A Hypothesis Test for the Evaluation of Soft Clusters for Classification

Rhonda D. Phillips,[1] **Layne T. Watson,**[1,2] **and Naren Ramakrishnan**[1]
Departments of Computer Science[1] and Mathematics[2], MC 0106
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Randolph H. Wynne
Department of Forestry, MC 0324
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Abstract – *In remote sensing and other disciplines, clustering is frequently used in classification to assign labels to data. In particular, the iterative guided spectral class rejection (IGSCR) classification algorithm uses labeled data and a statistical hypothesis test to determine which clusters should be used in classification. Rejected clusters (based on this evaluation method) are then refined. The hypothesis test used in IGSCR is based on the binomial distribution, which effectively models hard cluster and class memberships. This work proposes an analogous hypothesis test for soft cluster evaluation.*

Keywords: clustering, functional enrichment, statistical pattern recognition, remote sensing

1 Background

Clustering is frequently used in the classification of remotely sensed images in both unsupervised and hybrid classification methods. Unsupervised methods do not require prior information such as training data, but require that clusters be labeled with classes after clustering. Hybrid classification methods combine supervised and unsupervised techniques. One such method is the iterative guided spectral class rejection (IGSCR) classification method ([1], [2], [3]) that clusters data and uses labeled points to associate each cluster with a class automatically. Each cluster should correspond to only one class, but unfortunately, there is no guarantee that a cluster will be representative of only one class. IGSCR uses a statistical hypothesis test to determine whether a cluster should be associated with a class. If a cluster fails this test, it is subjected to further refinement. The test used in IGSCR is based on the binomial distribution, and is only appropriate for hard clusters. The purpose of this work is to develop a similar test for soft clusters created by fuzzy K-means that can be used in a soft version of IGSCR. A soft cluster evaluation method will have applicability in other methods that use soft clusters for classification.

The remainder of the paper is organized as follows. Section 2 describes fuzzy K-means, a soft clustering method used for the derivation of the soft cluster evaluation. Section 3 discusses evaluating clusters and Section 4 describes the statistical distribution used to model fuzzy cluster and class memberships for a sample. Section 5 introduces two potential hypothesis tests used for soft cluster evaluation, and Section 6 concludes the paper.

2 Fuzzy K-means

Consider a soft clustering algorithm that minimizes the objective function [4]

$$J(\rho) = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}^{p}\rho_{ij} \quad \text{subject to} \quad \sum_{j=1}^{K} w_{ij} = 1 \text{ for each } i \tag{1}$$

where n is the number of samples, K is the number of clusters, $w_{ij} \in (0,1)$ is the value in the ith row and jth column of the weight matrix $W \in \Re^{n\times K}$, $U^{(j)} \in \Re^{B}$ is the jth cluster prototype, $x^{(i)} \in \Re^{B}$ is the ith data sample, $p > 1$, and $\rho_{ij} = \rho(x^{(i)}, U^{(j)}) = ||x^{(i)} - U^{(j)}||_2^2$ is the Euclidean distance squared. The algorithm that minimizes this objective function first calculates

$$w_{ij} = \frac{(1/\rho_{ij})^{1/(p-1)}}{\sum_{k=1}^{K}(1/\rho_{ik})^{1/(p-1)}}$$

for all i and j followed by calculating updated cluster prototypes

$$U^{(j)} = \sum_{i=1}^{n} w_{ij}^{p} x^{(i)} \Big/ \sum_{i=1}^{n} w_{ij}^{p}.$$

This iteration (recalculation of the weights followed by recalculation of cluster prototypes, following by recalculation of the weights, etc.) is guaranteed to

converge (with these definitions of ρ_{ij}, $U^{(j)}$, and w_{ij}) for $p > 1$ [5].

3 Cluster Evaluation

A key component in the IGSCR clustering framework is the homogeneity test used to determine if a cluster contains a statistically significant proportion of one class. This test provides a basis for rejecting a cluster for further refinement. The test for cluster purity is performed using the labeled training set. Let $V_{c,j}$ be the binomial random variable denoting the number of labeled samples assigned to the jth cluster that are labeled with a particular cth class. Let p be the user-supplied cluster homogeneity threshold ($p = .9$ would indicate a cluster is 90% pure with respect to the majority class), and let α be the user-supplied acceptable one-sided Type-I error for a statistical hypothesis test. Then if c is the majority class represented in the jth cluster, the jth cluster is rejected if $P(Z < \hat{z}) < 1 - \alpha$ where Z is a standard normal random variable, m is the number of labeled samples in the jth cluster, and

$$\hat{z} = \frac{v_{c,j} - mp}{\sqrt{mp(1-p)}}. \quad (2)$$

(Typically a continuity correction of 0.5 is added in the numerator of (2).)

