SANDIA REPORT

SAND2018-10833

Preliminary Results on Applying Nonparametric Clustering and Bayesian Consensus Clustering Methods to Multimodal Data

Maximillian G. Chen, Michael C. Darling, and David J. Stracuzzi

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA0003525.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2018-10833

Preliminary Results on Applying Nonparametric Clustering and Bayesian Consensus Clustering Methods to Multimodal Data

Maximillian G. Chen
Data Science and Applications Department
Sandia National Laboratories
mgchen@sandia.gov

Michael C. Darling
Data-Driven and Neural Computing Department
Sandia National Laboratories
mcdarli@sandia.gov

David J. Stracuzzi
Data-Driven and Neural Computing Department
Sandia National Laboratories
djstrac@sandia.gov

Abstract

In this report, we present preliminary research into nonparametric clustering methods for multi-source imagery data and quantifying the performance of these models. In many domain areas, data sets do not necessarily follow well-defined and well-known probability distributions, such as the normal, gamma, and exponential. This is especially true when combining data from multiple sources describing a common set of objects (which we call *multimodal* analysis), where the data in each source can follow different distributions and need to be analyzed in conjunction with one another. This necessitates nonparametric density estimation methods, which allow the data to better dictate the distribution of the data. One prominent example of multimodal analysis is *multimodal image analysis*, when we analyze multiple images taken using different radar systems of the same scene of interest. We develop uncertainty analysis methods, which are inherent in the use of probabilistic models but often not taken advance of, to assess the performance of probabilistic clustering methods used for analyzing multimodal images. This added information helps assess model performance and how much trust decision-makers should have in the obtained analysis results. The developed methods illustrate some ways in which uncertainty can inform decisions that arise when designing and using machine learning models.

Acknowledgements

This work was funded by the Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program.

Contents

List of Figures

Chapter 1

Introduction

Many machine learning and statistical models output point estimates that provide answers to analysis questions. However, quantifying the uncertainty of the obtained estimates so the trustworthiness of the estimates can be assessed usually does not accompany the models' outputs. This uncertainty information can be very important in decision-making. For example, in the problem of detecting malicious URLs using supervised classification methods, uncertainty quantification of the classifiers' results allows us to gauge the reliability of the classifiers, particularly in deployed settings where observations may not be well covered by the original training and validation data (Darling and Stracuzzi, 2018).

In our primary motivating example for this report, we want to segment images containing different types of information about a particular scene (which we call *multimodal imaging data*). We want to know the quality of the analysis of each individual image and the value of each image to aiding the understanding of the scene of interest, which can have significant ramifications on potential decisions (Stracuzzi et al., 2017a, 2018). In the problem of multimodal image analysis, or multimodal data analysis in general, one of the primary issues is that different data sources can obey different probability distributions or different data structures, such as discrete and continuous. Furthermore, even if two data sources follow similar distributions independently, when we combine the data sources, the combined data may follow a completely different and unknown distribution.

Nonparametric statistical methods allow for flexible distributional modeling of datasets that are not best described by parametric distributions. We report preliminary results in several areas of analysis when applied to the problem of segmenting multimodal imagery data. First, we report results on applying nonparametric clustering models to the imagery data and compute and visualize their clustering uncertainty results. Second, we report results on a Bayesian consensus clustering (BCC) method, which considers the source-specific clusterings for each data source to obtain a "consensus" clustering that factors in the multiple data sources at hand. We apply the BCC method to the imagery data and compute and visualize the associated clustering uncertainty results.

In chapter 2, we provide some background information on nonparametric methods and the challenges in applying these methods. In chapter 3, we discuss our implemented methods in details and the results we obtain when applying the methods to examples of multimodal imagery data and the clustering uncertainty results we compute. We implement three methods: nonparametric modal clustering method (Li et al., 2007), nonparametric mixture model (Benaglia et al., 2009a), and Bayesian consensus clustering (BCC) (Lock and Dunson, 2013a). In chapter 4, we discuss our results and the meanings and implications of the obtained results, including comments on directions for future work, and we finish with concluding remarks.

Chapter 2

Background

Recent advances in the technologies of computation and data collection has given rise to models that exploit data in order to describe events and systems of interest. Data sets are often incomplete, imprecise and, increasingly, too large for human comprehension — expanding the need to quantify uncertainty about known and unknown influencing factors. In the context of machine learning, uncertainty quantification asks the question: *What is the range of responses that a model might make given the available data and what are the relative likelihoods of each?*

Models that produce probabilistic labels provide a deeper understanding than an all-or-nothing class assignment. A probability value indicates goodness-of-fit for candidate labels, so a label predicted with high probability indicates that it fits a data point well given the model. An uncertainty analysis increases understanding a step further by providing a measure of a model's credibility when it assesses particular examples. Therefore, a high uncertainty (low credibility) model output indicates that alternate valid interpretations of the data point exist and the degree to which the model can distinguish among them.

In order to obtain probabilistic labels and quantify the uncertainty of the labeling, we need to determine the distribution of the data. In many data sets, the distribution of the data is not best described by parametric distributions, such as the normal, exponential, and gamma. This necessitates methods like *nonparametric* density estimation methods. The classic nonparametric method for density estimation is *kernel density estimation* (Rosenblatt, 1956; Parzen, 1962). Let $(x_1, x_2, ..., x_n)$ be a univariate independent and identically distributed (i.i.d.) sample drawn from some distribution with an unknown density f. The *kernel density estimator* of the shape of the function f is

$$\hat{f}_h(x) = \frac{1}{nh} K(\frac{x - x_i}{h}),$$

where K is the *kernel*, a symmetric, non-negative function that integrates to one, and $h > 0$ is a smoothing parameter called the *bandwidth*. There are a range of kernel functions that are commonly used, such as the Gaussian, uniform, triangular, biweight, triweight, and Epanechnikov. The bandwidth is often chosen by Silverman's rule of thumb (Silverman, 1986) when the Gaussian kernel is used, or by selecting the bandwidth using the criterion of the expected L_2 risk function, also termed the mean integrated squared error:

$$\text{MISE}(h) = E[\int (\hat{f}_h(x) - f(x))^2 dx].$$

Rather than be fixed, the length of the bandwidth can also be adapted and varied. *Adaptive* or *"variable-bandwidth" kernel density estimation* varies the size of the bandwidth depending upon either the location of the samples or the location of the test point. It is a particularly effective technique when the sample space is multi-dimensional. A common method of varying the kernel width is to make it inversely proportional to the density at the test point $\vec{x}$, written as $P(\vec{x})$:

$$h = \frac{k}{[nP(\vec{x}))]^{1/D}},$$

where D is the number of dimensions of $\vec{x}$ and k is a constant (Terrell and Scott, 1992).

Depending on the nature of the data set, challenges can arise with nonparametric density estimation. If the distribution of the data set is very complicated, it can be difficult to select an appropriate kernel function and to accurately vary the bandwidth of the kernel density function. Due to this, kernel density estimation can be computationally expensive. Furthermore, being able to characterize a nonparametric density estimate is often difficult. Even with estimating the mean and variance of the density estimate, as is often done with parametric distributions, they cannot characterize all of the changes in the shape of the distribution. The sampling distributions of resulting analysis estimates are related to the nonparametric density function. However, the nonparametric density function is not very well-defined, which makes interpreting and using uncertainty estimates difficult when the data follows a well-defined parametric distribution.

Another problem more specific to the analysis task of segmenting multimodal imagery data is *consensus clustering*, which obtains one clustering factoring in multiple data sets that contain different types of information over a common set of objects, such as the same number of pixels over multiple images covering the scene of interest. Existing consensus clustering methods either cluster each image separately and then use distance metrics to find a consensus clustering, or the multiple data sets are concatenated together and the combined data set is segmented. However, we are interested in quantifying the uncertainty of each image's clustering and uncertainty the uncertainty contributions of each image to the overall uncertainty we have about a particular scene. This suggests a method that segments each individual image and borrows information from the image-specific clustering and the consensus clustering to obtain the best clusterings and uncertainty estimates. We will discuss this in depth in section 3.3.

Chapter 3

Methods and Results

3.1 Nonparametric Modal Clustering

We utilize the nonparametric clustering approach of Li et al. (2007), which forms clusters by sample points that ascend to the same local maximum (mode) of the nonparametric density function through the use of two algorithms. The data is modeled using kernel density functions. Given a density estimate in the form of a mixture, a new algorithm, the Modal EM (MEM) finds an increasing path from any point to a local maximum of the density, that is, a hilltop. The clustering algorithm groups data points into one cluster if they are associated with the same hilltop. This approach is called modal clustering. A new algorithm, the Ridgeline EM (REM), is also developed to find the ridgeline linking two hilltops, which is proven to pass trough all the critical points of the mixture density of the two hills (Li et al., 2007).

For our purposes, we only use the MEM algorithm to find the clusters in the data. We use this algorithm to cluster multimodal imagery data and investigate methods for quantifying the uncertainty of the clustering results and combining the clusterings of multiple images. We review the relevant algorithms and visualization techniques, and then we present the results of their implementation on multimodal imagery data.

3.1.1 Modal EM (MEM) Algorithm

The MEM algorithm solves a local maximum of a mixture density by ascending iterations starting from any initial point. The algorithm is named Modal EM because it comprises two iterative steps similar to the expectation and maximum steps in the EM algorithm (Dempster et al., 1977). However, the objective of the MEM algorithm is different from the EM algorithm. The EM algorithm aims to maximize the likelihood of data over the parameters of an assumed distribution. On the contrary, the MEM seeks to find the local maxima, that is, modes, of a given distribution.

Let a mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$, where $x \in \mathcal{R}^d$, π_k is the prior probability of mixture component k, and $f_k(x)$ is the density of component k. Given any initial value $x^{(0)}$, MEM solves a local maximum of the mixture by alterating the following steps until a stopping criterion is met. Start with $r = 0$.

1. Let

$$p_k = \frac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}, k = 1, ..., K.$$

2. Update

$$x^{(r+1)} = \arg\max_x \sum_{k=1}^{K} p_k \log f_k(x).$$

The first step is the "Expectation" step where the posterior probability of each mixture component k, $1 \leq k \leq K$, at the current point $x^{(r)}$ is computed. The second step is the "Maximization" step. We assume that $\sum_{k=1}^{K} p_k \log f_k(x)$ has a unique maximum, which is true when the $f_k(x)$ are normal densities. In the special case of a mixture of Gaussians with common covariance matrix, that is, $f_k(x) = \phi(x|\mu_k|\Sigma)$, where $\phi(\cdot)$ is the pdf of a Gaussian distribution, we $x^{(r+1)} = \sum_{k=1}^{K} p_k \mu_k$.

3.1.2 Mode Association Clustering (MAC) Algorithm

Given a data set $\{x_1, x_2, ..., x_n\}$, $x_i \in \mathcal{R}^d$, a probability density function for the data is esitmate nonparametrically using Gaussian kernels. As the kernel density estimate is in the form of a mixture distribution, MEM is applied to find a mode using every sample point x_i, $i = 1, ..., n$, as the intial value for the iteration. Two points x_i and x_j are grouped into one cluster if the same mode is obtained from both. When the variances of Gaussian kernels increase, the density estimate becomes smoother and tends to group more points into one cluster. A hierarchy of clusters can thus be constructed by gradually increasing the variances of Gaussian kernels.

Let the set of data to be clustered by $S = \{x_1, x_2, ..., x_n\}$, $x_i \in \mathcal{R}^d$. The Gaussian kernel density estimate is formed:

$$f(x) = \sum_{i=1}^{n} \frac{1}{n} \phi(x|x_i, \Sigma),$$

where the Gaussian density function is

$$\phi(x|x_i, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - x_i)^t \Sigma^{-1}(x - x_i)).$$

We use a spherical covariance matrix $\Sigma = D(\sigma^2) = \text{diag}(\sigma^2, \sigma^2, ..., \sigma^2)$. The standard deviation is also referred to as the *bandwidth* of the Gaussian kernel.

With a given Gaussian kernel covariance matrix $D(\sigma^2)$, data are clustered as follows:

1. Form kernel density
$$f(x|S,\sigma^2) = \sum_{i=1}^{n} \frac{1}{n}\phi(x|x_i, D(\sigma^2)).$$

2. Use $f(x|S,\sigma^2)$ as the density function. Use each x_i, $i = 1, 2, ..., n$, as the intial value in the MEM algorithm to find a mode of $f(x|S,\sigma^2)$. Let the mode identified by starting from x_i be $\mathcal{M}_\sigma(x_i)$.

3. Extract distinctive values from the set $\{\mathcal{M}_\sigma(x_i), i = 1, 2, ..., n\}$ to form a set G. Label the elements in G from 1 to $|G|$. In practice, due to finite precision, two modes are regarded equal if their distance is below a threshold.

4. If $\mathcal{M}_\sigma(x_i)$ equals the $k$the element in G, x_i is put in the $k$the cluster.

In the basic version of the algorithm, the density $f(x|S,\sigma^2)$ is a sum of Gaussian kernels centered at every data point. However, the algorithm can be carried out with any density estimate in the form of a mixture. The key step in the clustering algorithm is the identification of a mode starting from any x_i. MEM moves from x_i via an ascending path, or figuratively, via hill climbing, to a mode. Points that climb to the same mode are located on the same hill and hence grouped into one cluster. We call this the *Mode Association Clustering (MAC)* algorithm.

3.1.3 Density Estimation

The density of each cluster is not explicitly modeled by MAC, but a pdf for each cluster can be obtained. These density functions facilitate soft clustering as well as cluster assignment of samples outside the data set. Denote the set of points in cluster k, $1 \leq k \leq |G|$, by C_k. The density estimate for cluster k is

$$g_k(x) = \sum_{x_i : x_i \in C_k} \frac{1}{|C_k|}\phi(x|x_i, D(\sigma^2)). \tag{3.1}$$

Because we do not assume a parametric form for the densities of individual clusters, this methods tends to be more robust and characterizes clusters more accurately when the attempted parametric assumptions are violated.

It is known in the literature of mixture modeling that if the density of a cluster is estimated using only known points assigned to this cluster, the variance tends to be underestimated, although the effect on clustering may be small. The under estimation of variance becomes more severe for poorly separated clusters, which often decay towards zero too quickly on leaving the cluster. We will see a similar phenomenon here with $g_k(x)$ having over fast decaying tails. A correction to this problem is to use soft instead of hard clustering. Every point is allowed to contribute to every cluster by a weight computed from the posterior probability of the cluster.

Under this spirit, we can make an ad-hoc modification on the density estimation. With $g_k(x)$ in (3.1) as the initial cluster density, compute the posterior of cluster k given each x_i by $p_{ik} \propto \frac{|C_k|}{n} g_k(x)$, $k = 1, .., |G|$, subject to $\sum_{k'=1}^{|G|} p_{i,k'} = 1$. Form the updated density of cluster k by

$$\tilde{g}_k(x) = \frac{\sum_{i=1}^{n} p_{i,k} \phi(x|x_i, D(\sigma^2))}{\sum_{i=1}^{n} p_{i,k}}.$$

3.1.4 Hierarchical MAC (HMAC) Algorithm

When the bandwidth σ increases, the kernel density estimate $f(x|S, \sigma^2)$ in (1) becomes smoother and more points tend to climb to the same mode. This suggests a natural approach for hierarchical clustering. Given a sequence of bandwidths $\sigma_1 < \sigma_2 < ... < \sigma_n$, hierarchical clustering is performed in a bottom-up manner. We start with every point x_i being a cluster by itself. The set of cluster representatives is thus $G_0 = S = \{x_1, ..., x_n\}$. This extreme case corresponds to the limit when σ approaches zero. At any bandwidth σ_l, the cluster representatives in G_{l-1} obtained from the preceding bandwidth are input into MAC using the density $f(x|S, \sigma_l^2)$. Note that the kernel centers remain at all the original data points although modes are identified only for cluster representatives when $l > 1$. The modes identified at this level form a new set of cluster representatives G_l. This procedure is repeated across all σ_l's. This hierarchical clustering algorithm is the *Hierarchical MAC (HMAC)* algorithm and corresponds to the mappings $x_i \to \mathcal{M}_{\sigma_1}(x_i) \to \mathcal{M}_{\sigma_2}(\mathcal{M}_{\sigma_1}(x_i)) \to \cdots$.

