# Bayes Clustering Operators for Known Random Labeled Point Processes

Lori A. Dalton*, Marco E. Benalcázar†‡§, Marcel Brun§ and Edward R. Dougherty¶‖
*Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA
†Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), Ecuador
‡Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
§Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina
¶Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
‖Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, USA

***Abstract*—There is a widespread belief that clustering is inherently subjective. To quote A. K. Jain, "As a task, clustering is subjective in nature. The same dataset may need to be partitioned differently for different purposes." One is then left with a number of questions: Where do clustering algorithms account for statistical properties of the sampling procedure? How can one address the ability of a clusterer to make inferences without a definition of its predictive capacity? This work develops a probabilistic theory of clustering that fully parallels the well-developed Bayes decision theory for classification, making it possible to address these questions and transform clustering from a subjective activity to an objective operation.**

## I. Introduction

Clustering algorithms attempt to group objects based on some notion of similarity, with the hope of gaining some knowledge about the underlying classes in a problem. But is anything really inferred when applying these algorithms? Consider Fig. 1, where several popular clustering algorithms are applied to a simple mixture of a Gaussian and a ring-shaped distribution. Performance appears very poor. This is confirmed in the average error rates provided in Table I (we will subsequently define error), where we see that the best performance is achieved by fuzzy C-means with an error of about 14%. These classical clustering algorithms perform poorly here because they either make incorrect (implicit) assumptions about the geometry of clusters (usually that they are spherical) or they are too sensitive to outliers.

It has been asserted that clustering is inherently subjective [2]. The definition of clustering is itself not formulated mathematically, but rather explained subjectively. How can we address the ability of a clusterer to make inferences without a definition of its predictive capacity [3]? Duda et al. put the matter clearly, "The answer to whether or not it is possible in principle to learn anything from unlabeled data depends upon the assumptions one is willing to accept – theorems cannot be proved without premises" [4].

TABLE I
Clustering error (%) of algorithms under the mixture of a Gaussian and a ring model [1].

| Algorithm | Error |
|---|---|
| K-means | 18.9 |
| fuzzy C-means | 13.97 |
| hierarch. complete | 26.59 |
| hierarch., single | 46.14 |
| hierarch., average | 40.75 |
| random | 46.04 |

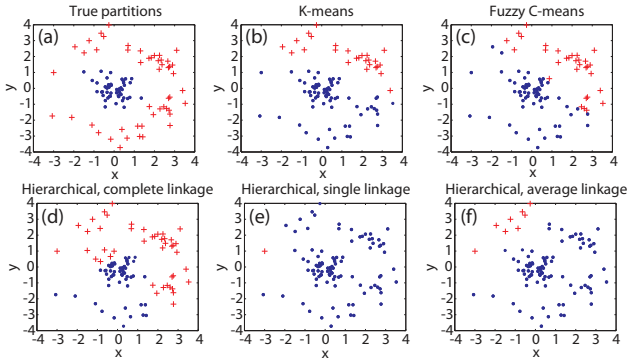


Fig. 1. Clustering point sets of size 50 from a mixture of a Gaussian and a ring [1]. (a) True labels, (b) K-means (24 errors), (c) fuzzy C-means (21 errors), (d) hierarchical with Euclidean distance and complete linkage (18 errors), (e) hierarchical with Euclidean distance and single linkage (49 errors), (f) hierarchical with Euclidean distance and average linkage (42 errors).

This is all in contrast to the probabilistic theory underlying classification, where given the distributions from which test points are drawn it is well known how to find the optimal classifier, and the corresponding minimal error. Even in a practical setting where the distributions are unknown and training data are used, there is always a backdrop of objective theory that serves in understanding fundamental limits of performance, the tradeoff between cost of constraint and cost of design when learning, consistency guarantees, and the accuracy of error estimation.

To date, the only part of this edifice that has been established for clustering is a probabilistic framework based on random set theory, including the definition of the error rate, the key realization being that, whereas a classifier operates on a random variable associated with a feature-label distribution, a clusterer operates on a random point set associated with a random labeled point process (RLPP) [5], making clustering theory inherently more difficult than classification. The development

in [5] does not go beyond the basic probabilistic model and operational definitions, and does not provide a representation of the optimal, or *Bayes cluster operator*. Moreover, the modeling problem was left open, with only trivial models having been considered.