A cluster might be composed of more than one class because the cluster itself is in fact composed of more than one cluster. A cluster might also contain more than one class because the initial clusters were determined in such a way as to prevent a cluster from moving toward a particular class. It would be useful to determine which clusters are not spectrally pure (contain more than one class with high probability) so that the cluster can be further refined, and if no refinement is possible (any number of iteration ending criteria are met), the cluster should not be used in the classification model. Statistical hypothesis tests provide a mechanism for determining class purity once an appropriate statistical model is selected for the data.

With hard clustering, the notion of a pure cluster is clear. Each sample will belong to one and only one cluster. A cluster can be 100% homogeneous when all labeled samples contained within that cluster belong to only one class. Although this is possible, it is unlikely that one cluster contains only one class because of inherent error in the labeling process and because two different class categories can contain spectrally similar samples. Once a homogeneity level is determined, a rigorous hypothesis test can be applied to select clusters that contain a certain percentage of one class, with that percentage unlikely to be observed in a particular cluster randomly.

Using soft clusters introduces complications to assessing and determining cluster purity. The first question might be whether a soft cluster can be spectrally pure, because being soft might indicate that clusters are naturally comprised of multiple classes. However, just as hard clusters can be representative of just one predominant class, soft clusters can be representative of a dominant class. Soft clusters are composed of different portions of each sample or pixel within an image, meaning that each sample has a positive probability of being in different individual classes or clusters. When samples labeled with different classes have a positive probability of belonging to the same cluster, that does not indicate that the cluster really contains two different classes, but rather perhaps that while the pixels have strong associations with different classes, there is also a positive (although possibly small) probability that each pixel actually belongs to or partially belongs to the majority class within the cluster. Both cases (the cluster is confused or the cluster is not confused but the pixels labeled with different classes still have small associations with the same class) are possible in soft clustering. The appropriate test for soft clusters is not which pixels "belong" to a particular cluster (they all "belong" to some degree), rather how strongly pixels from different classes belong to a particular cluster. If pixels from only one class have strong associations with a cluster when compared to pixels labeled with other classes, then the cluster should be labeled with that most strongly associated class. In this manner, each pixel/sample is associated by varying degrees with multiple spectrally pure clusters that are mapped to individual classes, ultimately producing a soft classification output when each sample is then mapped to different individual classes with varying probabilities.

4 Distribution

Developing a hypothesis test to assess purity of clusters requires a random variable and knowledge of the distribution of that random variable. In IGSCR, a cluster can be considered pure and labeled with a class if the number of labeled samples belonging to the class is high compared to the number of labeled samples not belonging to the class. The random variable of interest, $V_{c,j} = \sum_{i \in I_j} V_{ic}$, is the count of the number of labeled samples belonging to the cth class for a particular jth cluster where i is the pixel index, I_j is the index set of labeled pixels in the jth cluster, and V_{ic} is the Bernoulli random variable corresponding to the ith pixel being associated with the cth class.

Using soft clustering, the random variable and distribution are more complicated as there are class

memberships (either 0 or 1) and cluster memberships (between 0 and 1). Building a test on only the class memberships is not useful as each labeled sample will have some positive probability of belonging to a particular cluster, making the results of the test the same for each cluster unless memberships are also considered. In this case, the association of a sample to a particular class (the majority class, for example) is still a Bernoulli trial. Each pixel also has a weight vector, $w_{i\cdot}$, indicating the probability of membership to each cluster. The random variable of interest is the sum of the memberships for the cth class and weights to the jth cluster,

$$Y_{c,j} = V_{1c}W_{1j} + V_{2c}W_{2j} + \cdots + V_{nc}W_{nj},$$

where n is the total number of labeled samples. The labels of the classified pixels are independent of cluster assignment, making an assumption that V_{ic} and W_{ij} are independent reasonable. Furthermore, the training samples are labeled prior to clustering, making the random variable of interest

$$Y_{c,j}|(V_{1c}, V_{2c}, \ldots, V_{nc}) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c},$$

where $\phi(i)$ is the label of the ith pixel, and

$$\delta_{\phi(i),c} = \begin{cases} 0 \text{ if } \phi(i) \neq c, \\ 1 \text{ if } \phi(i) = c, \end{cases}$$

is the Kronecker delta. The probability density function (pdf) of $Y_{c,j}|(V_{ic}, i = 1, \ldots, n) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c}$ is the pdf of a sum of individual cluster weights.