Denote the partition of points obtained at bandwidth σ_l by $\mathcal{P}_l$, a function mapping x_i's to cluster labels. If K clusters labeled $1, 2, ..., K$, are formed at bandwidth σ_l, $\mathcal{P}_l(x_i) \in \{1, 2, ..., K\}$. HMAC ensures that $\mathcal{P}_l$'s are nested, that is, if $\mathcal{P}_l(x_i) = \mathcal{P}_l(x_j)$, then $\mathcal{P}_{l+1}(x_i) = \mathcal{P}_{l=1}(x_j)$. Recall that the set of cluster representatives at level l is G_l. HMAC starts with $G_0 = \{z_1, ..., x_n\}$ and solves G_l, $l - 1, 2, ..., \eta$, sequentially by the following procedure:

1. Form kernel density
$$f(x|S, \sigma_l^2) = \sum_{i=1}^{n} \frac{1}{n} \phi(x|x_i, D(\sigma_l^2)).$$

2. Cluster G_{l-1} by MAC using density $f(x|S, \sigma_l^2)$. Let the set of distinct modes obtained be G_l.

3. If $\mathcal{P}_{l-1}(x_i) = k$ and the kth element in G_{l-1} is clustered to the k'th mode in $\mathcal{G}_l$, then $\mathcal{P}_l(x_i) = k'$. That is, the cluster of x_i at level l is determined by its cluster representative in G_{l-1}.

3.1.5 Visualization

In order to visualize clusterings for data of higher than two dimensions, we need to project the results into lower dimensions. Principal component analysis (PCA), a widely used linear projection method, is not designed to reveal clustering structures. Section 5 of Li et al. (2007) describe a novel linear projection method that reveals clustering structures.

Modal clustering provides an estimated density function and a prior probability for each cluster. Suppose K clusters are generated. Let the cluster density function of x, $x \in \mathcal{R}^d$, be $g_k(x)$, and the prior probability be π_k, $k = 1, 2, ..., K$. For any $x \in \mathcal{R}^d$, its extent of association with each cluster k is indicated by the posterior probability $p_k(x) \propto \pi_k g_k(x)$. Tp determine the posterior probability $p_k(x)$, under a given set of priors, it suffices to specify the discriminant functions $\log \frac{g_1(x)}{g_K(x)}, ..., \log \frac{g_{K-1}(x)}{g_K(x)}$. Without loss of generality, we use $g_K(x)$ as the basis for computing the ratios. Our projection method attempts to find a plane such that $\log \frac{g_k(x)}{g_K(x)}$, $k = 1, ..., K-1$ can be well approximated if only the projection of data into the plane is specified. By preserving the discriminant functions, the posterior probabilities of clusters will remain accurate.

Let the data set be $\{x_1, x_2, ..., x_n\}$, $x_i \in \mathcal{R}^d$. Denote a particular dimension of the data set by $x_{\cdot l} = (x_{1,l}, x_{2,l}, ..., x_{n,l})^t$, $l = 1, ..., d$. For each k, $k = 1, ..., K-1$, the pairs $(x_i, \log \frac{g_k(x_i)}{g_K(x_i)})$, $i = 1, ..., n$, are computed. Let $y_{i,k} = \log \frac{g_i(x_i)}{g_K(x_i)}$. Linear regression is performed based on the pairs $(x_i, y_{i,k})$, $i = 1, ..., n$, to acquire a linear approximation for each discriminant function. Let $\beta_{k,0}, \beta_{k,1}, \beta_{k,2}, ..., \beta_{k,d}$ be the regression coefficients for the kth discriminant function. Denote $\beta_k = (\beta_{k,1}, \beta_{k,2}, ..., \beta_{k,d})^t$ and the fitted values for $\log \frac{g_k(x_i)}{g_K(x_i)}$ by $\hat{y}_{i,k} = \beta_{k,0} + \beta_k^t x_i$. Also denote $\tilde{y}_{i,k} = \beta_k^t x_i = \hat{y}_{i,k} - \beta_{k,0}$. For mathematical tractability, we convert the approximation of the discriminant functions to the approximation of the linearly regressed values $(\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,K-1})$, $i = 1, ..., n$, which is equivalent to approximate $(\tilde{y}_{i,1}, \tilde{y}_{i,2}, ..., \tilde{y}_{i,K-1})$, since the two only differ by a constant. To precisely specify $(\tilde{y}_{i,1}, \tilde{y}_{i,2}, ..., \tilde{y}_{i,K-1})$, we need the projection of x_i onto the $K-1$ directions, $\beta_1, \beta_2, ..., \beta_{K-1}$. If we are restricted to showing the data in a plane and $K-1 > 2$, further projection of $(\tilde{y}_{i,1}, \tilde{y}_{i,2}, ..., \tilde{y}_{i,K-1})$ is needed. At this stage, we employ PCA on the vectors $(\tilde{y}_{i,1}, \tilde{y}_{i,2}, ..., \tilde{y}_{i,K-1})$ (referred to as the discriminant vectors), $i = 1, ..., n$, to yield a two-dimensional projection. Suppose the two principal component directions for the discriminant vectors are $\gamma_j = (\gamma_{j,1}, ..., \gamma_{j,K-1})^t$, $j = 1, 2$. The two principal components v_j, $j = 1, 2$, are

$$\begin{pmatrix} v_{1,j} \\ v_{2,j} \\ \vdots \\ v_{n,j} \end{pmatrix} = \gamma_{j,1} \begin{pmatrix} \tilde{y}_{1,1} \\ \tilde{y}_{2,1} \\ \vdots \\ \tilde{y}_{n,1} \end{pmatrix} + ... + \gamma_{j,K-1} \begin{pmatrix} \tilde{y}_{1,K-1} \\ \tilde{y}_{2,K-1} \\ \vdots \\ \tilde{y}_{n,K-1} \end{pmatrix} = \sum_{l=1}^{d} [\sum_{k=1}^{K-1} \gamma_{j,k} \beta_{k,l}] x_{\cdot,l}.$$

To summarize, the two projection directions for x_i are

$$\tau_j = (\sum_{k=1}^{K-1} \gamma_{j,k} \beta_{k,1}, \sum_{k=1}^{K-1} \gamma_{j,k} \beta_{k,2}, ..., \sum_{k=1}^{K-1} \gamma_{j,k} \beta_{k,d})^t, j = 1, 2. \tag{3.2}$$

The two projection directions in (3.2) are not guaranteed to be orthogonal, but it is easy to find two orthonormal directions spanning the same plane.

3.1.6 Experiments on “Vee” Multimodal Imagery

We implement the nonparametric modal clustering method on a data set we refer to as the “Vee data,” which consists of one optical image of size 103 × 103 pixels and one lidar image of size 100 × 100 pixels covering a concrete “V” located on the south side of Kirtland Air Force Base in Albuquerque, New Mexico. One leg of the “V” consists of an elevated concrete strip extending to the right from the center of a concrete circle, and the other leg is a slightly elevated gravel path extending in a southwest direction from the lower left-hand curve of the concrete circle. Below in Figure 3.1 are the original optical and lidar images.

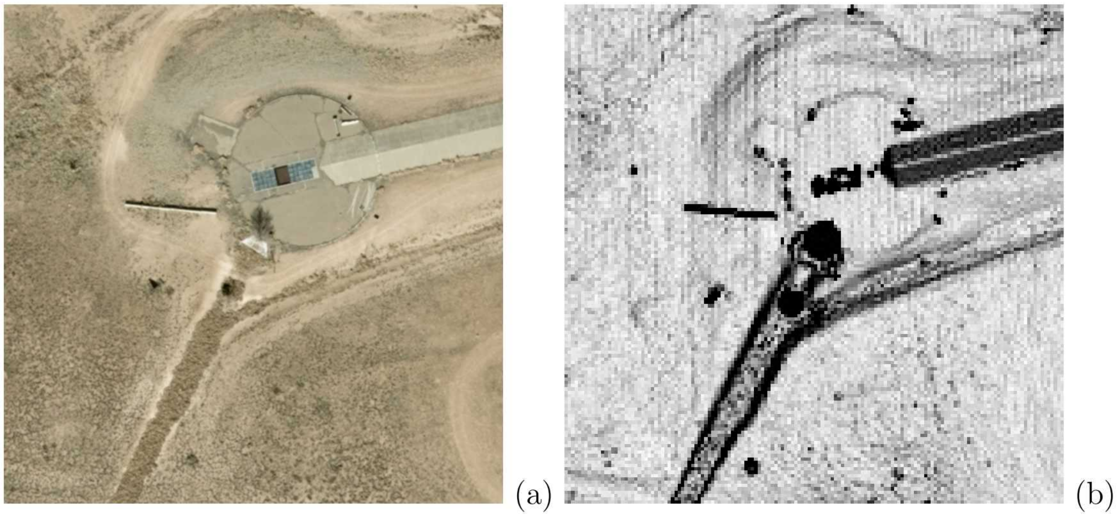


Figure 3.1: Original optical (a) and lidar (b) images.

In order to visualize the performance of the clustering model, we will need to reduce the size of the data to a dimension of two. However, we also want to plot the input data itself so the clusters can be interpreted. Since lidar data measures of the heights of the objects in a particular pixel, no changes to the data are necessary. However, for the optical image, which contains the colors of the objects in a pixel and contain the red-blue-green (RGB) values of each pixel, we decide to convert the optical image to a gray-scale image, so that each pixel’s data value is a scalar. Alternatively, we could have applied PCA to the optical image and obtain the first two principal component values to reduce the size of each data point from three dimensions to two. However, principal components are often not interpretable.

Below in Figure 3.2 are the cluster assignments and the contour plots for both the optical and lidar images. For the optical image, the nonparametric modal clustering methods finds the optimal number of clusters to be 33, and thus, they are all very small in size and do not really have any semantic meanings. The cluster assignments don’t correspond well at all with the contents of the image. We cannot make out any significant features, such as the “V” or the concrete circle. Two reasons why this may be the case are the color differences between the objects in the image are big enough for the HMAC algorithm to distinctly cluster those objects, and the principal components computed do not correspond well to the objects in the image. For the lidar image, the optimal number of clusters is only seven, and the most prominent clusters are at the bottom center of the image (by far) and at the top center of the image, to a much lesser extent. While the plot of the data seems to capture the "V" a bit better than with the plot of the optical data, and one of the clusters may be able to capture one leg of the "V", the HMAC algorithm doesn't capture the "V" that is the center of the scene very well. This is likely due to the fact that the "V" is only slightly elevated from the ground, and the height difference between the "V" and the ground may not have been big enough for the HMAC to distinguish the "V".

We note that the cluster assignments are based on the hard assignments made by the HMAC algorithm for each pixel. An alternative method is assigning clusters by selecting the maximum estimated density value over all of the possible clusters, $\tilde{g}_k(x)$. However, we find that the cluster assignments chosen from the maximum estimated value of $\tilde{g}_k(x)$ do not always agree with the cluster assignments made by the HMAC algorithm. This indicates that the density estimation method is not an accurate and robust method for estimating the cluster probabilities for each pixel. One possible explanation is that the density estimates, $\tilde{g}_k(x)$, are computed outside of the HMAC algorithm, but this does not necessarily describe the cluster probability, which is a parameter estimated as part of the clustering algorithm seen in methods such as the Gaussian mixture model and the nonparametric mixture model (Benaglia et al., 2009a). Because the clustering probability results are not trustworthiness, we do not perform an uncertainty analysis.

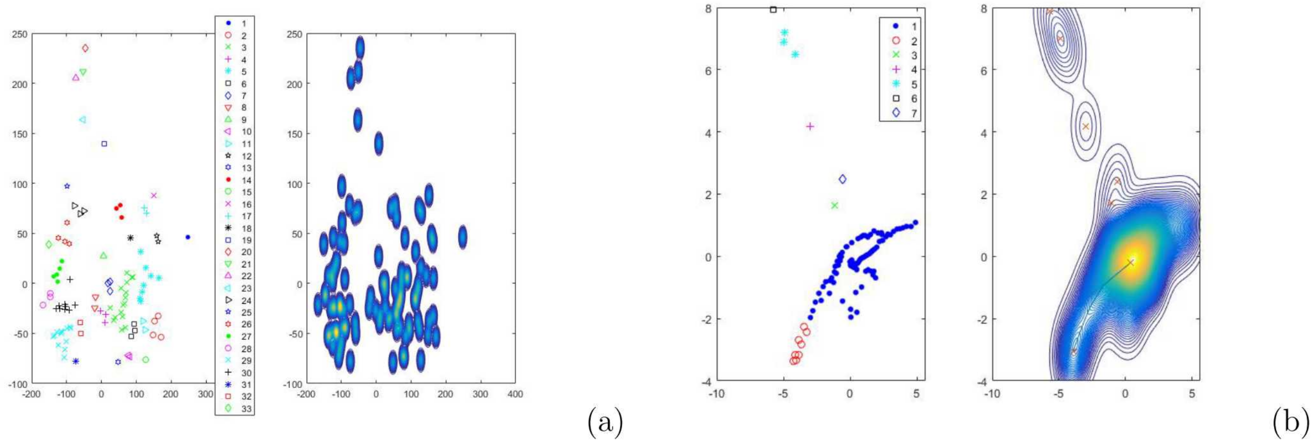


Figure 3.2: Cluster assignment and contour plots for the nonparametric modal clustering method applied to a 103 × 103 optical image converted to gray scale (a) and to a 100 × 100 lidar image (b). For the optical image, the image is segmented into 33 clusters. For the lidar image, the image is segmented into seven clusters.

3.1.7 Discussion and Future Work

One of the biggest drawbacks of the nonparametric modal clustering method is that the posterior clustering probabilities, which estimate the probabilities of each data point belonging to a specific cluster after the data point has been observed, is estimated outside of the clustering algorithm. This is unlike the EM algorithm for fitting Gaussian mixture models, when the posterior clustering probabilities are estimated as the number of clusters and the cluster assignments for each data point are determined. Therefore, modifying the method to estimate the posterior clustering probabilities is a big area of future work that would greatly aid in computing the uncertainty of this model's performance in clustering image data.

3.2 Nonparametric Mixture Model

Benaglia et al. (2009a) propose an EM-like algorithm for fitting a non-parametric mixture model to multivariate random vector data. Suppose the vectors $\mathbf{X}_1, ..., \mathbf{X}_n$ are a simple random sample from a finite mixture of $m > 1$ arbitrary distributions. The density of each $\mathbf{X}_i$ may be written

$$g_\varphi(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{x}_i), \tag{3.3}$$

where $\mathbf{x}_i \in \mathbb{R}^r$, $\varphi^t = (\lambda^t, \phi^t) = (\lambda_1, ..., \lambda_m, \phi_1, ..., \phi_m)$ denotes the parameter, and the λ_j are positive and sum to unity. We assume that the ϕ_j are drawn from some family $\mathcal{F}$ of multivariate density functions (say, absolutely continuous with respect to Lebesgue measure).

