – red and cyan. The cyan Gaussian is entirely contained by the red Gaussian, which is an unusual choice for a label.

Classification has also been performed for these datasets with *no* prior training. However in many cases an over-representation of mixtures results. It seems that some prior training of the IGMM is still necessary in order to properly infer the amount of aggregation in the data. This result demonstrates that the aggregation parameter can be reasonably inferred from a small amount of training data, in which not all habitats are represented, enabling the detection of new habitats during a mission.

V. Discussion

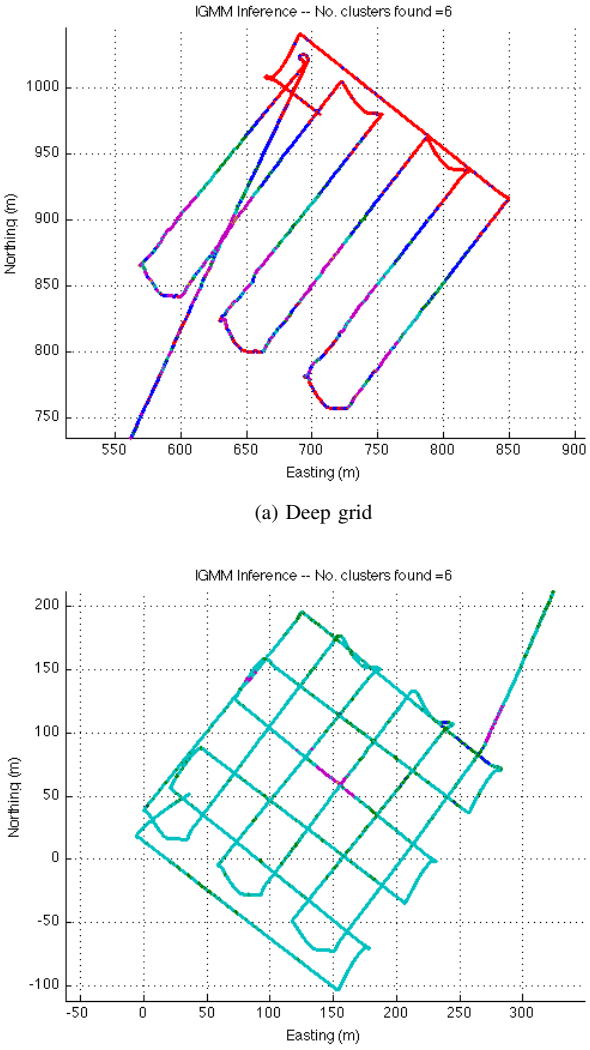
The IGMM training algorithm used in this paper, Gibbs sampling, is primarily intended for density estimation rather than classification [10]. Usually for density estimation multiple samples of the IGMM's parameters are combined linearly to then form a predictive distribution. There is no clear way to do this for classification since we need to have only one mixture component correspond to each habitat class. We



(a) Training (b) Complete

(c) Training (d) Observed (training and test)

Fig. 4: Histograms and learned distributions of the Great Barrier Reef dataset. The black dashed line is the estimated posterior density.



(a) Deep grid

(b) Shallow grid

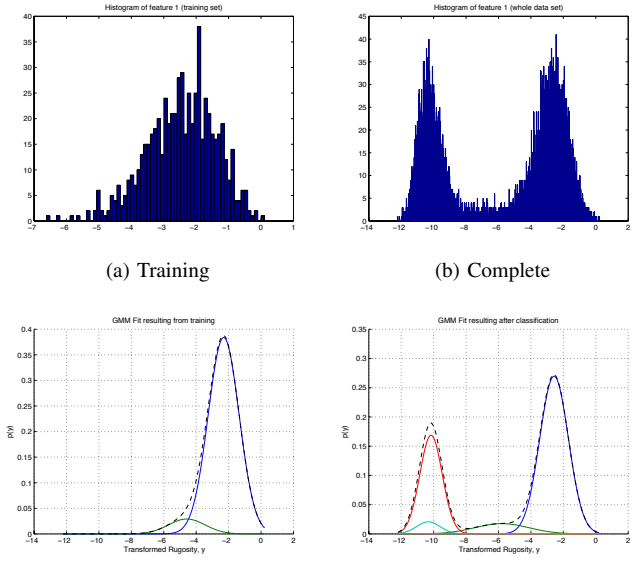
Fig. 5: Great Barrier Reef classification results.