Figures 1 and 2 contain experimental frequency histograms of weights w_{ij} for two clusters ($K = 2$) of a satellite image. The distribution of the cluster weights appears to be multimodal, which is consistent with the data having multiple inherent classes, indicating that W_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, K$ would not be identically distributed. A closed form distribution is not readily available for W_{ij}, but a closed form distribution, or at least a reasonable approximate closed form distribution, for $W_{+j} = \sum_{i=1}^{n} W_{ij}$ exists.

4.1 Normal Approximation to $Y_{c,j}$

Suppose an image x contains n pixels $x^{(i)} \in \Re^B$, $i = 1, \ldots, n$. For K fixed cluster centers $U^{(k)} \in \Re^B$, $k = 1, \ldots, K$, the assigned weight of the ith pixel to the jth cluster is

$$w_{ij} = \frac{1/||x^{(i)} - U^{(j)}||_2^2}{1\Big/\sum_{k=1}^{K} ||x^{(i)} - U^{(k)}||_2^2},$$

which is the inverse of the distance squared over the sum of the inverse squared distances. (Such inverse distance

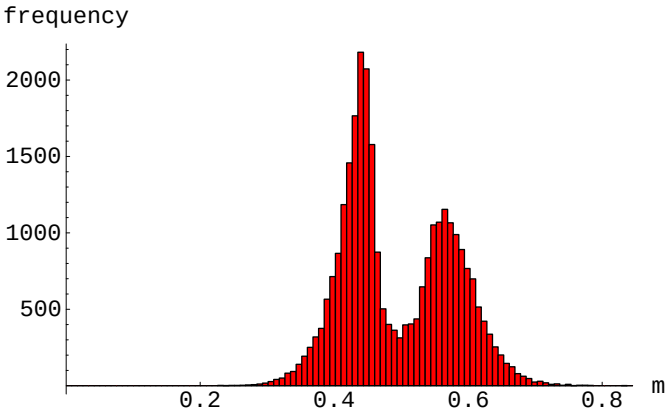

Figure 1. Histogram of cluster weights in one cluster, K=2.

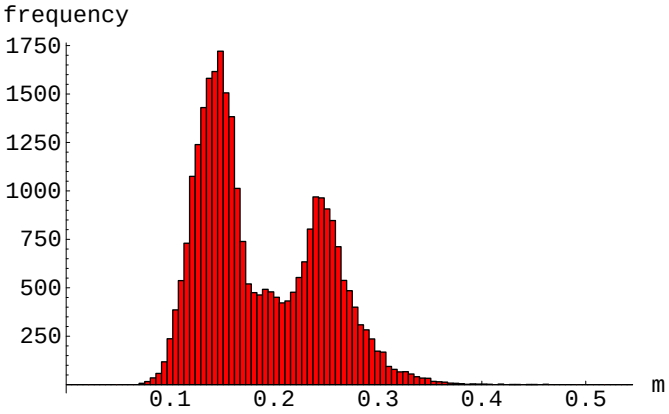

Figure 2. Histogram of cluster weights in one cluster, K=5.

weights are widely used, e.g., by Shepard's algorithm for sparse data interpolation.) Note this is the specific case in the soft clustering algorithm described above when $p = 2$. In this case where a remotely sensed image is to be clustered, it is reasonable to assume that $x^{(i)}$, $i = 1, \ldots, n$ are generated from a finite number of multivariate normal distributions. The act of clustering assumes that the data are generated from a finite number of distributions, and remotely sensed earth data are assumed to be generated from normal distributions. The following proof demonstrates that under these assumptions (pixels are generated from a finite number of normal distributions), the Lindeberg condition is satisfied and therefore the central limit theorem applies to the sum of a sequence of cluster weight random variables $\sum_{i=1}^{n} W_{ij}$. Let $q = \psi(i)$ denote the distribution from which $X^{(i)}$ was sampled.