A common restriction placed on $\mathcal{F}$ is that each joint density $\phi_j(\dot{})$ is equal to the product of its marginal densities. In other words, the coordinates of the $\mathbf{X}_i$ vectors are independent, conditional on the subpopulation or component (ϕ_1 through ϕ_m) from which $\mathbf{X}_i$ is drawn. Therefore, model (3.3) becomes

$$g_\varphi(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}), \tag{3.4}$$

where the function $f(\cdot)$, with or without subscripts, will always denote a univariate density function. Another special case of the model is in which the density $f_{jk}(\cdot)$ does not depend on k, that is, in which the $\mathbf{X}_i$ are not only conditionally independent but identically distributed as well:

$$g_\varphi(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_j(x_{ik}), \tag{3.5}$$

where we assume that $f_{j1}(\cdot) = \ldots = f_{jr}(\cdot)$ for all j.

To encompass both the special case (3.5) and the more general case (3.4) simultaneously, we allow that the coordinates of $\mathbf{X}_i$ are conditionally independent and that there exist *blocks* of coordinates that are also indetically distributed. These blocks may be all of size one so that case (3.4) is still covered, or there may exist only a single block of size r, which is base (3.5). If we let b_k denote the block to which the $k$the coordinate belongs, where $1 \leq b_k \leq B$

and B is the total number of such blocks, then equation (3.4) is replaced by

$$g_{\varphi}(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}). \tag{3.6}$$

The nonparametric EM algorithm begins with given initial values $\phi^0 = (\lambda^0, \mathbf{f}^0)$. Then for $t = 1, 2, ...,$ we follow these three steps:

1. **E-step:** Calculate the "posterior" probabilities (conditional on the data and ϕ^t) of component inclusion,

$$p_{ij}^t = P_{\phi^t}(Z_{ij} = 1|\mathbf{x}_i) \tag{3.7}$$

$$= \frac{\lambda_j^t \prod_{k=1}^r f_{jb_k}^t(x_{ik})}{\sum_{j'=1}^m \lambda_{j'}^t \prod_{k=1}^r f_{j'b_k}^t(x_{ik})} \tag{3.8}$$

for all $i = 1, ..., n$ and $j = 1, ..., m$.

2. **M-step:** Set

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^t \tag{3.9}$$

for $j = 1, ..., m$.

3. **Nonparametric density estimation step:** For any real u, define for each component $j \in \{1, ..., m\}$ and each block $l \in \{1, ..., B\}$

$$f_{jl}^{t+1}(u) = \frac{\frac{1}{h}\sum_{k=1}^r \sum_{i=1}^n p_{ij}^t I\{b_k = l\} K(\frac{u - x_{ik}}{h})}{\sum_{k=1}^r \sum_{i=1}^n p_{ij}^t I\{b_k = l\}} \tag{3.10}$$

$$= \frac{1}{nhC_l\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} p_{ij}^t I\{b_k = l\} K(\frac{u - x_{ik}}{h}), \tag{3.11}$$

where $K(\cdot)$ is a kernel density function, h is a bandwidth chosen by the user, and

$$C_l = \sum_{k=1}^{r} I\{b_k = l\}$$

is the number of coordinates in the $l$the block. Note that in the case in which $b_k = k$ for all k, equation (3.10) becomes

$$f_{jk}^{t+1}(u) = \frac{1}{nh\lambda_j^{t+1}} \sum_{i=1}^{n} p_{ij}^t K(\frac{u - x_{ik}}{h}). \tag{3.12}$$

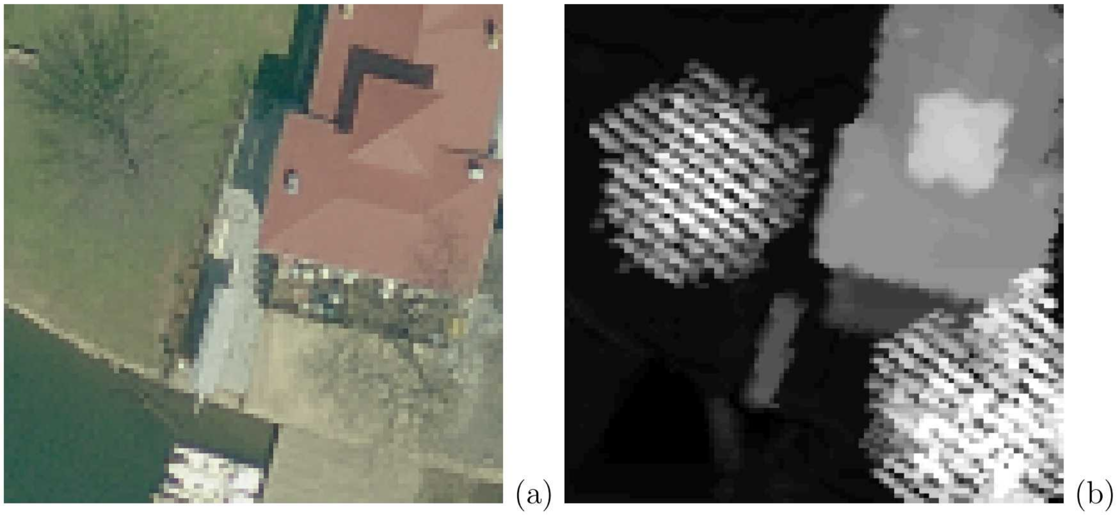


Figure 3.3: Original Philadelphia optical (a) and lidar (b) images.

3.2.1 Experiments with Philadelphia Multimodal Imagery

We fit the nonparametric mixture model to the Philadelphia optical and lidar images, which we reproduce in Figure 3.3 for convenience.

The model fit is done using the R package mixtools (Benaglia et al., 2009b), which is associated with Benaglia et al. (2009a). The bandwidth is chosen using the method of Benaglia et al. (2011). The implementation is very slow, and we are only able to run 10 iterations though the EM-like algorithm. Below in Figure 3.4 are the category confusion plots showing the entropy only (top row) and the entropy overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery. In the top plots of columns (a) and (b) that contain the black-and-white plots of the entropy values for the cluster probabilities at each pixel, where whiter colors indicate entropy values closer to one and thus, higher uncertainty, we see that there are many more whiter colors when we only consider the optical image (column (a)) than when we also incorporate the lidar image (column (b)). In the bottom plots of columns (a) and (b) that contain the cluster assignments overlaid on to the entropy values, we can see more distinct clusterings of objects, such as the shadows cast by the building, when we incorporate the information from the lidar image.

Below in Figure 3.5 are the category uncertainty plots showing the standard deviation of the posterior distribution alone (top row) and overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery. In the top plots of columns (a) and (b) that contain the black-and-white plots of the standard deviation values for the cluster probabilities of the assigned cluster at each pixel, where whiter colors indicate higher standard deviation values, and thus, higher uncertainty, we see that there are many more whiter colors when we also incorporate the lidar image (column (b)) as opposed to only considering the optical image (column (a)). This is especially true for the building, which is very distinct and has very lower uncertainty in the optical image. However, when we look at the bottom plots of columns (a) and (b) that contain the cluster assignments overlaid on

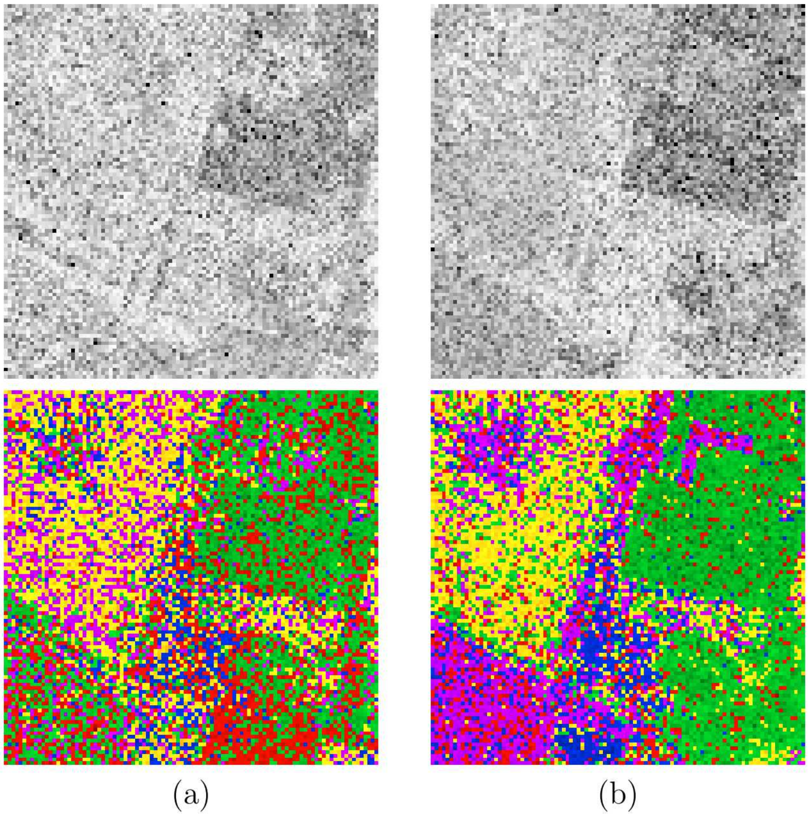


Figure 3.4: Category confusion plots showing the entropy only (top row) and the entropy overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

to the standard deviation values, we can see more distinct clusterings of objects, such as the shadows cast by the building, when we incorporate the information from the lidar image, even though there seems to be more variance in the estimated clustering probabilities.

Below in Figure 3.6 are the violin plots for a roof pixel showing the change in posteriors before (a) and after (b) incorporating lidar data into the analysis. For this particular pixel, we see that after incorporating lidar data, we now have five classes instead of only four for when we only consider the optical data. For the optical clustering probabilities, class two and class three are fairly close in variability (as indicated by the length of the bar), while classes one and four have much lower variability. For the combined data, we see much higher variabilities for classes one, two, and four, while there is much lower variabilities for classes three and five. This indicates that for this particular pixel, incorporating the lidar data increases the level of clustering uncertainty, as there is more variability in the cluster probabilities.

Finally, below in Figure 3.7 are the probability maps showing green category pixel probabilities based on optical data (a), combined optical and lidar (b), and the difference between the two (c). Panel (d) shows the K-L divergence between the category posteriors associated with panels (a) and (b). From panel (c), we see that the largest differences between the pixel

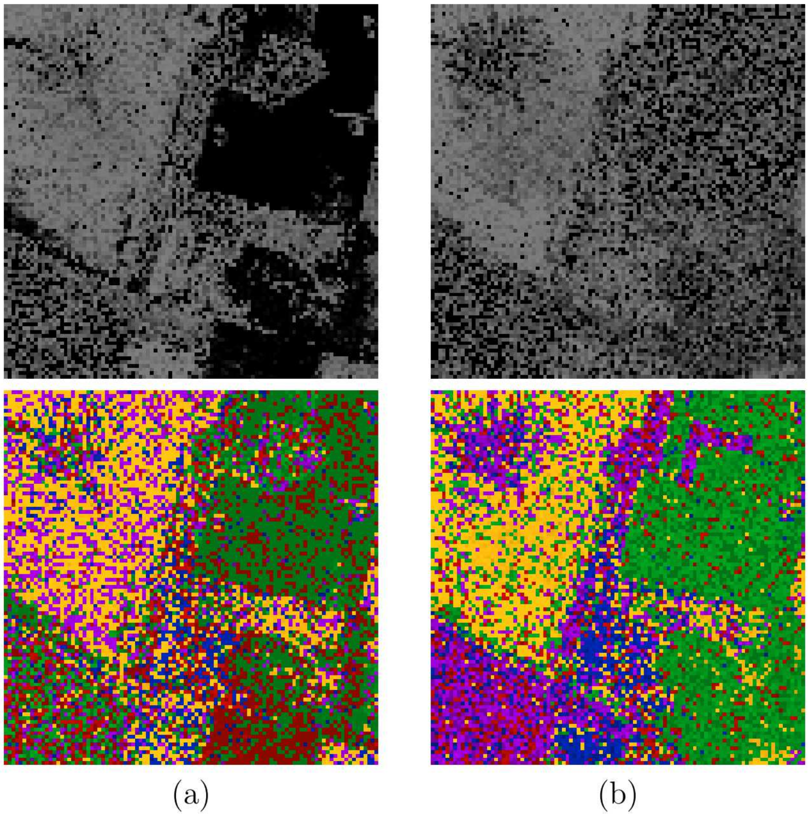


Figure 3.5: Category uncertainty plots showing the standard deviation of the posterior distribution alone (top row) and overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

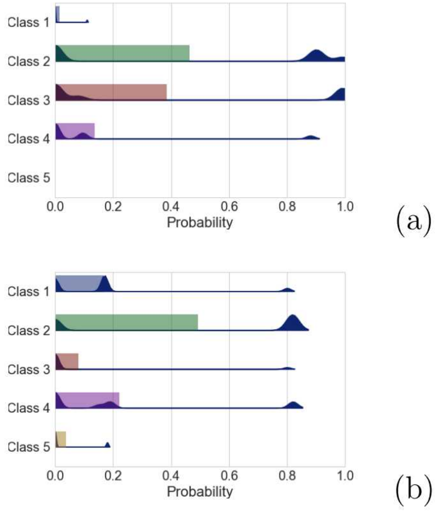


Figure 3.6: Violin plots for a roof pixel showing the change in posteriors before (a) and after (b) incorporating lidar data into the analysis.

probabilities between the optical data (panel (a)) and combined data (panel (b)) is in the roof region of the building. In the K-L divergence plot in panel (d), we see that the biggest differences between the probability distribution of the cluster probabilities lies in the roof region of the building.

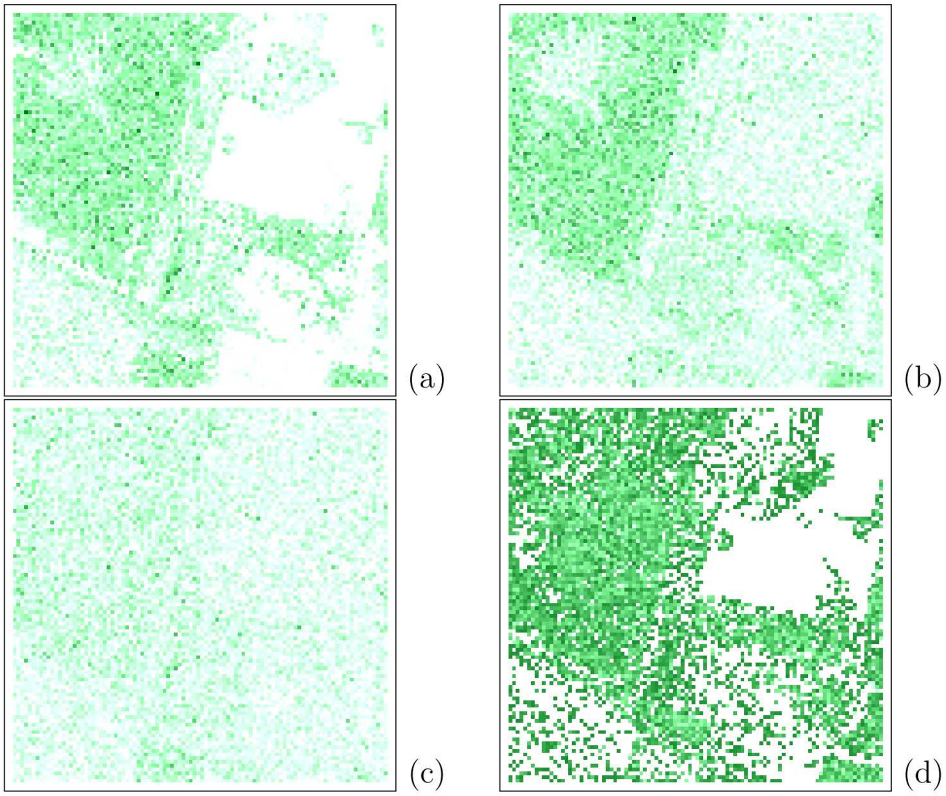


Figure 3.7: Probability maps showing green category pixel probabilities based on optical data (a), combined optical and lidar (b), and the difference between the two (c). Panel (d) shows the K-L divergence between the category posteriors associated with panels (a) and (b).