## II. Bayes Clustering

In clustering, the objective is to find the best partition of a point set. It was understood in the original paper [5] that the Bayes clusterer and Bayes error may be found pointwise, that is, they are defined for every possible input point set independently of other point sets. However, it was not realized that the partition error at point set $S$ for an arbitrary clustering algorithm $\zeta$ assigning partition $\zeta(S) = \mathcal{P}_S$ can be written in the form

$$\varepsilon[S, \mathcal{P}_S] = \sum_{\mathcal{Q}_S \in \mathbb{P}_S} c_{\mathcal{P}_S, \mathcal{Q}_S} P(\mathcal{Q}_S|S), \tag{1}$$

where $\mathbb{P}_S$ is the set of all partitions of $S$ (which we call *reference partitions*), $c_{\mathcal{P}_S,\mathcal{Q}_S}$ is a fixed deterministic partition cost for choosing the *candidate partition* $\mathcal{P}_S$ when the true partition is $\mathcal{Q}_S$, and $P(\mathcal{Q}_S|S)$ is the probability that partition $\mathcal{Q}_S$ is the true partition of $S$. The partition cost in clustering plays a role analogous to cost in Bayes decision theory for classification, and the partition error is analogous to the risk (error) of a given classifier label assignment at a fixed point.

In [5], the definition of error provided was equivalent to (1) with a partition cost function

$$c_{\mathcal{P}_S,\mathcal{Q}_S} = \frac{1}{n} \min_{\phi_S \in \mathcal{G}_{\mathcal{Q}_S}} \sum_{\mathbf{x} \in S} I_{\phi_S^*(\mathbf{x}) \neq \phi_S(\mathbf{x})}, \tag{2}$$

where $n$ is the number of points in $S$, $\phi_S^*$ is any fixed label function inducing $\mathcal{P}_S$, and $\mathcal{G}_{\mathcal{Q}_S}$ is the set of all label functions inducing the partition $\mathcal{Q}_S$. A label function is a mapping from points in $S$ to the set of all labels, $L = \{1, 2, \ldots, l\}$, and a label function, $\phi_S$, induces a partition $\mathcal{P}_S$ if $\mathcal{P}_S = \{S_1, S_2, \ldots, S_l\}$ where $S_i = \{\mathbf{x} \in S | \phi_S(\mathbf{x}) = i\}$. This definition of cost finds the proportion of points that are mislabeled in the best possible labeling of $S$ relative to the labeling $\phi_S^*$. It has been shown that this definition of cost is a valid metric in the space of all partitions of $S$.

Having defined the partition error, the Bayes clusterer outputs the partition $\mathcal{R}_S$ that minimizes the error, i.e., $\varepsilon[S, \mathcal{R}_S] \leq \varepsilon[S, \mathcal{P}_S]$ for all $\mathcal{P}_S \in \mathbb{P}_S$. Provided that $P(\mathcal{Q}_S|S)$ is available in a given model and the size of $\mathbb{P}_S$ is feasible, the Bayes clusterer can be found with an exhaustive search. Once the Bayes partition, $\mathcal{R}_S$, has been found, the Bayes partition error can be found using (1). This is completely parallel to classification, where the goal is to find a rule that will assign an optimal label to an observed object, where optimality is relative to the expected error of the rule. The duality goes deeper: Classification involves an feature-label distribution; clustering involves a random labeled point process. Bayes decision theory in classification provides exact Bayes classifiers with minimum expected error relative to known underlying processes. In this work, for the first time we do the same in clustering.

## III. A Gaussian Model

The main theoretical challenge is in determining the partition probabilities, $P(\mathcal{Q}_S|S)$, for a given RLPP, which models probabilistically how points in $S$, and their true labels, are realized. In general, $P(\mathcal{Q}_S|S)$ can be written in the form

$$P(\mathcal{Q}_S|S) = \sum_{\phi_S \in \mathcal{G}_{\mathcal{Q}_S}} P(\phi_S|S), \tag{3}$$

where $P(\phi_S|S)$ is the probability that the labeling $\phi_S$ of points in $S$ is the true labeling. This problem has been solved for several Gaussian models, and here we discuss a class of Gaussian RLPP models with random means and covariances.