Theorem: Let $X^{(i)}$, $i = 1, 2, \ldots$, be B-dimensional random vectors having one of Q distinct multivariate normal distributions. For $i = 1, 2, \ldots$ and $j = 1, \ldots, K$ define the random variables

$$W_{ij} = W_j(X^{(i)}) = \frac{1/||X^{(i)} - U^{(j)}||_2^2}{\sum_{k=1}^{K} 1/||X^{(i)} - U^{(k)}||_2^2},$$

where K is the number of clusters and $U^{(k)} \in \Re^B$ is the kth cluster center (and is considered fixed for weight calculation). Then for any $j = 1, \ldots, K$,

$$P\left\{\frac{1}{B_{nj}}\sum_{i=1}^{n}(W_{ij} - a_{ij}) < x\right\} \rightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}dz$$

as $n \rightarrow \infty$, where $a_{ij} = \mathrm{E}[W_{ij}]$, $b_{ij}^2 = \mathrm{Var}[W_{ij}]$, and $B_{nj}^2 = \sum_{i=1}^{n} b_{ij}^2$.

Proof. W_{ij} is a bounded ($0 \leq W_{ij} \leq 1$) measurable function of a normal random variable, and is therefore a random variable with finite mean and variance. Fix j for the remainder of the proof, and let $q = \psi(i)$ denote which of the Q distributions $X^{(i)}$ is from. In order to prove

$$P\left\{\frac{1}{B_{nj}}\sum_{i=1}^{n}(W_{ij} - a_{ij}) < x\right\} \rightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}dz,$$

it is sufficient to verify the Lindeberg condition [6]:

$$\lim_{n\rightarrow\infty}\frac{1}{B_{nj}^2}\sum_{i=1}^{n}\int_{|x-a_{ij}|>\tau B_{nj}}(x - a_{ij})^2 dF_{\psi(i),j}(x) = 0,$$

for any constant $\tau > 0$ where $F_{\psi(i),j}(x)$ is the cumulative distribution function for W_{ij}.

For each q, $1 \leq q \leq Q$, define $I_q = \psi^{-1}(q) = \{i \mid \psi(i) = q, 1 \leq i \leq n\}$, $n_q = |I_q|$, and for $i \in I_q$ let $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ and $\mathrm{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$. Now considering only the independent and identically distributed random variables W_{ij}, $i \in I_q$, the Lindeberg condition holds:

$$\lim_{n_q\rightarrow\infty}\frac{1}{n_q\beta_{qj}^2}\sum_{i\in I_q}\int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}}(x - \alpha_{qj})^2 dF_{qj}(x)$$

$$= \lim_{n_q\rightarrow\infty}\frac{1}{\beta_{qj}^2}\int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}}(x - \alpha_{qj})^2 dF_{qj}(x) = 0.$$

Since β_{qj} is positive and finite, and the integral is finite, the limit of the integral is zero as $\sqrt{n_q}\beta_{qj} \rightarrow \infty$.

$W_{ij}, i = 1, 2, \ldots,$ are random variables from Q iid distributions, $F_{qj}, q = 1, \ldots, Q$, where the mean of the qth distribution is α_{qj}, the variance is β_{qj}^2, and the number of random variables from that distribution is n_q, where $\sum_{q=1}^{Q} n_q = n$. As $n \rightarrow \infty$ there is at least one q for which $n_q \rightarrow \infty$. For this sequence of independent random variables from Q distributions, the Lindeberg condition is

$$\lim_{n\rightarrow\infty}\frac{1}{B_{nj}^2}\sum_{i=1}^{n}\int_{|x-a_{ij}|>\tau B_{nj}}(x - a_{ij})^2 dF_{\psi(i),j}(x)$$

$$= \lim_{n\rightarrow\infty}\frac{1}{\sum_{k=1}^{Q} n_k\beta_{kj}^2}\sum_{q=1}^{Q} n_q$$

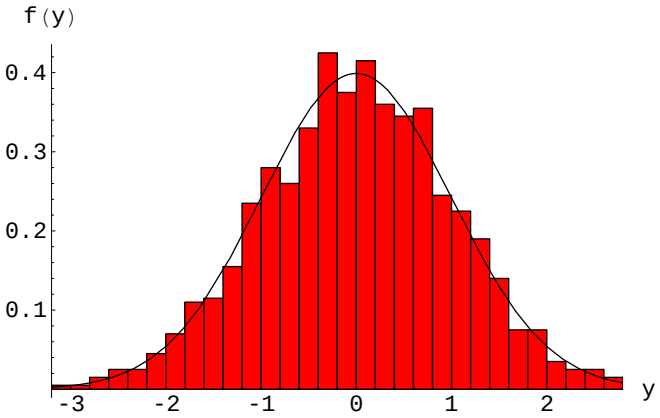

Figure 3. Pdf of Y (normalized) compared to a standard normal distribution.