Through using measures such as Shannon's entropy and standard deviation to represent the clustering uncertainty at the pixels of an image, we can assess the differences in cluster probabilities and the level of variance and uncertainty in the cluster probabilities at each pixel. This gives insight into the value of each image into the usefulness of our analysis towards decision-making. In the Philadelphia imagery example, we see that after incorporating the lidar image along with the optical image, we get more distinct cluster assignments and there are certain classes with much higher probabilities compared to other classes. However, there can be more variance in the cluster probabilities for the assignment cluster, which means greater uncertainty in the estimates of the cluster probabilities. Overall, in this example, lidar gives more information for distingishing the important features in this scene in Philadelphia.

3.2.2 Discussion and Future Work

We now visually compare the results of the NMM to the fit of the GMM (see Chapters 4 and 5 of Stracuzzi et al. (2017a) for an in-depth analysis of these results) by using the same visualization we use for the NMM results. For the GMM, the entropy plots are in Figure 3.8, the standard deviation plots are in Figure 3.9, the violin plots for the same roof pixel are in Figure 3.10, and the probability maps and KL-divergence plot are in Figure 3.11.

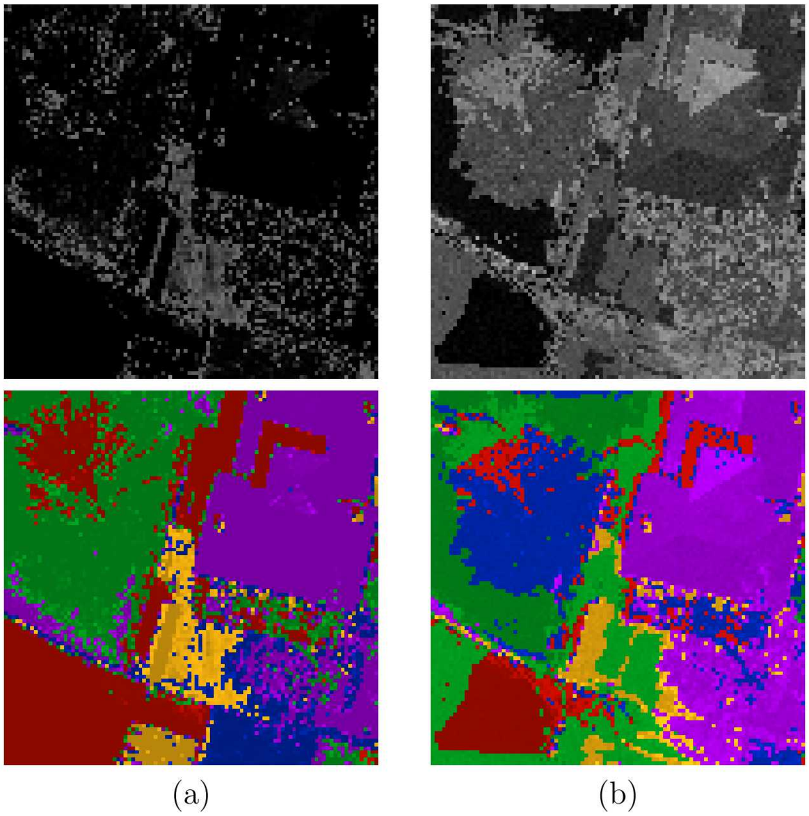


Figure 3.8: GMM category confusion plots showing the entropy only (top row) and the entropy overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

In the NMM, the clusterings are much messier and it is much more difficult to identify potential semantic meanings for the clusters. There is also higher uncertainty when applying the NMM as opposed to the GMM. This is evident when comparing the entropy, standard deviation, probability maps, and K-L Divergence plots. One possible reason that we speculate for this is the NMM allows the data to dictate the distributions of the clusters, and therefore, there is much more potential variability and error in the model fit. However, this needs to be investigated further. In addition, the results presented here are obtained after running one iteration through the EM algorithm and with 10 bootstrap samples. This is largely due to the computational cost and the very slow speed that the algorithm runs at. In addition to seeing how we can implement the NMM much faster, we also need to investigate how does uncertainty change with different numbers of iterations through the EM algorithm

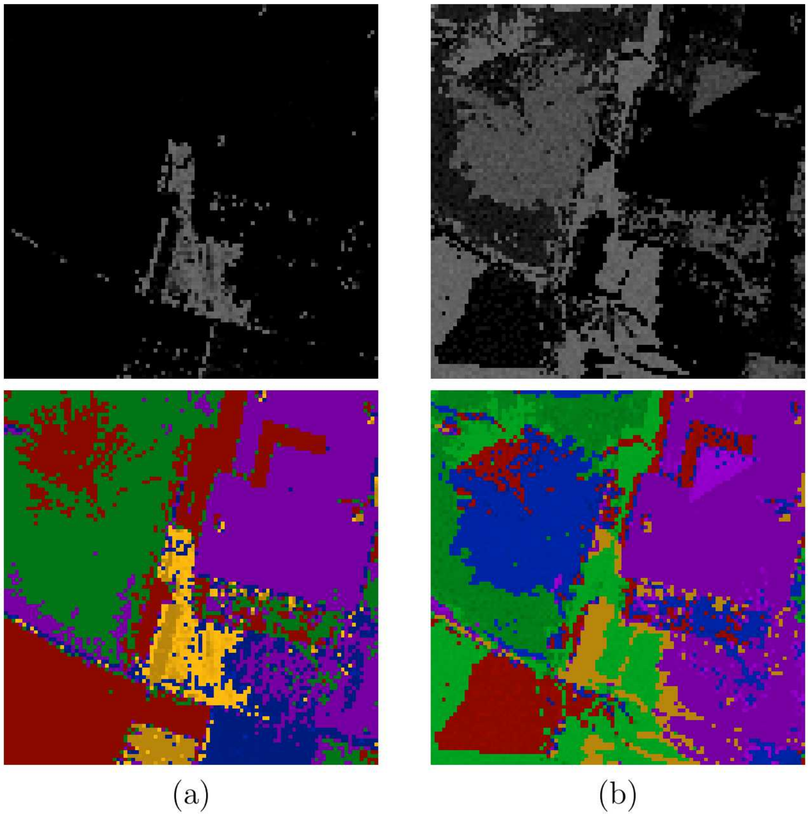
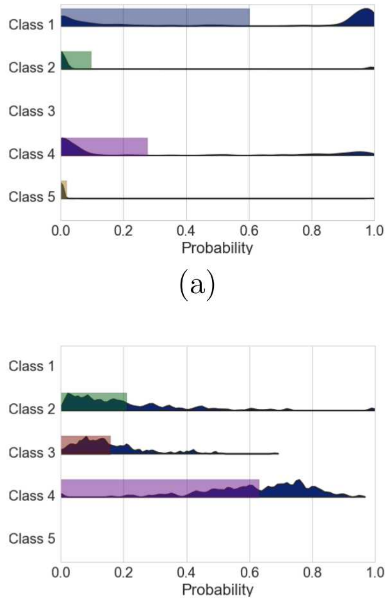


Figure 3.9: GMM category uncertainty plots showing the standard deviation of the posterior distribution alone (top row) and overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

and the number of bootstrap samples, and the reasons for these changes. Additional areas of future work include improved parameter tuning for the NMM, integration of supervised labels (like with Experiments 3 and 4 described in section 4 of Stracuzzi et al. (2018)), and a mathematical framework for multimodal uncertainty analysis with the NMM.



(a)

(b)

Figure 3.10: GMM violin plots for a roof pixel showing the change in posteriors before (a) and after (b) incorporating lidar data into the analysis.

3.3 Bayesian Consensus Clustering

In multimodal data analysis, *consensus clustering* (also called *ensemble clustering*) determines an overall partition of the objects that agree with the most source-specific clusterings. In other words, after clustering each data source, what is the overall clustering that factors in all data sources?

There have been exploratory methods that simultaneously model shared features and features that are specific to each data source have been developed as flexible alternatives to separate analyses of each data source and methods that perform joint analysis while ignoring the heterogeneity of the data (Lock et al., 2013; Löfstedt and Trygg, 2011; Ray et al., 2014; Zhou et al., 2012). Most application of clustering multisource data follow one of two general approaches:

1. Clustering of each data source separately, potentially followed by a post hoc integration of these separate clusterings.

2. Combining all data sources to determine a single 'joint' clustering.

Under approach (1), several functions and algorithms to perform consensus clustering have been proposed [see Nguyen and Caruana (2007) for a survey]. Most of these methods do not inherently model uncertainty, and statistical models assume that the separate clusterings are known in advance (Wang et al., 2010, 2011). Consensus clustering is most commonly used to combine multiple clustering algorithms, or multiple realizations of the

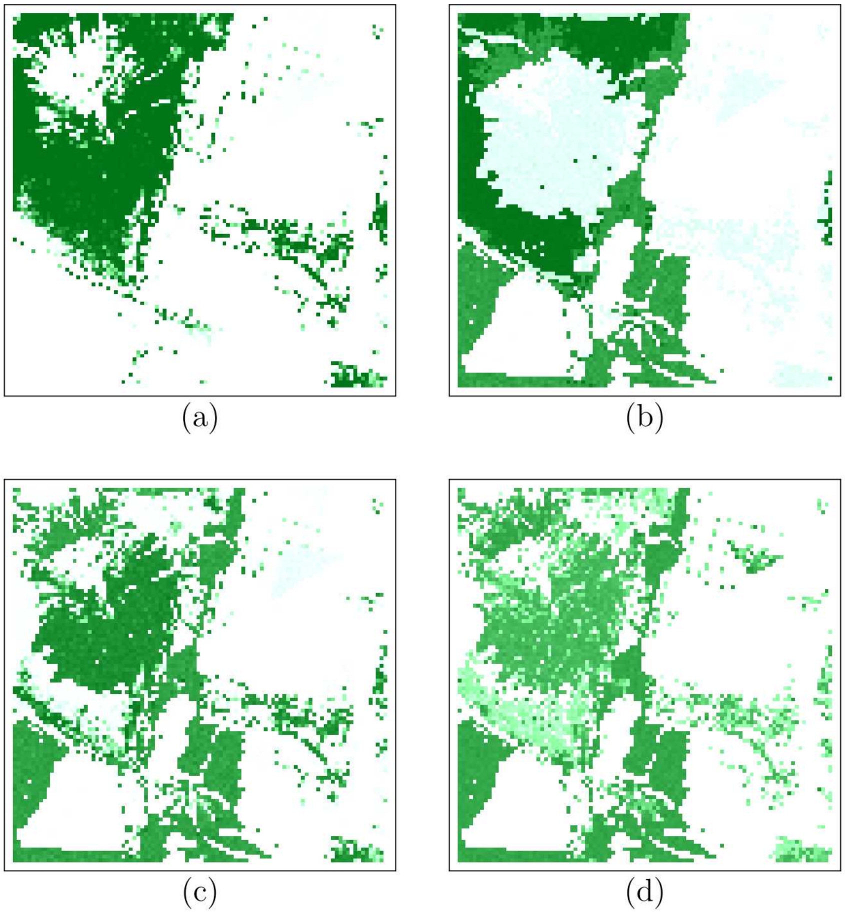


Figure 3.11: GMM probability maps showing green category pixel probabilities based on optical data (a), combined optical and lidar (b), and the difference between the two (c). Panel (d) shows the K-L divergence between the category posteriors associated with panels (a) and (b).

same clustering algorithm, on a single dataset. These approaches models source-specific features and determines an overall clustering. However, the two-stage process of performing entirely separate clusterings followed by post hoc integration limits the power to identify and exploit shared structure.

Approach (2) effectively exploits shared structure, at the expense of failing to recognize features that are specific to each data source. Within a model-based statistical framework, one can find the clustering that maximizes a joint likelihood. Assuming that each source is conditionally independent given the clustering, the joint likelihood is the produce of the likelihood functions for each data source. This approach has been used in the context of integrating gene expression and DNA methylation data (Kormaksson et al., 2012). The *iCluster* method (Mo et al., 2013; Shen et al., 2009) performs clustering by first fitting a Gaussian latent factor model to the joint likelihood, and then the clusters are determined by K-means clustering of the factor scores. Rey and Roth (2012) propose a dependency-seeking model in which the goal is to find a clustering that accounts for associations across the data sources.

More flexible methods allow for separate but dependent source clusterings. Dependent models have been used to simultaneously cluster gene expression and proteomic data (Rogers et al., 2008), gene expression and transcription factor binding data (Savage et al., 2010), and gene expression and copy number data (Yuan et al., 2011). Kirk et al. (2012) describe a more general dependence model for two or more data sources. Their approach, called *Multiple Dataset Integration* (MDI), uses a statistical framework to cluster each data source while simultaneously modeling the pairwise dependence between clusters. However the pairwise dependence model does not explicitly model adherence to an overall clustering.

We use the *Bayesian consensus clustering* (BCC) approach (Lock and Dunson, 2013a), which we summarize below. BCC differs from traditional consensus clustering in three key aspects.

1. Both the source-specific clusterings and the consensus clustering are modeled in a statistical way that allows for uncertainty in all parameters.
2. The source-specific clusterings and the consensus clustering are estimated simultaneously, rather in two stages. This permits borrowing of information across sources for more accurate cluster assingments.
3. The strength of association to the consensus clustering for each data source is learned form the data and accounted for in the model.

3.3.1 Finite Dirichlet mixture models

The Bayesian consensus clustering (BCC) method is an extension of the Dirichlet mixture model to accommodate data from multiple data sources. We begin with a description of the finiste Dirichlet mixture model for clustering a single dataset. Given data X_n for N objects $(n = 1, ..., N)$, the goal is to partition these objects into at most K clusters. Typically, X_n is a multidimensional vector, but it can assume more complex data strictures. Let $f(X_n|\theta)$ define a probability model for X_n given parameter(s) θ. For example, f may be a Gaussian density defined by the mean and variance $\theta = (\mu, \sigma^2)$. Each X_n is drawn independently from a mixture distribution with K components, specified by the parameters $\theta_1, ..., \theta_K$. Let $C_n \in \{1, ..., K\}$ represent the component corresponding to X_n, and π_k be the probability that an arbitrary object belongs to cluster k:

$$\pi_k = P(C_n = k).$$

Then, the generative model is

$$X_n \sim f(\cdot|\theta_k) \text{ with probability } \pi_k.$$

Under a Bayesian framework, one can put a prior distribution on $\Pi = (\pi_1, ..., \pi_K)$ and the parameter set $\Theta = (\theta_1, ..., \theta_K)$. It is natural to use a Dirichlet prior distribution for Π.