Assume that the true clusters in a point set we wish to partition are each drawn from Gaussian distributions in a $d$-dimensional space with independent unknown means, $\mu_i$, and covariances, $\Sigma_i$, where $i \in L$ indexes each of $l$ groups. For each cluster define the following hyperparameters: a real number $\nu_i > 0$, a length $d$ real vector $\mathbf{m}_i$, a real number $\kappa_i > d-1$, and a positive definite matrix $\Psi_i$. Then we assume $\Sigma_i$ is inverse-Wishart distributed,

$$f(\Sigma_i) = \frac{|\Psi_i|^{\frac{\kappa_i}{2}}}{2^{\frac{\kappa_i d}{2}} \Gamma_d(\frac{\kappa_i}{2})} |\Sigma_i|^{-\frac{\kappa_i + d + 1}{2}} \exp\left(-\tfrac{1}{2}\mathrm{tr}(\Psi_i \Sigma_i^{-1})\right),$$

where $\Gamma_d$ is the multivariate gamma function and, given $\Sigma_i$, the $\mu_i$ have Gaussian distributions with mean $\mathbf{m}_i$ and covariance $\frac{1}{\nu_i}\Sigma_i$. With $\Sigma_i$ and $\mu_i$ fixed, each point $\mathbf{x} \in S$, having label $i$, is independent and Gaussian with density $f_i(\mathbf{x}) \sim \mathcal{N}(\mu_i, \Sigma_i)$.

Under this model, it can be shown that

$$P(\phi_S|S) \propto P(n_1, n_2, \ldots, n_l) \times \prod_{i=1}^{l} \frac{\Gamma_d\left(\frac{\kappa_i + n_i}{2}\right)}{|\kappa_i + n_i|^{\frac{d}{2}} |\Psi_i + \Psi_i^*|^{\frac{\kappa_i + n_i}{2}}}, \tag{4}$$

where $n_i$ is the number of points in cluster $S_i = \{\mathbf{x} \in S | \phi_S(\mathbf{x}) = i\}$, $P(n_1, n_2, \ldots, n_l)$ is the prior probability of observing exactly $n_i$ points in cluster $i$ for all $i \in L$,

$$\Psi_i^* = (n_i - 1)\widehat{\Sigma}_i + \frac{n_i \nu_i}{n_i + \nu_i} (\widehat{\mu}_i - \mathbf{m}_i)(\widehat{\mu}_i - \mathbf{m}_i)^T, \tag{5}$$

and $\widehat{\mu}_i$ and $\widehat{\Sigma}_i$ are the usual sample mean and covariance for points that are assigned to cluster $S_i$.

If we additionally assume there are exactly $l$ clusters where all clusters are known to contain the same number of points, $n_1 = n_2 = \cdots = n_l = m$, then for any partition $\mathcal{Q}_S$ with the correct sized clusters, from (4),

$$P(\phi_S|S) \propto \prod_{i=1}^{l} |\Psi_i + \Psi_i^*|^{-\frac{\kappa_i + m}{2}}. \tag{6}$$

The above models assume some information is known about the means and covariances through use of a proper prior density. It is possible to carry out a similar derivation using a RLPP model with improper priors over the means and covariances, resulting in a *generalized Bayes cluster operator*. In particular, set $\nu_i = 0$ for all $i$, so that the $\mu_i$ have improper distributions of the form $f(\mu_i|\Sigma_i) \propto |\Sigma_i|^{-\frac{1}{2}}$, and

set $\kappa_i \geq -d-1$ and $\Psi_i = 0$ for all $i$, resulting in an improper density:

$$f(\Sigma_i) \propto \frac{|\Sigma_i|^{-\frac{\kappa_i+d+1}{2}}}{2^{\frac{\kappa_i d}{2}}\Gamma_d(\frac{\kappa_i}{2})}. \tag{7}$$

As long at $n_i \geq 1$ for all $i \in L$, or equivalently as long as every cluster is known to have at least one point associated with it so that there are exactly $L$ clusters, $P(\phi_S|S)$ is still given by (4) and (6) with $\Psi_i^* = (n_i - 1)\widehat{\Sigma}_i$ for $n_i \geq 2$ and $\Psi_i^* = 0$ for $n_i = 1$. For instance, if all clusters have exactly $m$ points and $\kappa_1 = \kappa_2 = \cdots = \kappa_l = \kappa$,

$$P(\phi_S|S) \propto \left(\prod_{i=1}^{l} |\widehat{\Sigma}_i|\right)^{-\frac{\kappa+m}{2}}. \tag{8}$$

Under these assumptions, a labeling $\phi_S$ has maximum probability if the product of the determinant of the sample covariance of each cluster is minimized. If $n_i = 0$ for any $i \in L$, $P(\phi_S|S)$ cannot be solved, and thus a generalized RLPP model must assign $P(n_1, n_2, \ldots, n_l) = 0$ for any label function that assigns no points to one of the labels.