$$\cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}}(x - \alpha_{qj})^2 dF_{qj}(x)$$

$$= \lim_{n\rightarrow\infty}\sum_{q=1}^{Q}\frac{n_q}{\sum_{k=1}^{Q} n_k\beta_{kj}^2}$$

$$\cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}}(x - \alpha_{qj})^2 dF_{qj}(x)$$

$$\leq \lim_{n\rightarrow\infty}\sum_{q=1}^{Q}\frac{1}{\beta_{qj}^2}\int_{|x-\alpha_{qj}|>\tau B_{nj}}(x - \alpha_{qj})^2 dF_{qj}(x) = 0.$$

Since each variance β_{qj}^2 is positive and finite, and $B_{nj} = \sqrt{n_1\beta_{1j}^2 + \cdots, +n_Q\beta_{Qj}^2} \rightarrow \infty$ as at least one $n_q \rightarrow \infty$, each integral converges to zero as $n \rightarrow \infty$, and the Lindeberg condition is verified. Q.E.D.

Remark: The assumption that the $X^{(i)}$, $i = 1, 2, \ldots,$ are generated from a finite number of normal distributions is stronger than necessary. This proof holds if $X^{(i)}$, $i = 1, 2, \ldots,$ are generated from a finite number of arbitrary distributions.

Experimental results match this theoretical result, as illustrated by one experiment in Figure 3.

5 Hypothesis Test

The hypothesis test used in IGSCR to assess the significance of a cluster association to a class is based on the normal approximation to the binomial distribution (2). The null hypothesis is that the true probability of a pixel belonging to the majority class (for the cluster of interest) is less than p_0, a user supplied value. If $P(Z > \hat{z}) < \alpha$, where α is the user provided Type-I error, then the null hypothesis is rejected. The null hypothesis corresponds to the case when the cluster is impure, and rejecting the null hypothesis equates with labeling the cluster pure; if the null hypothesis is

not rejected, the cluster is impure and the cluster is "rejected."

The hypothesis test for soft clusters is different as the Bernoulli trials are fixed and testing the probability p of a success is no longer relevant. A pure soft cluster should have large weights for the majority class and comparatively small weights for other classes. One possible hypothesis test compares the average weight for one particular cth class with the overall average weight for all classes in the jth cluster. Starting with the normal approximation for the sum of the cluster weights, the standard normal test statistic would be

$$\hat{z} = \frac{\sum_{i \in J_c} \big(w_{ij} - \mathrm{E}[W_{ij}]\big)}{\sqrt{\sum_{i \in J_c} \mathrm{Var}[W_{ij}]}},$$

where J_c is the index set of pixels prelabeled with the cth class. $\mathrm{E}[W_{ij}]$ and $\mathrm{Var}[W_{ij}]$ are unknown, but can be reasonably approximated using the sample mean

$$\overline{w}_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$$

and sample standard deviation

$$S_{\overline{w}_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (w_{ij} - \overline{w}_j)^2}.$$

The Wald statistic is then

$$\hat{z} = \frac{\sqrt{n_c}(\overline{w}_{c,j} - \overline{w}_j)}{S_{\overline{w}_j}}, \tag{3}$$

where $n_c = |J_c|$ and

$$\overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}.$$

Since $\hat{z}$ is generated (approximately) by the standard normal distribution, a hypothesis test can be formed where the null hypothesis is that the average cluster weights corresponding to the cth class *are not* significantly different from the average cluster weights corresponding to all classes, and the alternate hypothesis is that the average cluster weights corresponding to the cth class *are* significantly different from the average cluster weights corresponding to all classes. Again, since class memberships are known a priori and all pixels have some positive membership with all clusters, testing for class memberships is not meaningful, but testing for significantly different cluster weights is meaningful. If $P(Z > \hat{z}) < \alpha$, the probability of observing the difference in the average cluster weights associated with c and the average cluster weights associated with all classes in the jth cluster is significant, and the null hypothesis is rejected. If the null hypothesis is *not* rejected, the cluster itself is rejected as impure, and further refinement is necessary.