are then limited to only using one sample of the IGMM for classification, with no guarantees that this sample will be appropriate for classification. Often low weight mixture components and mixture components with arbitrarily close means are present in this sample. The low weight mixtures are a normal part of Gibbs sampling. They will tend to only have a short lifetime if the parameters of the IGMM have reached convergence and there is a low likelihood associated with their existence. Inappropriate clustering is also a result of Gibbs sampling. This can be attributed to the incremental nature of Gibbs sampling, in that it only re-samples one observation label at a time [17]. This makes training the IGMM susceptible to getting trapped in local modes, resulting in the mixtures that have arbitrarily close means when only one mixture should be present, and also mixtures that do not split into multiple mixtures when they should. Furthermore, actually gauging whether the Gibbs sampling algorithm has converged or not is somewhat arbitrary. Monitoring the log-likelihood is not a robust way to gauge convergence since it does not increase in a monotonic fashion like maximum likelihood methods.

Removing low weight mixtures by performing a hard EM assignment iteration somewhat improves the IGMM for classification. However, this EM iteration does not solve the inappropriate clustering problem, and it does not actually sample from the true posterior distribution since it is a maximum likelihood approach. Learning the IGMM using variational techniques [18], or running Metroplis-Hastings 'split-merge' proposals for multiple observation labels interspersed with Gibbs sampling of the IGMM, [17], may be better, and faster, approaches to training.

Since classification of test observations is effectively equivalent to running one Gibbs sweep over the new observations, many of the problems mentioned previously are encountered. Mixtures with arbitrarily close means can manifest, which has happened in Figure 6d for the sand habitat – this should really be one cluster. This is probably because a label can never change once it has been sampled by the classifier. There should be a mechanism in place that can update all of the observation labels when appropriate. A variation of Jain's 'split-merge' Metropolis-Hastings algorithm [17] may be an appropriate mechanism. Low weight mixtures can also be a result of this classification procedure. Although having good prior knowledge of the aggregation parameter, α, from training helps to prevent these mixtures from appearing.

An alternative DPMM for classification is a generative classification model based upon mixtures of multinomial logit regression models [7]. This may be more appropriate for classification tasks. However it requires at least partially labelled data which is not appropriate for truly autonomous learning.



(a) Training (b) Complete

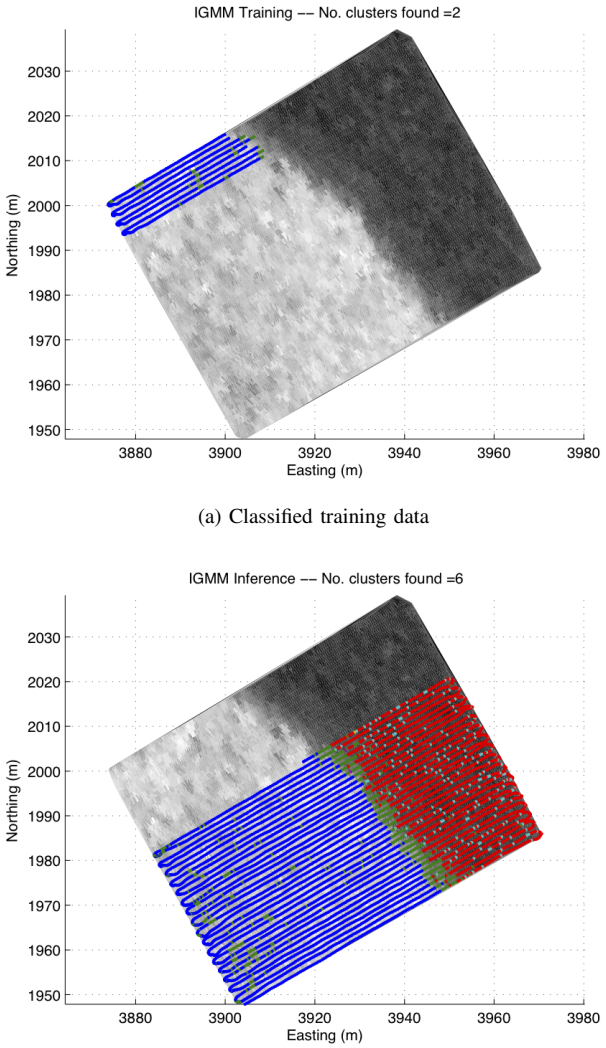
(c) Training (d) Observed (training and test)

Fig. 6: Histograms and learned distributions of the Scott Reef dataset. The black dashed line is the estimated posterior density.