## IV. Suboptimal Algorithm

As $n$ grows, the problem of finding the Bayes clusterer quickly becomes intractable due to the number of partitions that must be considered. Even overlooking the complexity of computing the sum in (1) for every partition $\mathcal{P}_S \in \mathbb{P}_S$, computing the probability $P(\mathcal{Q}_S|S)$ for all reference partitions can itself become infeasible. Hence, we must consider suboptimal algorithms to approximate the Bayes clusterer, the idea being to constrain the space of candidate and reference partitions to a subset of partitions representing a high concentration of the probability mass over all partitions. This boils down to the problem of identifying partitions that have high probability without evaluating the probability for every partition. The methods discussed herein assume that partitions "near" each other have close probabilities, so that we can search for high probability partitions by looking at only neighborhoods of high probability partitions. This tends to be the case for the Gaussian RLPP model we have introduced. We measure the distance between two partitions, say $\mathcal{P}_S^1$ and $\mathcal{P}_S^2$, by the minimum Hamming distance between labels inducing the partitions or, equivalently, the scaled partition cost, $n \times c_S(\mathcal{P}_S^1, \mathcal{P}_S^2)$, with cost as defined in (2). We loosely refer to this as the Hamming distance between partitions.

Algorithm 1 is a greedy method that searches for the highest probability partition by evaluating the scaled probability of a seed partition (the right hand side of (4) or (6), ignoring the scaling factor), evaluating the scaled probability of all partitions within a closed ball of radius $k$ centered on the seed partition for a given fixed integer $k$ (call these partitions $\mathcal{Q}_S^1, \mathcal{Q}_S^2, \ldots, \mathcal{Q}_S^K$), identifying the partition among these, including the seed, with highest probability, and repeating using the highest probability partition as the new seed until there is no improvement. This procedure is guaranteed to converge to a local maximum in a bounded number of steps,

**Data**: $\mathcal{P}_S^0$ = seed partition
$k$ = Hamming distance
**Result**: $\mathcal{P}_S$ = local maximum probability partition
$i = 0$;
$p_0$ = scaled probability of $\mathcal{P}_S^0$;
**repeat**
  $i = i + 1$;
  $\{\mathcal{Q}_S^1, \mathcal{Q}_S^2, \ldots, \mathcal{Q}_S^K\} = \text{Neighborhood}(\mathcal{P}_S^{i-1}, k)$;
  **for** $j = 1$ *to* $K$ **do**
    $q_j$ = scaled probability of $\mathcal{Q}_S^j$;
  **end**
  $j^* = \arg\max(q_1, q_2, \ldots, q_s)$;
  **if** $q_{j^*} > p_{i-1}$ **then**
    $\mathcal{P}_S^i = \mathcal{Q}_S^{j^*}$;
    $p_i = q_{j^*}$;
  **end**
**until** $q_{j^*} \leq p_{i-1}$;
$\mathcal{P}_S = \mathcal{P}_S^{i-1}$;

**Algorithm 1:** Search for the maximum probability partition. Neighborhood$(\mathcal{P}_S, k)$ finds all unique partitions in a closed ball of radius $k$ centered on a seed partition, $\mathcal{P}_S$.

since the search is over a finite number of partitions and the probabilities can only increase. It is also guaranteed to find the maximum probability partition when $k$ is large enough. The entire procedure may be repeated a number of times with different seeds, and a final partition with highest scaled probability selected.

