Finding a Better k: A psychophysical investigation of clustering

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Abstract

Finding the number of groups in a data set, k, is an important problem in the field of unsupervised machine learning with applications across many scientific domains. The problem is difficult however, because it is ambiguous and hierarchical, and current techniques for finding k often produce unsatisfying results. Humans are adept at navigating ambiguous and hierarchical situations, and this paper measures human performance on the problem of finding k across a wide variety of data sets. We find that humans employ multiple strategies for choosing k, often simultaneously, and the number of possible interpretations of even simple data sets with very few ($N < 20$) samples can be quite high. In addition, two leading machine learning algorithms are compared to the human results and methods for improving these techniques are discussed.

Keywords: machine learning; clustering; psychophysics.

Introduction

Within the field of unsupervised machine learning, clustering is a technique used to separate an arbitrary collection of data points into groups (commonly called clusters). Clustering is usually comprised of two steps. First, one must choose the number of groups, represented by the variable k, for which to look. Second, one must assign each data point to one or more groups while ensuring that there are no empty groups. This second step has received the majority of attention from researchers, with techniques such as the venerable k-means and spectral clustering [1] focusing exclusively on assignment and leaving the task of choosing k up to other algorithms.

There are a few reasons why choosing k is a less attractive problem for researchers as compared to the assignment problem. In some application domains for clustering algorithms, researchers may approach their data with a particular value for k already in mind. In this case, they can simply enter the desired number of clusters into an algorithm like k-means and have a solution without bothering to find which values for k have statistical support. Beyond this practical matter, choosing k has all the hallmarks of a difficult computer science problem. For most data there is no one right answer for what k should be. In fact there may be many answers, some more likely than others. Thus k is an inherently ambiguous quantity, causing much algorithmic difficulty. Some of this ambiguity comes from multiple possible hierarchical interpretations of the data. For the data in Fig. 1, for example, the value of k depends on whether one wants to focus on the details (that there are seven or eight small groups) versus the broad trend (that there are two clear larger groups).

Beyond challenges with ambiguity and hierarchy, there is also the issue of profligacy. Naïvely, one might want to represent the effectiveness of a certain k by calculating an assignment based on that k and then measuring the sum squared distance between each data point and the centroid of the cluster that it is assigned to. This seems like a reasonable strategy, but imagine choosing a k equal to the number of points in one's data. In this situation, each cluster will have exactly one data point, which will be located at the cluster centroid. Obviously the sum squared distance in this case will be zero, indicating a good fit but providing an answer completely useless in terms of data analysis. In general, $k+1$ will always fit data better than k based on this simple measure. A solution to this limitation is to design a more sophisticated statistical test to determine when to stop increasing the number of clusters in order to better fit the data. Unfortunately, statistical tests are based on assumptions about the underlying distribution of the data and if these assumptions are incorrect the test will fail to provide a reasonable result.

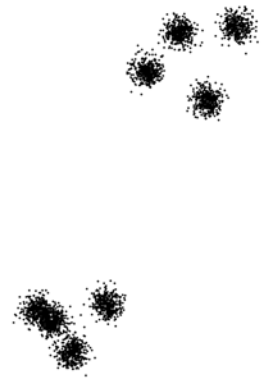

Figure 1: Are there 8, 7 or 2 groups?

The challenges presented in choosing k might lead one to wonder whether humans are accomplished at the task. Humans are adept at navigating ambiguous and hierarchical situations, and we generally cringe at the thought of laboriously counting large numbers of objects, so perhaps we are k-choosing experts. There is a distinct (and often implicit) trend in the clustering literature to use the human visual system as a standard against which the performance of clustering algorithms should be judged. In one prominent spectral clustering paper, the authors state, "The results are surprisingly good... the algorithm reliably finds clusterings consistent with what a human would have chosen [1]." Given that

our visual system is an adept and powerful data processing system (surprisingly resistant to myriad forms of algorithmic mimicry) it is reasonable to solicit its judgments on a thorny problem for which it seems particularly well-suited.