One potential issue with the above test is that the sample mean and standard deviation calculations assume the sample is identically distributed, which is specifically *not* the assumption in this case. A better hypothesis test acknowledges that the data are not identically distributed, but are generated from a finite number of distributions. Since the number of distributions and the distributions are unknown, the number of classes and the individual class labels, which are assumed to correspond to inherent structure of the data, are used to approximate the true mean and variance of multiple clusters. Precisely, assume that all labeled pixel indices i with distribution index $\psi(i) = q$ correspond to the same class label $\phi(i) = c$. If $i \in \psi^{-1}(q)$, then $i \in \phi^{-1}(c)$, but $i \in \phi^{-1}(c)$ does not imply $i \in \psi^{-1}(q)$ (more than one distribution can map to one class), and $J_c = \phi^{-1}(c) = \{i \mid \phi(i) = c, 1 \le i \le n\}$. The above hypothesis test requires modification to use class information. In the previous test,

$$\sum_{i \in J_c} w_{ij} = \sum_{i=1}^{n} w_{ij} \delta_{\phi(i),c},$$

$$\hat{z} = \frac{\sum_{i=1}^{n} \big(w_{ij}\delta_{\phi(i),c} - \mathrm{E}[W_{ij}\delta_{\phi(i),c}]\big)}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}[W_{ij}\delta_{\phi(i),c}]}},$$

$$\begin{aligned}
&\sum_{i=1}^{n} (w_{ij}\delta_{\phi(i),c} - \mathrm{E}[W_{ij}\delta_{\phi(i),c}]) \\
&= \sum_{i=1}^{n} (w_{ij}\delta_{\phi(i),c} - a_{ij}\delta_{\phi(i),c}) \\
&= \sum_{i=1}^{n} (w_{ij}\delta_{\phi(i),c} - \alpha_{qj}\delta_{\phi(i),c}),
\end{aligned}$$

recalling that $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ for $i \in I_q$. Assume when $\phi(i) = c$, and distribution index $q = \psi(i)$ corresponds to $c = \phi(i)$, then α_{qj} can be approximated by γ_{cj}, the mean of class $c = \phi(i)$. Ideally α_{qj} should be approximated directly, but there is no way to know $\psi^{-1}(q)$, so essentially $\psi^{-1}(q) \subset \phi^{-1}(c)$ is being approximated by $\phi^{-1}(c)$. Unfortunately, using the sample mean of the cth class and the jth cluster to approximate γ_{cj} and therefore α_{qj} breaks down because the sample mean of the cth class and the jth cluster is both the random variable on the left side and the approximation of the expected value on the right side of the minus sign. This

is illustrated below. Approximating γ_{cj} (and α_{qj}) with the sample mean for the cth class,

$$\gamma_{cj} \approx \overline{w}_{c,j} = \frac{\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n} \delta_{\phi(k),c}},$$

the numerator of the test statistic $\hat{z}$ becomes

$$\begin{aligned}
&\sum_{i=1}^{n}\big(w_{ij}\delta_{\phi(i),c} - \overline{w}_{c,j}\delta_{\phi(i),c}\big) \\
&= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \frac{\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n} \delta_{\phi(k),c}} \sum_{i=1}^{n} \delta_{\phi(i),c} \\
&= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c} = 0.
\end{aligned}$$

Thus this test statistic does not work because the value being tested is the same as the estimated mean for the cth class when using the Kronecker delta instead of Bernoulli random variables. Recall that $Y_{c,j} = \sum_{i=1}^{n} V_{ic}W_{ij}$, where V_{ic}, $i = 1, \ldots, n$ are known prior to classification/clustering. Consider now the test statistic

$$\hat{z} = \frac{y_{c,j} - \mathrm{E}[Y_{c,j}]}{\sqrt{\mathrm{Var}[Y_{c,j}]}}.$$