VI. Conclusion and Future Work

This paper has shown that models based upon Dirichlet Process mixture models are appropriate for truly autonomous learning and classification of benthic habitats. While the implementation of these models in this paper can be made more appropriate for classification, we have demonstrated the first step in creating completely autonomous methods for adaptive exploration of the benthos. In the future we wish to evaluate whether variational or other Metropolis-Hastings methods are more appropriate for training and classification when using IGMMs.

Dirichlet Process mixtures of regression models also exist. An example is the Infinite Mixture of Gaussian Process Experts model [19] which is a generalisation of an IGMM that has been made spatially dependent. This model is primarily for regression in that it estimates $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$, and is not suited to classification tasks in its current form. However it may provide a starting point for making spatial DPMM classifiers where it is possible to model $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$. These sorts of models will then allow for inference of habitat classes at unobserved locations, they can be learned in-situ, and will inform a decision process for adaptive sampling.



(a) Classified training data

(b) Classified test data

Fig. 7: Scott Reef training and classification results.

Acknowledgments

This work is supported by the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC), the New South Wales State Government and the Integrated Marine Observing System (IMOS) through the DIISR National Collaborative Research Infrastructure Scheme. The authors of this work would like to thank Ariell Friedman for providing his rugosity code, and the Australian Institute for Marine Science and the Marine National Facility (CSIRO) for making ship time available to support this study. The crews of the R/V Solander and R/V Southern Surveyor were instrumental in facilitating successful deployment and recovery of the AUV. We also acknowledge the help of all those who have contributed to the development and operation of the AUV, including Duncan Mercer, George Powell, Matthew Johnson-Roberson, Ian Mahon, Stephen Barkby, Ritesh Lal, Paul Rigby, Jeremy Randle, Bruce Crundwell and the late Alan Trinder.

References

[1] D. R. Yoerger, M. Jakuba, A. M. Bradley, and B. Bingham, "Techniques for deep sea near bottom survey using an autonomous underwater vehicle," *The International Journal of Robotics Research*, vol. 26, no. 1, pp. 41–54, 2007.

[2] P. Rigby, "Autonomous spatial analysis using Gaussian process models," Ph.D. dissertation, The University of Sydney, 2008.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cambridge, UK: Springer Science+Business Media, 2006.

[4] T. Ferguson, "A Bayesian Analysis of some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[5] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974. [Online]. Available: http://www.jstor.org/stable/2958336

[6] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000. [Online]. Available: http://www.jstor.org/stable/1390653

[7] B. Shahbaba and R. Neal, "Nonlinear models using Dirichlet process mixtures," *Journal of Machine Learning Research*, vol. 10, pp. 1829–1850, August 2009.

[8] T. Ferguson, "Bayesian density estimation by mixtures of normal distributions," *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pp. 287–302, 1983.

[9] D. Blackwell and J. MacQueen, "Ferguson distributions via Polya urn schemes," *The Annals of statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[10] C. Rasmussen, "The Infinite Gaussian Mixture Model," *Advances in Neural Information Processing Systems*, vol. 12, pp. 554–560, 2000.

[11] W. Gilks and P. Wild, "Adaptive rejection sampling for Gibbs sampling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 41, no. 2, pp. 337–348, 1992.

[12] E. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[13] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[14] I. Mahon, S. Williams, O. Pizarro, and M. Johnson-Roberson, "Efficient view-based SLAM using visual loop closures," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1002–1014, 2008.

[15] M. Johnson-Roberson, O. Pizarro, S. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2009.

[16] A. Friedman, O. Pizarro, and S. B. Williams, "Rugosity, slope and aspect from bathymetric stereo image reconstructions," in *OCEANS 2010*. Sydney: IEEE, May 2010.

[17] S. Jain and R. Neal, "A Split-Merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 13, pp. 158–182, 2000.

[18] D. Blei and M. Jordan, "Variational methods for the Dirichlet process," in *Proceedings of the twenty-first International Conference on Machine Learning*. ACM New York, NY, USA, 2004.

[19] C. Rasmussen and Z. Ghahramani, "Infinite mixtures of Gaussian process experts," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002, pp. 881–888.