When the Bayes error is not high, most of the probability mass over partitions tends to be concentrated around a neighborhood of the maximum probability partition. To take advantage of this, we propose a suboptimal clusterer in which the set of candidate and reference partitions is constrained to a closed ball of radius $h$, a fixed integer threshold, about a given seed partition, for instance, the partition output by Algorithm 1. Note $k$ controls the complexity of the maximal probability partition search in Algorithm 1 and $h$ controls the complexity of the Bayes partition search. There are two extremes: If $h = 0$, then the only reference partition is the seed partition, so the suboptimal clusterer trivially produces the seed partition. If $h$ is large enough, then the space of reference partitions includes all partitions, so no matter what seed partition is used we produce the exact Bayes partition. This suboptimal method can be made as accurate as desired by increasing $h$.

Approximating the partition error for a given point set is more difficult than clustering itself. To illustrate, suppose that the partition error is approximated by considering only the constrained set of reference partitions, and we approximate the probability of each reference partition by normalizing the sum of the scaled probabilities to one. If $h = 0$, then only one reference partition is considered, and since the normalized probability of this partition is 1 (it is the only partition), we would report an approximate error of zero for this partition. In

fact, the approximate partition error computed in this fashion is upper bounded by the maximum cost between partitions in the candidate and reference sets, which is upper bounded by $2h/n$ when the candidate and reference sets are both contained in a ball of radius $h$. Put another way, we must choose $h \geq \frac{n}{2}\varepsilon^*$, where $\varepsilon^*$ is the partition error, for it to even be possible to report the Bayes error correctly. In this way, difficult problems with higher Bayes error require higher computational complexity, or higher $h$, for accurate error estimation. The necessary value of $h$ for accurate error estimation is typically much larger than that required for accurate clustering.

## V. Performance

Consider a 2-feature 2-cluster mixture of Gaussians model with unknown means and covariances. Point sets contain 20 points, 10 points per cluster. Set $\mathbf{m}_1 = [0,0]$, $\mathbf{m}_2 = [1,1]$, $\nu_1 = 1$, $\nu_2 = 2$, $\kappa_1 = 2$, $\kappa_2 = 3$, $S_1 = S_2 = 0.5I_2$. These hyperparameters correspond to a proper density over the parameters of the distributions generating points in $S$. An example of a point set realization of this model is shown in Fig. 2, along with the result from clustering using the Bayes clusterer, our suboptimal clustering algorithm (based on Algorithm 1 with constrained reference partitions) and several classical clustering methods.

Average partition errors from 500 realizations of point sets are shown in Fig. 3. Performance graphs are shown for the Bayes clusterer (Optimal), a suboptimal algorithm constraining the set of candidate and reference partitions to a neighborhood of radius $h$ centered on the true maximum probability partition (Subopt. Pmax), a suboptimal algorithm constraining the set of candidate and reference partitions to a neighborhood of radius $h$ centered on a seed generated by running Algorithm 1 five times, and further constrained to only partitions with clusters of the correct size (Subopt. Pseed), and several popular algorithms. The performance of random labeling is also shown, representing worst-case performance. The optimal and suboptimal methods we propose all assume the correct number of clusters, and the correct number of points per cluster. The $x$-axis in Fig. 3 corresponds to $h$, the Hamming distance determining the number candidate/reference partitions considered in our suboptimal algorithms. At $h = 0$, Pmax reports the maximum probability partition, while Pseed reports the output of Algorithm 1. At $h = 10$, Pmax is equivalent to the optimal clusterer, and Pseed is very close (in rare cases the optimal partition may not have the same number of points in each cluster). Note the significant performance gain of the Bayes clusterer relative to the other algorithms. The RMS of the approximate partition error found using Pseed is shown in Fig. 4. For small $h$, the RMS is quite poor, although for large $h$ the RMS converges to zero.

Next we cluster 70 points in 2 dimensions, with 28 point in cluster 1 and 42 points in cluster 2. Set hyperparameters $\mathbf{m}_1 = [0,0]$, $\mathbf{m}_2 = [5,5]$, $\nu_1 = 2$, $\nu_2 = 5$, $\kappa_1 = 2$, $\kappa_2 = 3$, $S_1 = S_2 = 4I_2$. An example of a point set realization of this model is shown in Fig. 5, along with partitions produced by our suboptimal clustering algorithm (Pseed) and several classical clustering methods. For this many points, the Bayes clusterer and maximum probability partition usually cannot be found exactly because the set of all partitions of the point set is very large. For this particular example, a cluster with small variance is embedded in the middle of a cluster with large variance. Pseed very successfully clusters the point sets, while all of the classical methods have a high error rate.