Standard computational methods can then be used to approximate the prior distribution for Π, Θ, and $\mathbb{C} = (C_1, ..., C_N)$. The Dirichlet prior is characterized by a K-dimensional concentration parameter β of positive reals.

3.3.2 Integrative model

The integrative model is the extension of the Dirichlet mixture model to accommodate data from M sources $\mathbb{X}_1, ..., \mathbb{X}_M$. Each data source is available for a common set of N objects, where X_{mn} represents data m for object n. Each data source requires a probability model $f_m(X_n|\theta_m)$ parameterized by θ_m. Under the general framework presented here, $\mathbb{X}_m$ may have disparate structure. For example, X_{1n} may give an image where f_1 defines the spectral density for a Gaussian random field, while X_{2n} may give a categorical vector where f_2 defines a multivariate probability mass function.

We assume there is a separate cluster of the objects for each data source, but that there adhere loosely to an overall clustering. Formally, each X_{mn}, $n = 1, ..., N$, is drawn independently from a K-component mixture distribution specified by the parameters $\theta_{m1}, ..., \theta_{mK}$. Let $L_{mn} \in \{1, ..., K\}$ represent the component corresponding to X_{mn}. Furthermore, let $X_n \in \{1, ..., K\}$ represent the overall mixture component for object n. The source-specific clusterings $\mathbb{L}_m = (L_{m1}, ..., L_{mN})$ are dependent on the overall clustering $\mathbb{C} = (C_1, ..., C_N)$:

$$P(L_{mn} = k|C_n) = \nu(k, C_n, \alpha_m),$$

where α_m adjusts the dependence function ν. The data $\mathbb{X}_m$ are independent of $\mathbb{C}$ conditional on the source-specific clustering $\mathbb{L}_m$. Hence, $\mathbb{C}$ serves only to unify $\mathbb{L}_1, ..., \mathbb{L}_M$. The conditional model is

$$P(L_{mn} = k|X_{mn}, C_n, \theta_{mk}) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn}|\theta_{mk}).$$

Throughout this article, we assume ν has the simple form

$$\nu(k, C_n, \alpha_m) = \begin{cases} \alpha_m, & \text{if } C_n = L_{mn} \\ \frac{1-\alpha_m}{K-1}, & \text{otherwise} \end{cases} \tag{3.13}$$

where $\alpha_m \in [\frac{1}{K}, 1]$ controls the adherence of data source m to the overall clustering. More simply, α_m is the probability that $L_{mn} = C_n$, so that if $\alpha_m = 1$, then $\mathbb{L}_m = \mathbb{C}$. The α_m are estimated from the data together with $\mathbb{C}$ and $\mathbb{L}_1, ..., \mathbb{L}_m$.

Let π_k be the probability that an object belongs to the overall cluster k:

$$\pi_k = P(C_n = k).$$

We assume a Dirichlet(β) prior distribution for $\Pi = (\pi_1, ..., \pi_K)$. The probability that an object belongs to a given source-specific cluster is

$$P(L_{mn} = k|\Pi) = \pi_k \alpha_m + (1 - \pi_k)\frac{1 - \alpha_m}{K - 1}. \tag{3.14}$$

A simple application of Bayes rule gives the conditional distribution of $\mathbb{C}$:

$$P(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m),$$

where ν is defined as in (3.13). Integrating over the overall clustering C gives the joint marginal distribution of $\mathbb{L}_1, ..., \mathbb{L}_M$:

$$P(\{L_{mn} = k_m\}_{m=1}^{M}|\Pi, \alpha) \propto \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} \nu(k_m, k, \alpha_m).$$

3.3.3 Estimation

We implement a general Bayesian framework for estimation of the integrative clustering model. A Gibbs sampling procedure is used to estimate the posterior distribution for the parameters. No specific form for the f_m and the parameters θ_{mk} is assumed.

We assume that $\mathbb{X}_i$ has a normal-gamma mixture distribution with cluster-specific mean and variance. Mathematically,

$$X_{mn}|L_{mn} = k \sim N(\mu_{nm}, \Sigma_{mk}),$$

where

- μ_{mk} is a D_m dimensional mean vector, where D_m is the dimension of data source m.
- Σ_{mk} is a $D_m \times D_m$ diagonal covariance matrix, $\Sigma_{mk} = \text{Diag}(\sigma_{mk1}, ..., \sigma_{mkD_m})$.

We use a D_m dimensional normal-inverse-gamma prior distribution for $\theta_{mk} = (\mu_{mk}, \Sigma_{mk})$. That is,

$$\theta_{mk} \sim N\Gamma^{-1}(\eta_{m0}, \lambda_0, A_{m0}, B_{m0}),$$

where η_{m0}, λ_0, A_{m0}, and B_{m0} are hyperparameters. It follows that μ_{mk} and Σ_{mk} are given by

- $\frac{1}{\sigma^2_{mkd}} \sim \text{Gamma}(A_{m0d}, B_{m0d})$, and
- $\mu_{mkd} \sim N(\eta_{m0}, \frac{\sigma^2_{mkd}}{\lambda_0})$ for $d = 1, ..., D_m$.

We set $\lambda_0 = 1$ and estimate μ_{m0}, A_{m0}, and B_{m0} from the mean and variance of each variable in $\mathbb{X}_m$.

We also use conjugate prior distributions for α_m and Π.

- $\alpha_m \sim \text{TBeta}(a_m = 1, b_m = 1, \frac{1}{K})$, the $\text{Beta}(a_m, b_m)$ distribution truncated below by $\frac{1}{K}$. By default, we choose $a_m = b_m = 1$, so that the prior for α_m is uniformly distributed between $\frac{1}{K}$ and 1.

- $\Pi \sim \text{Dirichlet}(\beta_0 = (1, 1, ..., 1))$, so that the prior for Π is uniformly distributed on the standard $(M-1)$-simplex.

Markov Chain Monte Carlo (MCMC) iteratively samples from the following conditional distributions:

- $\Theta_m | \mathbb{X}_m, \mathbb{L}_m \sim p_m(\theta_{mk} | \mathbb{X}_m, \mathbb{L}_m)$ for $k = 1, ..., K$

$$\theta_{mk} \sim N\Gamma^{-1}(\eta_{mk}, \lambda_k, A_{m0}, B_{m0})$$

- $\mathbb{L}_m | \mathbb{X}_m, \Theta_m, \alpha_m, \mathbb{C} \sim P(k | X_{mn}, C_n, \theta_{mk}, \alpha_m)$ for $n = 1, ..., N$, where

$$P(k | X_{mn}, C_n, \theta_{mk}, \alpha_m) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn} | \theta_{mk}).$$

- $\alpha_m | \mathbb{C}, \mathbb{L}_m \sim \text{TBeta}(a_m + \tau_m, b_m + N - \tau_m, \frac{1}{K})$, where τ_m is the number of samples n satisfying $L_{mn} = C_n$.

- $\mathbb{C} | \mathbb{L}_m, \Pi, \alpha \sim P(k | \Pi, \{L_{mn}, \alpha_m\}_{m=1}^M)$ for $n = 1, ..., N$, where

$$P(k | \Pi, \{L_{mn}, \alpha_m\}_{m=1}^M) \propto \pi_k \prod_{m=1}^{M} \nu(k, L_{mn}, \alpha_m)$$

- $\Pi | \mathbb{C} \sim \text{Dirichlet}(\beta_0 + \rho)$, where ρ_k is the number of samples allocated to cluster k in $\mathbb{C}$ Lock and Dunson (2013a,b)

3.3.4 Multimodal Uncertainty Quantification

We extend the work of Lock and Dunson (2013a) by quantifying the uncertainty of the parameter estimates and mathematically relating the uncertainties between the source-specific clustering results to the uncertainty in the overall clustering results. We use the *variance* as our measure of uncertainty in the results obtained after implementing the BCC method. Thus, the uncertainty of the overall clustering is the variance of the distribution of the cluster probabilities. This can be written as $\text{Var}[P(C_n = k)]$ for the overall consensus clustering. The uncertainty of the source-specific clusterings are $\text{Var}[P(L_{mn} = k_m)]$ for data source m. We wish to relate the uncertainty of the overall consensus clustering to the uncertainty of the source-specific clusterings.

The concrete example we use is the analysis of the optical and lidar images of Philadelphia's Schuylkill River. Each image represents a data source, so we have $M = 2$. We will derive our mathematical expressions for this specific case. The derivations for the general case of any number of M data sources is an area of future work.

Recall that in BCC, the conditional distribution of the overall consensus clustering $\mathbb{C}$ is

$$P(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m).$$

Therefore, the uncertainty of the overall clustering is

$$\mathrm{Var}[P(C_n = k|\mathbb{L}, \Pi, \alpha)] \propto \mathrm{Var}[\pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m)].$$

In the case of $M = 2$, from the definition of the dependence function ν in (3.13), $\mathrm{Var}[P(C_n = k|\mathbb{L}, \Pi, \alpha)]$ can be written more specifically as

$$\begin{aligned}\mathrm{Var}[P(C_n = k|\mathbb{L}, \Pi, \alpha)] &\propto \mathrm{Var}\{\pi_k[\nu(L_{1n}, k, \alpha_1)][\nu(L_{2n}, k, \alpha_2)]\} \\ &\propto \begin{cases} \mathrm{Var}[\pi_k \alpha_1 \alpha_2], & \text{if } L_{1n} = k, L_{2n} = k \\ \mathrm{Var}[\pi_k \alpha_1 \frac{1-\alpha_2}{K-1}], & \text{if } L_{1n} = k, L_{2n} \neq k \\ \mathrm{Var}[\pi_k \frac{1-\alpha_1}{K-1} \alpha_2], & \text{if } L_{1n} \neq k, L_{2n} = k \\ \mathrm{Var}[\pi_k \frac{1-\alpha_1}{K-1} \frac{1-\alpha_2}{K-1}], & \text{if } L_{1n} \neq k, L_{2n} \neq k. \end{cases}\end{aligned}$$

For the sake of simplicity, we will omit the conditioned variables for each posterior, but in all future expressions, they are for the conditional distributions. Since π_k, α_1, and α_2 are all dependent,

$$\begin{aligned}\mathrm{Var}(\pi_k \alpha_1 \alpha_2) &= \mathrm{Cov}(\pi_k^2, \alpha_1^2 \alpha_2^2) + (\mathrm{Var}(\pi_k) + [E(\pi_k)]^2)(\mathrm{Var}(\alpha_1 \alpha_2) + [E(\alpha_1 \alpha_2)]^2) \\ &\quad - [\mathrm{Cov}(\pi_k, \alpha_1 \alpha_2) + E(\pi_k) E(\alpha_1 \alpha_2)]^2 \\ \mathrm{Var}(\alpha_1 \alpha_2) &= \mathrm{Cov}(\alpha_1^2, \alpha_2^2) + (\mathrm{Var}(\alpha_1) + [E(\alpha_1)]^2)(\mathrm{Var}(\alpha_2) + [E(\alpha_2)]^2) \\ &\quad - [\mathrm{Cov}(\alpha_1, \alpha_2) + E(\alpha_1) E(\alpha_2)]^2\end{aligned}$$

$$\begin{aligned}&\mathrm{Var}(\pi_k \alpha_1 \alpha_2) = \mathrm{Cov}(\pi_k^2, \alpha_1^2 \alpha_2^2) + (\mathrm{Var}(\pi_k) + [E(\pi_k)]^2)(\mathrm{Cov}(\alpha_1^2, \alpha_2^2) \\ &+ (\mathrm{Var}(\alpha_1) + [E(\alpha_1)]^2)(\mathrm{Var}(\alpha_2) + [E(\alpha_2)]^2) - \underbrace{[\mathrm{Cov}(\alpha_1, \alpha_2) + E(\alpha_1) E(\alpha_2)]^2}_{[E(\alpha_1 \alpha_2)]^2} \\ &+ [E(\alpha_1 \alpha_2)]^2) - [\mathrm{Cov}(\pi_k, \alpha_1 \alpha_2) + E(\pi_k) E(\alpha_1 \alpha_2)]^2 \\ &= \mathrm{Cov}(\pi_k^2, \alpha_1^2 \alpha_2^2) + (\mathrm{Var}(\pi_{\mathbf{k}}) + [E(\pi_k)]^2)(\mathrm{Cov}(\alpha_1^2, \alpha_2^2) \\ &+ (\mathrm{Var}(\alpha_{\mathbf{1}}) + [E(\alpha_1)]^2)(\mathrm{Var}(\alpha_{\mathbf{2}}) + [E(\alpha_2)]^2)) - [\mathrm{Cov}(\pi_k, \alpha_1 \alpha_2) + E(\pi_k) E(\alpha_1 \alpha_2)]^2\end{aligned}$$

Relating $\mathrm{Var}(\pi_k)$, $\mathrm{Var}(\alpha_1)$, and $\mathrm{Var}(\alpha_2)$, we can conclude that the **uncertainty for the overall clustering is directly proportional to the uncertainties for the adherence**

of the source-specific clusterings to the overall clustering, which affect the results of the source-specific clusterings.

Using the conjugate priors for fitting the BCC to the imagery data described in section 3.3.3, we can specify some of the expressions. Recall that the conditional distribution for Π, the vector of overall clustering probabilities, is

$$\Pi|\mathbb{C} \sim \text{Dirichlet}(\beta_0 + \rho), \beta_0 = (1, 1, ..., 1),$$

and since the marginal distribution of each element of a Dirichlet-distributed vector is the beta distribution, the conditional distribution of $\pi_k|\mathbb{C}$ is

$$\pi_k|\mathbb{C} \sim \text{Beta}(1 + \rho_k, \sum_{i=1}^{K}(1 + \rho_i) - (1 + \rho_k)) = \text{Beta}(1 + \rho_k, K - 1 + \sum_{i \neq k} \rho_i).$$

The conditional expectation of $\pi_k|\mathbb{C}$ is

$$E(\pi_k|\mathbb{C}) = \frac{1 + \rho_k}{K - 1 + \sum_{i \neq k} \rho_i},$$

and its conditional variance is

$$\text{Var}(\pi_k|\mathbb{C}) = \frac{(1 + \rho_k)(K - 1 + \sum_{i \neq k} \rho_i)}{(K + \sum_{i=1}^{K} \rho_i)^2 (1 + K + \sum_{i=1}^{K} \rho_i)}.$$

For the adherence of each data source to the overall clustering, α_m, recall the conditional distribution of $\alpha_m|\mathbb{C}, \mathbb{L}_m$ is

$$\alpha_m|\mathbb{C}, \mathbb{L}_m \sim \text{TBeta}(a_m + \tau_m, b_m + N - \tau_m, \frac{1}{K}) = \text{TBeta}(1 + \tau_m, 1 + N - \tau_m, \frac{1}{K}).$$