Fixing j and c, and recalling that $n_q = |I_q|$, the number of indices i for which $X^{(i)}$ has the qth distribution,

$$\begin{aligned}
\mathrm{E}[Y_{c,j}] &= \mathrm{E}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}[W_{ij}V_{ic}] \\
&= \sum_{i=1}^{n} \mathrm{E}[W_{ij}]\mathrm{E}[V_{ic}] \\
&= \sum_{q=1}^{Q} n_q \alpha_{qj} p_c \\
&= p_c \sum_{q=1}^{Q} n_q \alpha_{qj},
\end{aligned}$$

where p_c is the probability that $V_{ic} = 1$. Assuming all the pixels are independent and recalling that $\mathrm{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$ where $i \in I_q$,

$$\begin{aligned}
\mathrm{Var}[Y_{c,j}] &= \mathrm{Var}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n} \mathrm{Var}[W_{ij}V_{ic}] \\
&= \sum_{i=1}^{n}\big(\mathrm{E}[W_{ij}^2V_{ic}^2] - \mathrm{E}[W_{ij}V_{ic}]^2\big) \\
&= \sum_{i=1}^{n}(p_c\mathrm{E}[W_{ij}^2] - p_c^2 a_{ij}^2) \\
&= \sum_{i=1}^{n}(p_c(b_{ij}^2 + a_{ij}^2) - p_c^2 a_{ij}^2) \\
&= \sum_{q=1}^{Q} n_q\big(p_c(\beta_{qj}^2 + \alpha_{qj}^2) - p_c^2\alpha_{qj}^2\big) \\
&= p_c \sum_{q=1}^{Q} n_q(\beta_{qj}^2 + (1-p_c)\alpha_{qj}^2).
\end{aligned}$$

In the above formula, p_c would be approximated by its maximum likelihood estimate $n_c/n = |J_c|/n$. In order to estimate α_{qj}, assume that the qth distribution corresponds to the cth class, $\psi^{-1}(q) \subset \phi^{-1}(c)$, and

$$\alpha_{qj} \approx \overline{w}_{c,j} = \frac{1}{n_c}\sum_{i\in J_c} w_{ij}, \quad c = 1, \ldots, C,$$

where C is the number of classes. Then

$$\begin{aligned}
\mathrm{E}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} n_q\alpha_{qj} \approx p_c \sum_{d=1}^{C} n_d \cdot \frac{1}{n_d}\sum_{i\in J_d} w_{ij} \\
&= \frac{n_c}{n}\sum_{i=1}^{n} w_{ij} = n_c\overline{w}_j,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} n_q(\beta_{qj}^2 + (1-p_c)\alpha_{qj}^2) \\
&\approx p_c \sum_{d=1}^{C} n_d(S_{\overline{w}_{d,j}}^2 + (1-p_c)\overline{w}_{d,j}^2),
\end{aligned}$$

where

$$S_{\overline{w}_{d,j}}^2 = \frac{1}{n_d - 1}\sum_{i\in J_d}(w_{ij} - \overline{w}_{d,j})^2.$$

Using these expressions for the mean and variance of $Y_{c,j}$, the Wald statistic is

$$\hat{z} = \frac{y_{c,j} - n_c\overline{w}_j}{\sqrt{p_c \sum_{d=1}^{C} n_d\big(S_{\overline{w}_{d,j}}^2 + (1-p_c)\overline{w}_{d,j}^2\big)}}, \tag{4}$$

and the null hypothesis is rejected if $P(Z > \hat{z}) < \alpha$.

6 Conclusions

This paper introduced two possible statistical hypothesis tests for the evaluation of soft clusters for classification. Test (3) may not work for data sampled from multiple distributions (the random variables are assumed to be independent and identically distributed). Test (4) attempts to use class information for the estimation of means and variances since the distributions of the data are unknown. Test (4) can be used to determine if a soft cluster has a statistically significant association with one particular class. Numerical results for the application of these two hypothesis tests to soft cluster evaluation, upon which classification is based, for several large scale remotely sensed images are presented in a companion paper.

References

[1] J.P. Wayman, R.H. Wynne, J.A. Scrivani, and G.A. Reams, "Landsat TM-based forest area estimation using Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, 67(2001), 1155–1166.

[2] R.F. Musy, R.H. Wynne, C.E. Blinn, J.A. Scrivani, and R.E. McRoberts, "Automated Forest Area Estimation via Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, 72(2006), 949–960.

[3] R.D. Phillips, L.T. Watson, and R.H. Wynne, "Hybrid image classification and parameter selection using a shared memory parallel algorithm," *Computers & Geosciences*, 33(2007), 875–897.

[4] J. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. Thesis, Cornell University, Ithaca, NY, 1974.

[5] J.C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2,1(1980), 1–8.

[6] B.V. Gnedenko, *Theory of Probability (Sixth ed.)*, Gordan and Breach Science Publishers, The Netherlands, 1997, 497pp.