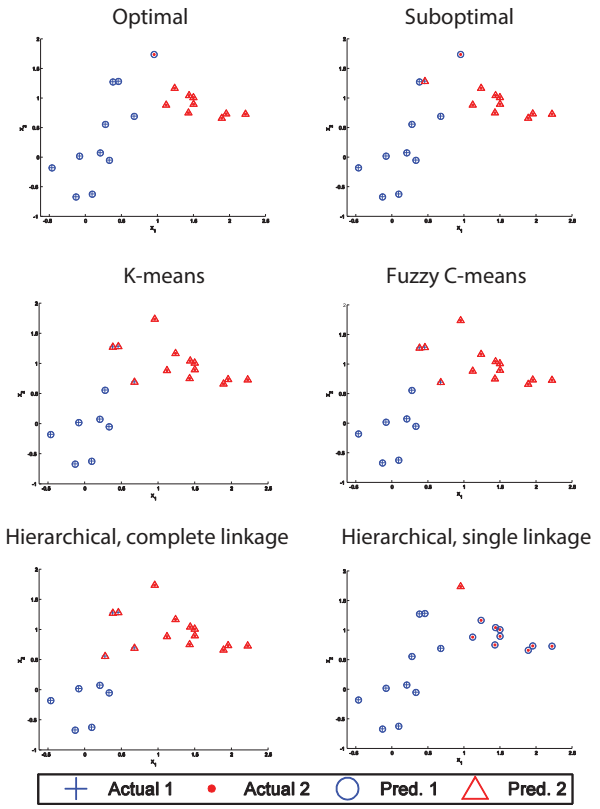


Fig. 2. Examples of clustering mixtures of Gaussians with unknown means and covariances.

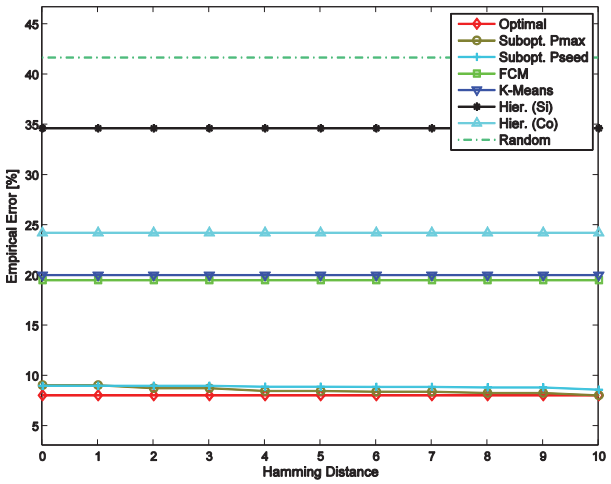


Fig. 3. Average partition error.

Average partition errors from 500 realizations of point sets is shown in Fig. 6 with respect to the Hamming distance, $h$. Although the simulation can only be carried out for small $h$, note the significant performance gain of the suboptimal Bayes clusterer relative to the other algorithms.

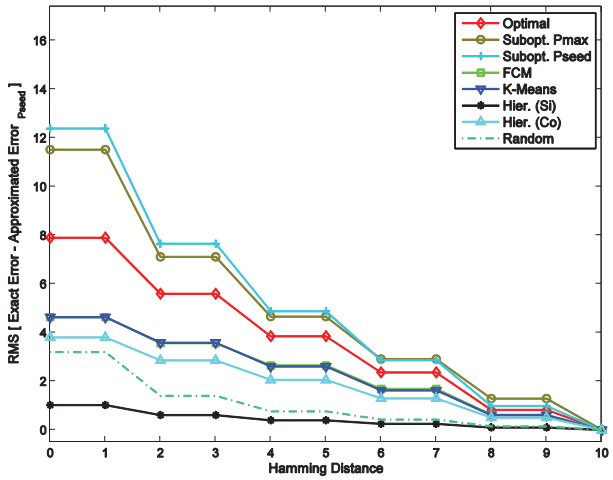


Fig. 4. RMS of approximated error.