Let $B(x; a, b) = \int_0^x t^{a-1}(1 - t)^{b-1} dt$ be the incomplete Beta function. Then the conditional first and second moments are variance are

$$\begin{aligned}
E[\alpha_m|\mathbb{C}, \mathbb{L}_m] &= \frac{B(\frac{1}{K}; 2 + \tau_m, 1 + N - \tau_m) - B(1; 2 + \tau_m, 1 + N - \tau_m)}{B(\frac{1}{K}; 1 + \tau_m, 1 + N - \tau_m) - B(1; 1 + \tau_m, 1 + N - \tau_m)} \\
E[\alpha_m^2|\mathbb{C}, \mathbb{L}_m] &= \frac{B(\frac{1}{K}; 3 + \tau_m, 1 + N - \tau_m) - B(1; 3 + \tau_m, 1 + N - \tau_m)}{B(\frac{1}{K}; 1 + \tau_m, 1 + N - \tau_m) - B(1; 1 + \tau_m, 1 + N - \tau_m)} \\
\text{Var}(\alpha_m|\mathbb{C}, \mathbb{L}_m) &= \frac{B(\frac{1}{K}; 3 + \tau_m, 1 + N - \tau_m) - B(1; 3 + \tau_m, 1 + N - \tau_m)}{B(\frac{1}{K}; 1 + \tau_m, 1 + N - \tau_m) - B(1; 1 + \tau_m, 1 + N - \tau_m)} \\
&\quad - \{\frac{B(\frac{1}{K}; 2 + \tau_m, 1 + N - \tau_m) - B(1; 2 + \tau_m, 1 + N - \tau_m)}{B(\frac{1}{K}; 1 + \tau_m, 1 + N - \tau_m) - B(1; 1 + \tau_m, 1 + N - \tau_m)}\}^2.
\end{aligned}$$

Therefore, the uncertainty of the overall clustering is

$$
\begin{aligned}
&\mathrm{Var}(\pi_k\alpha_1\alpha_2)\\
&= \mathrm{Cov}(\pi_k^2, \alpha_1^2\alpha_2^2) + \{\frac{(1+\rho_k)(K-1+\sum_{i\neq k}\rho_i)}{(K+\sum_{i=1}^{K}\rho_i)^2(1+K+\sum_{i=1}^{K}\rho_i)} + [\frac{1+\rho_k}{K-1+\sum_{i\neq k}\rho_i}]^2\}\times\\
&\{\mathrm{Cov}(\alpha_1^2, \alpha_2^2) + (\frac{B(\frac{1}{K};3+\tau_1,1+N-\tau_1)-B(1;3+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)}\\
&-\{\frac{B(\frac{1}{K};2+\tau_1,1+N-\tau_1)-B(1;2+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)}\}^2+\\
&[\frac{B(\frac{1}{K};2+\tau_1,1+N-\tau_1)-B(1;2+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)}]^2)\times\\
&(\frac{B(\frac{1}{K};3+\tau_2,1+N-\tau_2)-B(1;3+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)}\\
&-\{\frac{B(\frac{1}{K};2+\tau_2,1+N-\tau_2)-B(1;2+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)}\}^2)+\\
&[\frac{B(\frac{1}{K};2+\tau_2,1+N-\tau_2)-B(1;2+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)}]^2\}
\end{aligned}
$$

We do not have expressions for the covariance term because that would require the derivation of the joint distributions between the variables in the covariance terms, which we need to figure out how to derive. That is an area of future work (Chen et al., 2018).

3.3.5 Simulations

We demonstrate our derivations through two simulation studies that both generate two datasets and data points that belong in two clusters. In the first simulation, we generate data where the two clusters have high separability, and one dataset has perfect adherence to the overall clustering, while the other dataset has no relationship to the overall clustering. In the second simulation, the clusters are much less separated and the boundary between the two clusters is much more muddled. Also, both datasets have the same level of adherence to the overall clustering.

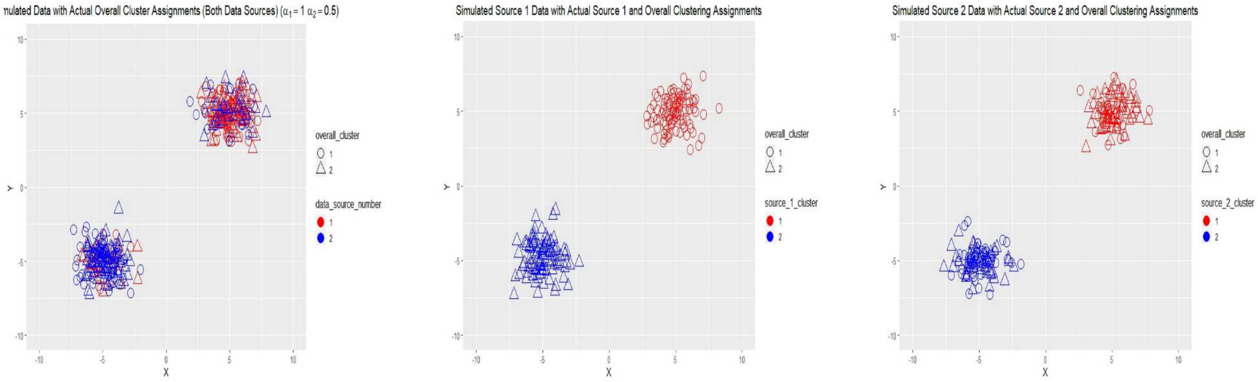
Simulation One

Following the simulation setup in section 3 of Lock et al. (2013), we generate two simulated datasets, denoted as $\mathbb{X}_1$ and $\mathbb{X}_2$, each with 200 observations ($N = 200$) and each is a two-dimensional vector. We generate the simulated datasets $\mathbb{X}_1 : 2 \times 200$ and $\mathbb{X}_2 : 2 \times 200$ as follows:

1. Let $\mathbb{C}$ define two clusters, where $C_n = 1$ for $n \in \{1, ..., 100\}$ and $C_n = 2$ for $n \in \{101, ..., 200\}$.
2. Set $\alpha_1 = 1$ (perfect relationship) and $\alpha_2 = 0.5$ (no relationship).
3. For $m = 1, 2$ and $n = 1, ..., 200$, generate $L_{mn} \in \{1, 2\}$ with probabilities $P(L_{mn} = C_n) = \alpha$ and $P(L_{mn} \neq C_n) = 1 - \alpha$.
4. For $m = 1, 2$, draw values X_{mn} from a $N_2([5, 5]', I_2)$ distribution if $L_{mn} = 1$ and from a $N_2([-5, -5]', I_2)$ distribution if $L_{mn} = 2$.

Note that I_2 is an identity matrix of size 2×2. We run the BCC method to obtain overall and source-specific clusterings of two clusters each ($K = 2$), and we run 10,000 MCMC iterations using the bayesCC package in R, which is associated with Lock and Dunson (2013a). The point estimates of the parameters are the maximum a posteriori (MAP) estimates, which is taken by averaging the estimated values of the parameters after the burn-in sample to the last iteration. Since the burn-in is half of the number of iterations, the MAP estimate is the average of the estimated values over the last 5,000 MCMC iterations.

Below in Figure 3.12 are visualizations of the simulated data with the actual overall clustering and the source-specific clusterings for $\mathbb{X}_1$ and $\mathbb{X}_2$. We will compare the estimated clusterings from the BCC method and the quantified uncertainty to these results.



Overall Source 1 Source 2

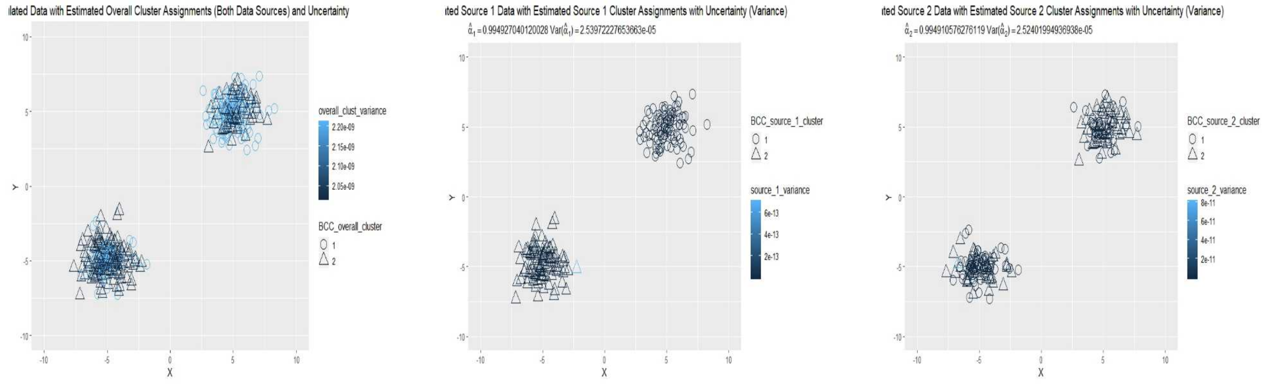
Figure 3.12: Simulated Data and Clusterings

Below are the estimated adherences of sources one and two to the overall clustering, denoted by α_1 and α_2, respectively, as well as their variances to measure the uncertainty of the adherence values for the two data sources.

$\hat{\alpha}_1 = 0.994927$ Var$(\hat{\alpha}_1)$=2.5397×10^{-5}
$\hat{\alpha}_2 = 0.9949106$ Var$(\hat{\alpha}_2)$=2.52402×10^{-5}

Below in Figure 3.13 are visualizations of the overall and source-specific clusterings estimated by BCC and the point-wise clustering uncertainties. The estimated adherences and the uncertainties of the adherences for the two data sources are both very similar. The primary difference between the two data sources is in the variance of the clusterings. We see that for source one, most of the variances in the clusterings are between 2×10^{-13} and 4×10^{-13}. On the other hand, for source two, most of the variances in the clusterings are between 2×10^{-11} and 4×10^{-11}. Thus, the variance in the clusterings of source two have the most influence on the uncertainties in the overall clustering.

Since the estimated adherences for both data sources are near one, we see that the overall clustering is pretty similar to the source-specific clusterings of data sources one and two. While we have not yet done an extensive analysis as to why the estimated adherence for source two (true $\alpha_2 = 0.5$) is so different than the true value, it appears this may be contradictory to the established accuracy of the estimates discussed in section 3.1 of Lock et al. (2013). However, in our simulation, the source-clustering errors (proportion of differences between the actual source-specific clusterings and the BCC estimated source-specific clusterings) and the overall error (proportion of differences between the actual overall clusterings and the BCC estimated overall clusterings) are pretty low (almost zero), which is consistent with the findings in section 3.2 of Lock et al. (2013).



Overall Source 1 Source 2

Figure 3.13: BCC Estimated Clusterings and Uncertainty

Simulation Two

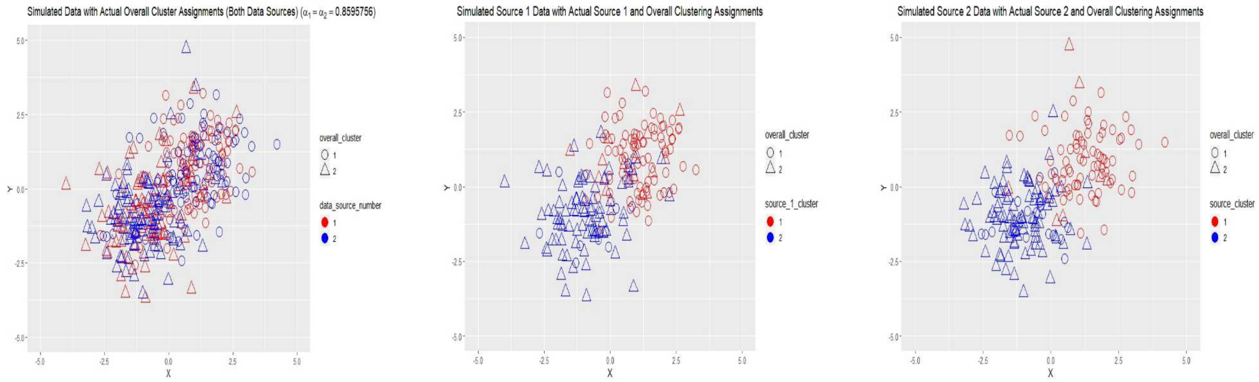
We generate the simulated datasets $\mathbb{X}_1 : 2 \times 200$ and $\mathbb{X}_2 : 2 \times 200$ as follows:

1. Let $\mathbb{C}$ define two clusters, where $C_n = 1$ for $n \in \{1, ..., 100\}$ and $C_n = 2$ for $n \in \{101, ..., 200\}$.
2. Draw α from a Uniform(0.5,1) distribution. Let $\alpha_1 = \alpha_2 = \alpha$. The true $\alpha = 0.8595756$.
3. For $m = 1, 2$ and $n = 1, ..., 200$, generate $L_{mn} \in \{1, 2\}$ with probabilities $P(L_{mn} = C_n) = \alpha$ and $P(L_{mn} \neq C_n) = 1 - \alpha$.

4. For $m = 1, 2$, draw values X_{mn} from a $N_2([1.5, 1.5]', I_2)$ distribution if $L_{mn} = 1$ and from a $N_2([-1.5, -1.5]', I_2)$ distribution if $L_{mn} = 2$.

Note that I_2 is an identity matrix of size 2×2. We run the BCC method to obtain overall and source-specific clusterings of two clusters each ($K = 2$), and we run 10,000 MCMC iterations using the bayesCC package. The point estimates of the parameters are the MAP estimates.

Below in Figure 3.14 are visualizations of the simulated data with the actual overall clustering and the source-specific clusterings for $\mathbb{X}_1$ and $\mathbb{X}_2$. As compared to Simulation One, the boundaries between the clusters are much less distinct because the centroids of the clusters are much closer to one another. We will compare the estimated clusterings from the BCC method and the quantified uncertainty to these results.



Overall Source 1 Source 2

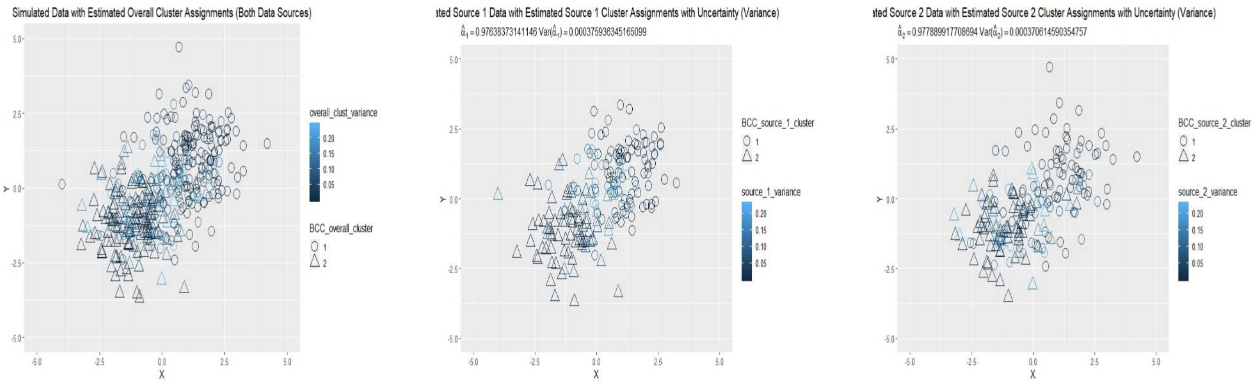
Figure 3.14: Simulated Data and Clusterings

Below are the estimated adherences of sources one and two to the overall clustering, denoted by α_1 and α_2, respectively, as well as their variances to measure the uncertainty of the adherence values for the two data sources.

$\hat{\alpha}_1 = 0.97638$	$\text{Var}(\hat{\alpha}_1)$=0.0003759
$\hat{\alpha}_2 = 0.9778899$	$\text{Var}(\hat{\alpha}_2)$=0.0003706

Below in Figure 3.15 are visualizations of the overall and source-specific clusterings estimated by BCC and the point-wise clustering uncertainties. The estimated adherences and the uncertainties of the adherences for the two data sources are fairly similar. The estimated adherence for source one ($\hat{\alpha}_1 = 0.97638$) is slightly lower than that of source two ($\hat{\alpha}_2 = 0.9778899$), with the estimated adherence uncertainty for source one ($\text{Var}(\hat{\alpha}_1)$=0.0003759) beings slightly higher than that for source two ($\text{Var}(\hat{\alpha}_2)$=0.0003706). From these estimated statistics, it is very difficult to tell if one source has more influence than the other on the overall clustering. When we look at the visualizations in Figure 3.15, both data sources have seemingly equal correspondence to the overall clustering. We note that compared to Simulation One, the magnitudes of uncertainties for the clustering probabilities and the source adherences are both higher. This is an expected result, as the two clusters have much more overlap than in Simulation One, where the clusters are much more distinct.

While we have not yet done an extensive analysis as to why the estimated adherences for both sources (true $\alpha_2 = 0.8595756$) are noticeably different than the true value, it appears this may be contradictory to the established accuracy of the estimates discussed in section 3.1 of Lock et al. (2013). However, in our simulation, the source-clustering errors (proportion of differences between the actual source-specific clusterings and the BCC estimated source-specific clusterings) and the overall error (proportion of differences between the actual overall clusterings and the BCC estimated overall clusterings) are both around 15%, which is consistent with the findings in section 3.2 of Lock et al. (2013).



Overall

Source 1

Source 2

Figure 3.15: BCC Estimated Clusterings and Uncertainty

3.3.6 Experiments with Philadelphia Multimodal Imagery

We obtain source-specific and consensus clusterings of the Philadelphia imagery data using BCC. We have two data sources (images), so $M = 2$. Each image is of size 100×100 pixels for a total of 10,000 pixels, so $N = 10,000$. We use BCC with six clusters for each image and the overall clustering, so $K = 6$. This is implemented using the bayesCC package. We run 1,000 MCMC iterations to approximate the posteriors. The point estimates of the parameters are the maximum a posteriori (MAP) estimates, taken at the modes of the conditional posterior distributions.

Below in Figure 3.16 are visualizations of the overall clustering when we factor in both the optical and lidar images. The three visualizations are the cluster assignments of each pixel (a), the variance of the sampling distribution of the probability that the pixel belongs in the assigned cluster (white represents higher variance and thus, higher uncertainty, while black represents lower variance and thus, lower uncertainty) (b), and the variance overlayed with the cluster assignments (c). We see that there is greater variance around the boundary of the lawn and the building, as well as around where the boats are docked in the river. The overall clustering appears to do a good job of segmenting the trees in the upper left and lower right hand corners of the image, but it does not segment the different parts of the roof very well.

In Figure 3.16, as well as in the visualizations for the individual clusterings of the optical (Figure 3.17) and lidar (Figure 3.18) images, we also include the visualizations using the standard deviation and Shannon's entropy as measures of uncertainty. We have not yet derived mathematically the relationship between the standard deviations and the entropies of the overall clustering and the individual clusterings of the optical and lidar images, but we note this as an area of future work. However, we will still perform an empirical analysis of the visualizations as confirmation of what we have derived with the variance as the uncertainty measure.

Below in Figure 3.17 are the visualization for the clustering of the optical image. We compute the adherence value of the optical clustering to the overall clustering, which is represented by α, to be 0.646. The variance of α is 0.0001248. The model does a good job of capturing the shadows around the building, the roof, and the trees. It also does a better job of separating the boats from the water, but it seems to cluster the boats with the pavement, even though the boats are whiter than the pavement.

Below in Figure 3.18 are the visualization for the clustering of the optical image. We compute the adherence value of the lidar clustering to the overall clustering, which is represented by α, to be 0.721. The variance of α is 0.0001689. We see that the model's clustering of the lidar image looks very similar to the original lidar image. It segments the trees and the portions of the building's roof the same as the original image. It also differentiates the pavement from the water.

The adherence values of the optical and lidar clusterings and their variances indicate the overall clustering adheres more to the lidar clustering, and the level of adherence has greater uncertainty than we see for the adherence between the optical and overall clusterings. This is confirmed by the similarity in the visualizations for the overall clustering (Figure 3.16) and the lidar clustering (Figure 3.18). The overall clustering can differentiate between the trees and the lawn and pavement that surrounds the trees, which is what we see in the optical image. However, since the overall clustering adheres more to the lidar clustering, it does not capture the shadows around the building and the different sections of the roof as well as the optical image clustering. For this particular dataset, it appears that the lidar image has more influence and value towards our overall knowledge of this scene in Philadelphia. This is confirmed with the standard deviation and entropy plots, as well.

3.3.7 Discussion and Future Work

We present preliminary work on mathematically relating the uncertainty in the results of source-specific clusterings to the uncertainty of the results of an overall consensus clustering. For the Philadelphia imagery example, the source-specific clusterings are the individual and separate clusterings of the optical and lidar images, and the overall consensus clustering is one clustering that takes into account the two individual clusterings of the images. Using variance as our measure for uncertainty and the BCC framework, we find that the **overall clustering uncertainty is directly proportional to the uncertainties in the adherences of each source-specific clustering to the overall clustering**. We confirm this through visualizing the results of the clusterings and seeing how the uncertainty results for the source-specific clustering and the uncertainty of the estimated adherences factors into the uncertainty in the overall consensus clustering.

There are many avenues of future work that needs to be done. We have previously identified deriving all of the mathematical expressions when relating the variance of the overall clustering to the variance of the source-specific clusterings, as well as the derivations for any number of M data sources. We have also previously identified deriving the relationship between the standard deviation and the entropy of the overall clustering and the source-specific clusterings. Related to the BCC implementation, we can try other distributional assumptions other than the ones specified in section 3.3.3. For example, we can try nonparametric distributions, where we let the data dictate the distributions of the clusters. We can also investigate the case when we have different numbers of clusters for each data source and the overall clustering, and when the semantics meanings of the clusters are different for all the clusterings. There is also the issue of computational scalability when implementing the BCC.

Another area of future work is a frequentist approach for consensus clustering and uncertainty quantification. The closest frequentist method for doing consensus clustering is the Probabilistic Feature Fusion (PFF) (Simonson, 1998; Simonson et al., 2017). It is a method that combines evidence arising from multiple features and classifiers expressed in the form of (generally dependent) hypothesis tests. It is used in one-class classification problems, where we want to find one target class. Here is a summary of how PFF works.

Suppose our data consists of N features each of length K. Let X_i, $i = 1, ..., K$, denote the $i$the feature. The marginal probability distribution function of X_i for the target class is denoted F_i. For X_i drawn from the target distribution, the quantity the quantity $F_i(X_i)$ will be uniformly distributed between zero and one (also written as $U[0, 1]$). For each feature, compute the p-value $1 - F_i(X_i)$, and then transformation the p-values using the transformation $Y_i = -\log(1 - F_i(X_i))$, which has a standard exponential distribution under the null hypothesis. Next, Sum the transformed values

$$S_{\text{fused}} = \sum_{i=1}^{K} Y_i,$$

which follows the gamma distribution with shape parameter $\alpha = K$ and $\beta = 1$ when the individual tests are independent. However, since the tests are likely dependent, we have an approximate gamma distribution. Let $\hat{r}_{ij}$ be the sample correlation coefficient between exponential random variables Y_i and Y_j. We estimate the mean and variance of the sum S_{fused} with the quantities $\hat{E}_k$ and $\hat{V}_k$, respectively:

$$\hat{E}_K = K$$

$$\hat{V}_K = K + 2\sum_{i=1}^{K}\sum_{j>i}\hat{r}_{ij}$$

The estimates of the shape and rate parameters are

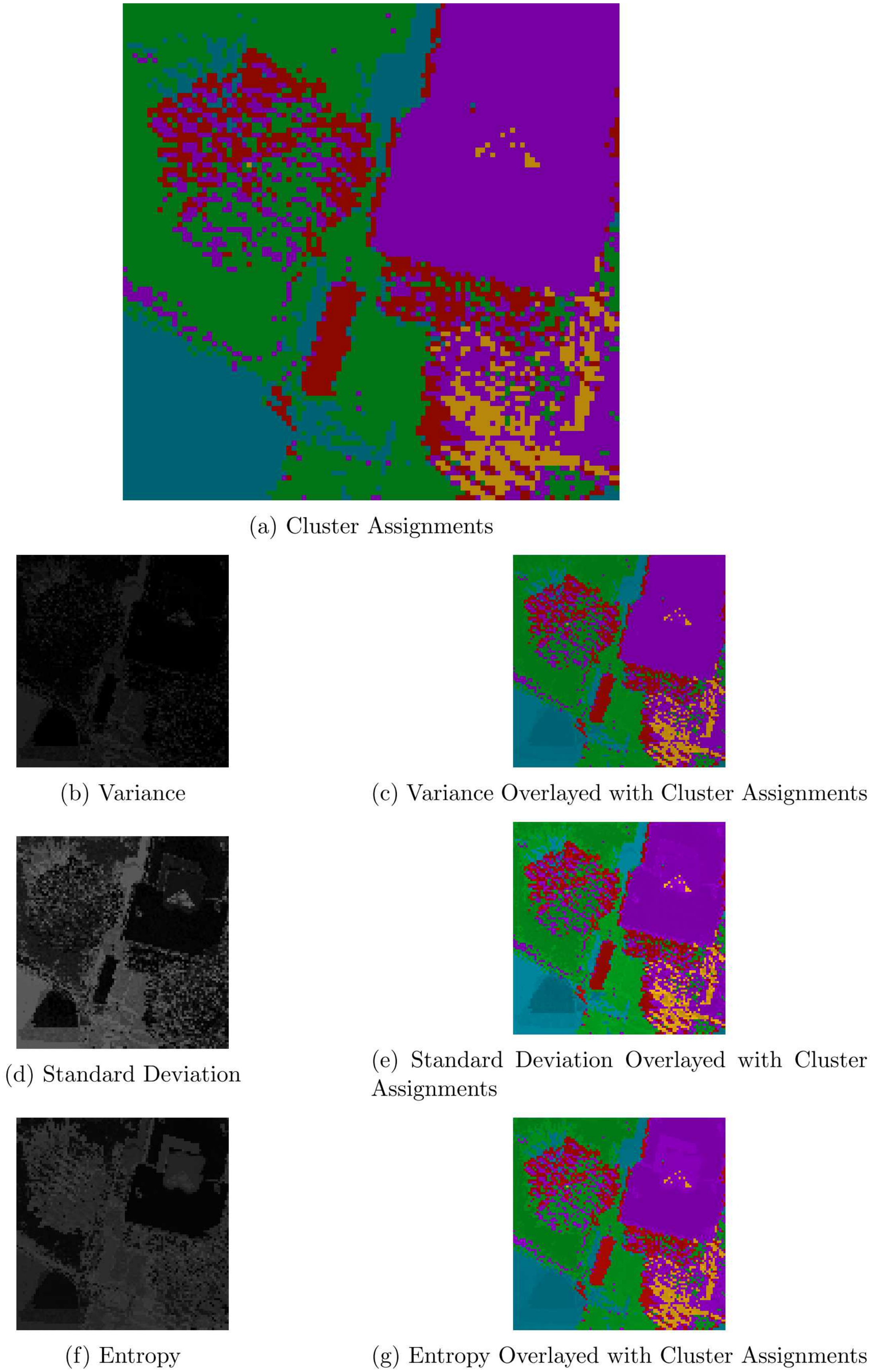
$$\hat{\alpha} = \frac{\hat{E}_K^2}{\hat{V}_K}$$

$$\hat{\beta} = \frac{\hat{E}_K}{\hat{V}_K}.$$

Finally, compute the fused p-value as $P_{\text{fused}}(s_{\text{fused}}) = 1 - F_{\text{fused}}(s_{\text{fused}})$ (Simonson, 1998; Simonson et al., 2017).

Unfortunately, PFF does not address certain issues related to multimodal image segmentation uncertainty quantification. First, PFF only works when you are testing for the same class. In the context of inference, you must test for the same hypothesis. Unfortunately, when we fit probabilistic clustering models, such as the GMM, NMM, or BCC, to each image, even if we cluster each image with the same number of clusters, the semantic meanings of the clusters in each image will likely be different. Therefore, in order to formulate a frequentist method for multimodal uncertainty quantification, whether or not we extend the PFF method, we have to address three questions:

1. How do we formulate a frequentist framework for fusing p-values together when they are testing different hypotheses, but they are related because they are describing the same scene of interest?
2. Can we do the same type of fusing for uncertainty measures, such as variance or entropy?
3. Is there a frequentist method for doing consensus clustering where information is shared between the overall clustering and the source-specific clusterings?



(a) Cluster Assignments

(b) Variance

(c) Variance Overlayed with Cluster Assignments

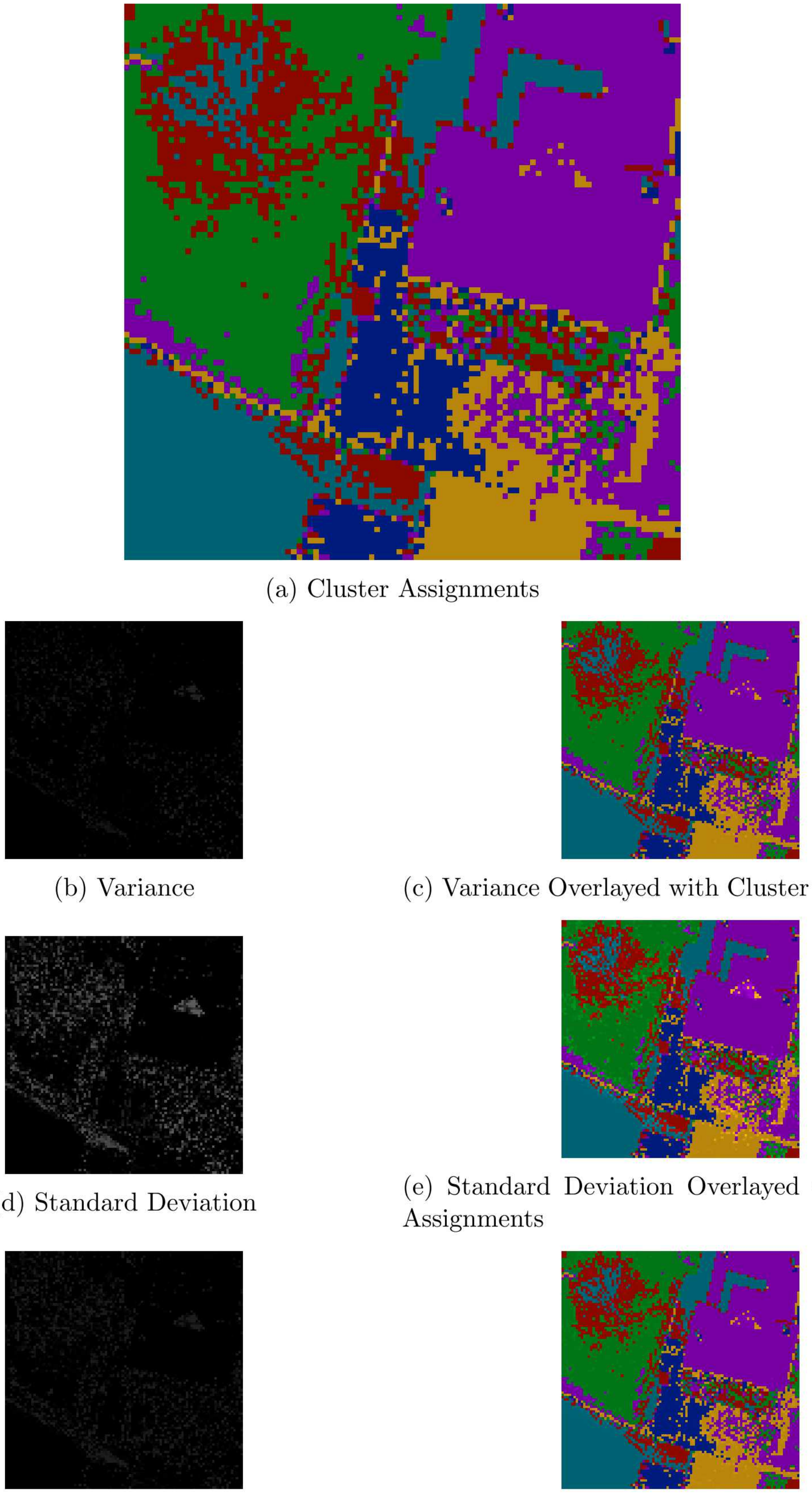
(d) Standard Deviation

(e) Standard Deviation Overlayed with Cluster Assignments

(f) Entropy

(g) Entropy Overlayed with Cluster Assignments

Figure 3.16: BCC Overall Clustering and Uncertainty Results Combining Optical and Lidar Images for Philadelphia Data