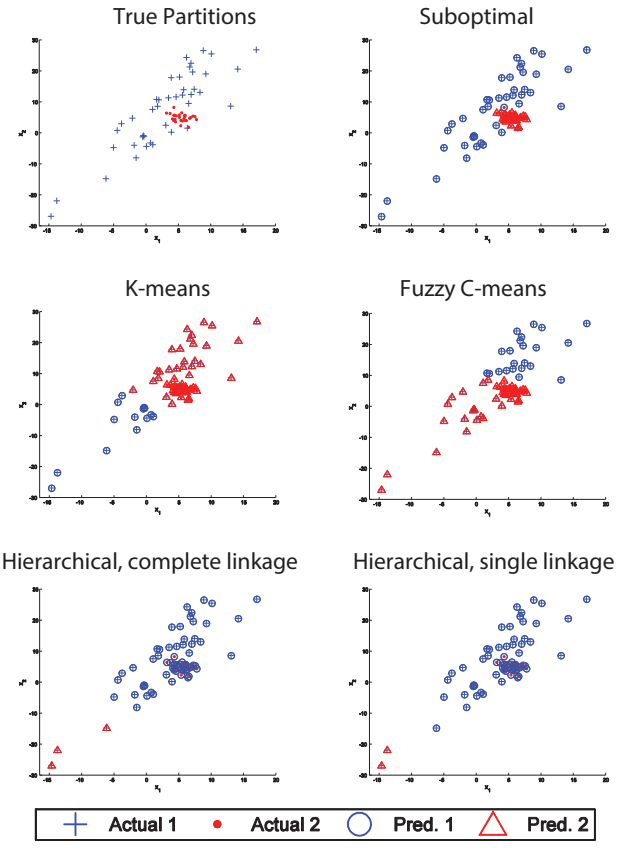


Fig. 5. Examples of clustering mixtures of Gaussians with unknown means and covariances.

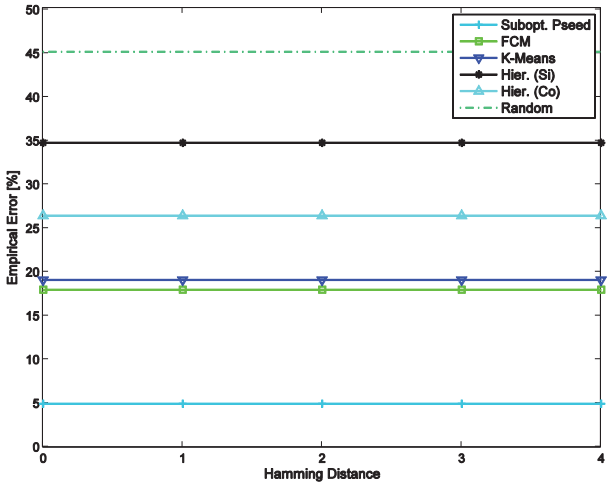


Fig. 6. Average partition error.

## VI. Conclusion

Two basic requirements for developing a rigorous theory of clustering are now in place: An appropriate probabilistic framework including a definition of clustering error, and a Bayes decision theory within that framework. We have shown that Bayes clustering can be formulated analogously to Bayes risk in classification, where a cost function between a predicted partition versus actual partition pair must be specified. The Bayes clustering error can also be objectively quantified for a given RLPP model, turning clustering into an objective operation. One difficulty is that of evaluating the probability (or scaled probability) for any partition. Given a RLPP model, we have shown that this probability can be theoretically formulated in closed form. A second difficulty is the size of the search space, consisting of all partitions of a point set. Though the size of this space is intractable for even moderately sized point sets, this problem can be alleviated with suboptimal search algorithms tailored to a given RLPP model. Many issues remain uncharted in clustering, for instance the practical problem of model uncertainty and learning from examples. The current work lays out tools necessary to begin addressing these fundamental epistemological issues.

## References


[1] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty, "Model-based evaluation of clustering validation measures," *Pattern Recognition*, vol. 40, no. 3, pp. 807–824, 2007.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, September 1999.

[3] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments," *Proc Natl Acad Sci U S A*, vol. 98, no. 16, pp. 8961–8966, 2001.

[4] R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, 2001.

[5] E. R. Dougherty and M. Brun, "A probabilistic theory of clustering," *Pattern Recognition*, vol. 37, no. 5, pp. 917–925, 2004.