(a) Cluster Assignments

(b) Variance

(c) Variance Overlayed with Cluster Assignments

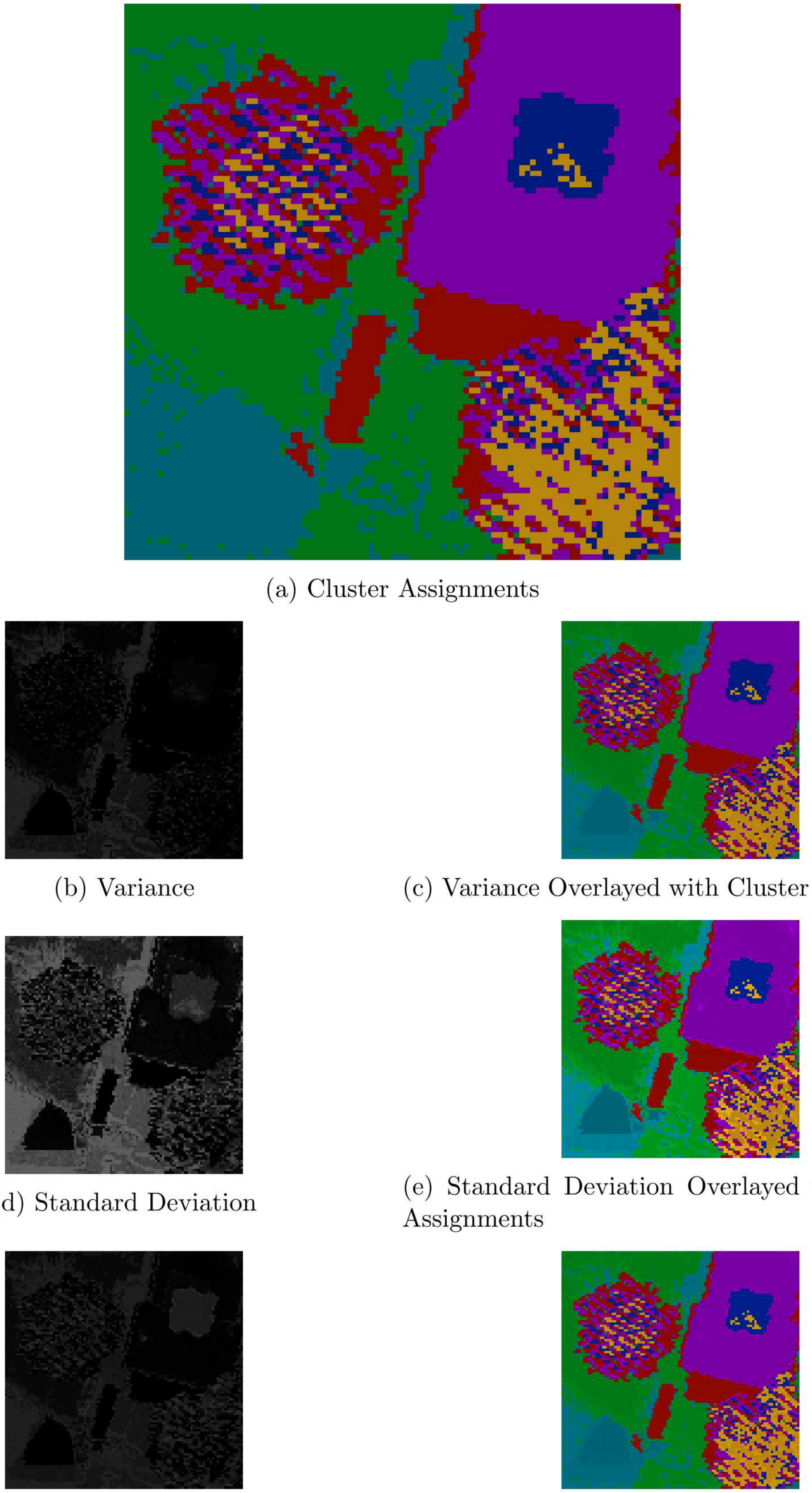
(d) Standard Deviation

(e) Standard Deviation Overlayed with Cluster Assignments

(f) Entropy

(g) Entropy Overlayed with Cluster Assignments

Figure 3.17: BCC Clustering and Uncertainty Results for Optical Image



(a) Cluster Assignments

(b) Variance

(c) Variance Overlayed with Cluster Assignments

(d) Standard Deviation

(e) Standard Deviation Overlayed with Cluster Assignments

(f) Entropy

(g) Entropy Overlayed with Cluster Assignments

Figure 3.18: BCC Clustering and Uncertainty Results for Lidar Image

Chapter 4

Discussion and Conclusion

From our preliminary work, we have learned several lessons, including that there is still much work to be done to improve the effectiveness and efficiency of nonparametric methods, as well as integrating semantic information into probabilistic clustering methods.

From our preliminary results when applying the nonparametric modal clustering method to the Vee imagery data (section 3.1), as well as our implementations of the nonparametric mixture model (NMM) method to the Philadelphia imagery data (section 3.2), to which we have also applied the Gaussian mixture model (GMM) (Stracuzzi et al., 2017b), we need to estimate the cluster probabilities as part of the clustering algorithm in order to obtain robust estimates of the cluster probabilities. From the implementations of the NMM and GMM methods, we learn that uncertainty arises not only from the model choice, but also from the model implementation. This is evident from the much more ambiguous clusterings offered by the NMM results.

In particular, for nonparametric methods, the need for being able to efficiently and precisely estimate the cluster densities is vital to obtaining sound clustering and uncertainty results. An added difficulty seen with nonparametric methods is distributional statistics that can adequately describe parametric distributions, such as the mean, covariance, and moments, are not enough for describing the distribution of the data and the sampling distribution of a parameter of interest, such as the cluster probability.

Related to nonparametric probabilistic clustering, one area of future work is improving the implementation of nonparametric methods, both in terms of computational speed with fitting nonparametric probabilistic clustering models and with best capturing the cluster distributions. Another area of future work is best describing the nonparametric density estimates of the cluster distributions. Unlike the GMM, when each cluster follows a multivariate normal distribution and the distribution be adequately described by the mean vector and covariance matrix, global distribution statistics do not accurately capture all characteristics of a nonparametric distribution.

For the BCC method, our uncertainty analysis represents a new contribution from existing research done with multiple data set analysis. Uncertainty analysis is not done with any of the existing consensus clustering methods that are cited in the first part of section 3.3. There have been a number of techniques proposed for dimension reduction of multi-view data. However, they assume that all data views consists of the same types of variables, such as continuous data, so they are not very suitable for mixed multi-view data (Cao et al., 2008; Witten and Tibshirani, 2009; Acar et al., 2011; Lock et al., 2013; Di et al., 2009). There have also been a number of papers published on data for genomic data, including the use of graphical models (Morris and Baladandayuthapani, 2017; Yang et al., 2012, 2014a,b). See Allen (2017) for a more comprehensive discussion. However, none of these methods then examine the uncertainty of the performance of the models with jointly analyzing multiple sources of data.

When we compare the results of both the nonparametric mixture model and BCC, we see that there are more distinct clusterings with the BCC method. This leads to the belief that when we iterate and borrow information between the source-specific and consensus clusterings, we can get more accurate clusterings and uncertainty information. This also calls in to question whether or not the nonparametric mixture model method is the best method, while also raising the question of what results we get if we do a nonparametric implementation of the BCC method. Finally, this calls into question whether or not concatenating data sets is truly the appropriate way for combining data sets, and if clustering the concatenated data set is the best way of obtaining a consensus clustering when we have multiple data sets. Therefore, one area of future work is a nonparametric implementation of BCC. To formulate this implementation, we will need to look into existing Bayesian nonparametric methods such as Bayesian nonparametric models (Orbanz and Teh, 2011) and latent Dirichlet allocation (Blei et al., 2003) to see if we can adapt the BCC method for use with nonparametric distributions.

In addition to the three future research directions we have mentioned above addressing current open problems with nonparametric probabilistic clustering and the BCC methods, there are many more areas of future work. A fourth area of future work is a method to compare the performance of the clustering algorithms and uncertainties in order to determine if a particular clustering model is better than another, particularly in the unconstrained supervised learning task when each data source's clustering can have a different number of clusters and the clusters have different semantic meanings. A fifth area of future work is a probabilistic formulation for hard clustering algorithms, where the clustering probabilities are not estimated as part of the clustering algorithm, such as in the nonparametric modal clustering method. These methods have been shown to achieve good clustering results; however, they don't carry any probabilistic information with their algorithmic output, which does not allow for uncertainty quantification of the clustering results. This would help to address the current shortcomings with obtaining robust estimates of the cluster probabilities in the nonparametric modal clustering method, and it would help ensure robust cluster probability estimates for all hard clustering algorithms. A sixth area of future work is expanding the current work on quantifying uncertainty of the clustering assignment of a specific pixel to quantifying the uncertainty of an entire cluster's boundary on an image level. Finally, there are many sources of variability and dependency with multimodal imaging data, such as the dependency between the components of the feature vector at each pixel, the dependency between pixels, and dependency between data sources, and the variance in all of these data sources. This presents many additional possible sources of uncertainty that needs to be accounted for. Therefore, a final area of future work is being able to decompose the uncertainty we have computed and attributing various uncertainty measures to their respective sources. This would provide even more valuable information to decision-makers when evaluating analysis results.

Nonparametric distributional assumptions have the potential to allow more accurate modeling of available multiple data sets and can be integrated in to methods for clustering multiple data sets. Allowing the data to estimate the distributions of the data and the clusters can provide more accurate segmentations relative to the actual semantic features of the data. Furthermore, when clustering multiple data sets and seeking to determine the value of information from each data set towards an overall consensus clustering, iterating between all of the source-specific clusterings and the overall consensus clustering seems to obtain the best results.

References

Acar, E., Kolda, T. G., and Dunlavy, D. M. (2011). All-at-once Optimization for Coupled Matrix and Tensor Factorizations. *ArXiv e-prints*.

Allen, G. I. (2017). Statistical data integration: Challenges and opportunities. *Statistical Modeling*, 17(4-5):332–337.

Benaglia, T., Chauveau, D., and Hunter, D. (2011). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger, World Scientific Publishing Co.*, pages 15–27. ¡hal-00353297¿.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Cao, K.-A. L., Rossouw, D., Robert-Grani, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 35.

Chen, M. G., Stracuzzi, D. J., and Darling, M. C. (2018). A mathematical framework for uncertainty quantification in multimodal image analysis via probabilistic clustering models. Joint Statistical Meetings, Vancouver, BC, Canada.

Darling, M. C. and Stracuzzi, D. J. (2018). Toward uncertainty quantification for supervised classification. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia National Lab.(SNL-CA), Livermore, CA (United States).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.*, 3(1):458–488.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297.

Kormaksson, M., Booth, J. G., Figueroa, M. E., and Melnick, A. (2012). Integrative model-based clustering of microarray methylation and expression data. *Ann. Appl. Stat.*, 6(3):1327–1347.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, 8:1687–1723.

Lock, E. F. and Dunson, D. B. (2013a). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Lock, E. F. and Dunson, D. B. (2013b). Supplement to "bayesian consensus clustering". *Bioinformatics*.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, 7(1):523–542.

Löfstedt, T. and Trygg, J. (2011). Onplsa novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, 25(8):441–455. Article first published online: 25 APR 2011.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Power, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):4245–4250.

Morris, J. S. and Baladandayuthapani, V. (2017). Statistical contributions to bioinformatics: Design, modelling, structure learning and integration. *Statistical Modelling*, 17(4-5):245–289.

Nguyen, N. and Caruana, R. (2007). Consensus clusterings. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 607–612, Washington, DC, USA. IEEE Computer Society.

Orbanz, P. and Teh, Y. W. (2011). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376.

Rey, M. and Roth, V. (2012). Copula mixture model for dependency-seeking clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 275–282, USA. Omnipress.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., and Wiley, H. S. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, 24(24):2894–2900.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837.

Savage, S., Ghahramani, Z., Griffin, J. E., Cruz, B. J. D. L., Wild, D. L., Savage, R. S., Ghahramani, Z., Griffin, J. E., Cruz, B. J. D. L., and Wild, D. L. (2010). Discovering transcriptional modules by bayesian data integration. *Bioinformatics*.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Simonson, K. M. (1998). Probabilistic fusion of atr results. Technical Report SAND98-1699, Sandia National Laboratories.

Simonson, K. M., Derek West, R., Hansen, R. L., LaBruyere, T. E., and Van Benthem, M. H. (2017). A statistical approach to combining multisource information in one-class classifiers. *Stat. Anal. Data Min.*, 10(4):199–210.

Stracuzzi, D. J., Chen, M. G., Darling, M. C., and Peterson, M. G. (2018). Data-driven uncertainty quantification for multi-sensor analytics. In *SPIE Proceedings*.

Stracuzzi, D. J., Chen, M. G., Darling, M. C., Peterson, M. G., and Vollmer, C. (2017a). Uncertainty quantification for machine learning. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia National Lab.(SNL-CA), Livermore, CA (United States).

Stracuzzi, D. J., Chen, M. G., Darling, M. C., Peterson, M. G., and Vollmer, C. (2017b). Uncertainty quantification for machine learning. Technical Report SAND2017-6776, Sandia National Laboratories.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *Ann. Statist.*, 20(3):1236–1265.

Wang, H., Shan, H., and Banerjee, A. (2011). Bayesian cluster ensembles. *Stat. Anal. Data Min.*, 4(1):54–70.

Wang, P., Domeniconi, C., and Laskey, K. B. (2010). Nonparametric bayesian clustering ensembles. In Balcázar, J. L., Bonchi, F., Gionis, A., and Sebag, M., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 435–450, Berlin, Heidelberg. Springer Berlin Heidelberg.

Witten, D. M. and Tibshirani, R. J. (2009). Extensons of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8:1–27.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1358–1366. Curran Associates, Inc.

Yang, E., Baker, Y., Ravikumar, P., Allen, G., and Liu, Z. (2014a). Mixed Graphical Models via Exponential Families. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1042–1050, Reykjavik, Iceland. PMLR.

Yang, E., Ravikumar, P., Allen, G. I., Baker, Y., Wan, Y.-W., and Liu, Z. (2014b). A general framework for mixed graphical models.

Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLOS Computational Biology*, 7(10):1–12.

Zhou, G., Cichocki, A., and Xie, S. (2012). Common and individual features analysis: Beyond canonical correlation analysis.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)

Sandia National Laboratories