This paper takes inspiration from a computer vision project undertaken at UC Berkeley [2]. Faced with the challenging task of determining how to segment images such that objects are separated from one another by outlines, researchers enlisted human subjects to manually outline the objects in several hundred images of real-world scenes. The problem of image segmentation is very similar to choosing k in that ambiguity and hierarchy play a major role in determining reasonable answers. Detail oriented subjects might outline the leaves on a tree whereas others might just outline the branches. Through this effort researchers collected what is known as the Berkeley Segmentation Dataset, a large collection of human image segmentation data. These data have motivated and assisted several research projects and continue to be a valuable resource in the computer vision field. Studies explicitly measuring human clustering judgements are rare, but at least one study exists that focuses on the developmental changes in human visual grouping of synthetic data sets [3].

This paper presents human judgements on a diverse set of clustering stimuli. The motivation for this undertaking is twofold. First and foremost, we hope to gain intuitions about the methods humans use to choose k and use those intuitions to develop better k-choosing algorithms. The results of this endeavor will be discussed later on. Second, we hope to create a comprehensive and detailed data set representing human clustering behavior that can be used as a standard against which to measure algorithmic performance, and to fuel innovation in this branch of machine learning.

Human Data

Eighteen undergraduate human subjects were recruited for this project, 11 female and 7 male, to determine the number of groups present in 50 distinct point light displays. Each point light display was presented at two different scales and two different rotations, for a total of four presentations per display and 200 trials per subject. Subjects were asked to determine the number of groups in each display and were encouraged to give more than one answer if appropriate. There was no time limit for response. Subjects were told to ignore answers above 20 and to focus on "the bigger picture" to find a reasonable answer less than 20. In addition to k judgments, response times and sequence information were recorded. The sequence of trial presentations was structured into four blocks of 50 randomly ordered trials each, with each block consisting of a unique permutation of every point light display. After the subjects completed all 200 trials, they were interviewed in order to gain insight into their techniques. The interview consisted of two questions:

- What strategies did you use for this task?
- Were any of the displays harder than the others?

While there are likely many interesting phenomena to investigate in the human data, such as consistency, reaction time, the relationship between reaction time and consistency, the relationship between reaction time and k, etc., this paper is mostly concerned with the overall gist of the human responses, their relationship to state-of-the-art k-choosing algorithms, and the new k-choosing methods they inspire. To that end, the human data were analyzed and will be presented collapsed across subjects, scales and rotations. The results are presented in normalized bar plots meant to represent a probability distribution over k, based on the number of responses at each particular k. For each display there are at least 72 responses represented, assuming one answer per subject per trial. The actual number of responses might be larger if subjects were inclined to give multiple answers.

The 50 point light displays used in this experiment were chosen to provide a mixture of depth and breadth within the extremely large space of possible point light displays. Sixteen of the displays consisted of various riffs on mixtures of Gaussians, while another three were mixtures of Gaussians overlaid with uniformly distributed random noise. Nine displays consisted solely of uniformly distributed random noise (with differing number of samples between eight and 10,000). Three displays depicted two-dimensional embeddings of real data. Eight displays contained lines, circles or a combination of the two. The final 11 displays consisted of other synthetic data transformed by a variety of nonlinear distortions. See Fig. 2 for thumbnails of all the displays used. Subjects always saw the displays as white points on a black background, but for the sake of presentation the displays in this paper are black on white and the points have been increased in size.

We focused heavily on mixture of Gaussian data sets due to the prominence of the Gaussianity assumption in the machine learning literature [4][5][6]. We also used several data sets with uniform noise in order to investigate how subject responses varied with sample size and to what extent subjects saw patterns where none were justified by the underlying distribution. Our shape-based and distorted displays were included for breadth and represent a case where the data are drawn from no standard underlying distribution.

Though all of the data sets are two-dimensional, we anticipate that insights gained from this study will lead to algorithmic improvements even in high-dimensional spaces. Certain algorithms (such as the Eigengap algorithm discussed below) operate over affinity matrices that are insensitive to the underlying dimensionality of a data set. Thus, improvements in these algorithms as measured by similarity to human performance in two dimensions will likely scale to high-dimensional data.

Results

Several interesting trends emerge in the human responses. In the interview section of the study, subjects predominantly report two central strategies: looking for areas of greatest density, perhaps separated by empty space ($N = 13$), and count-

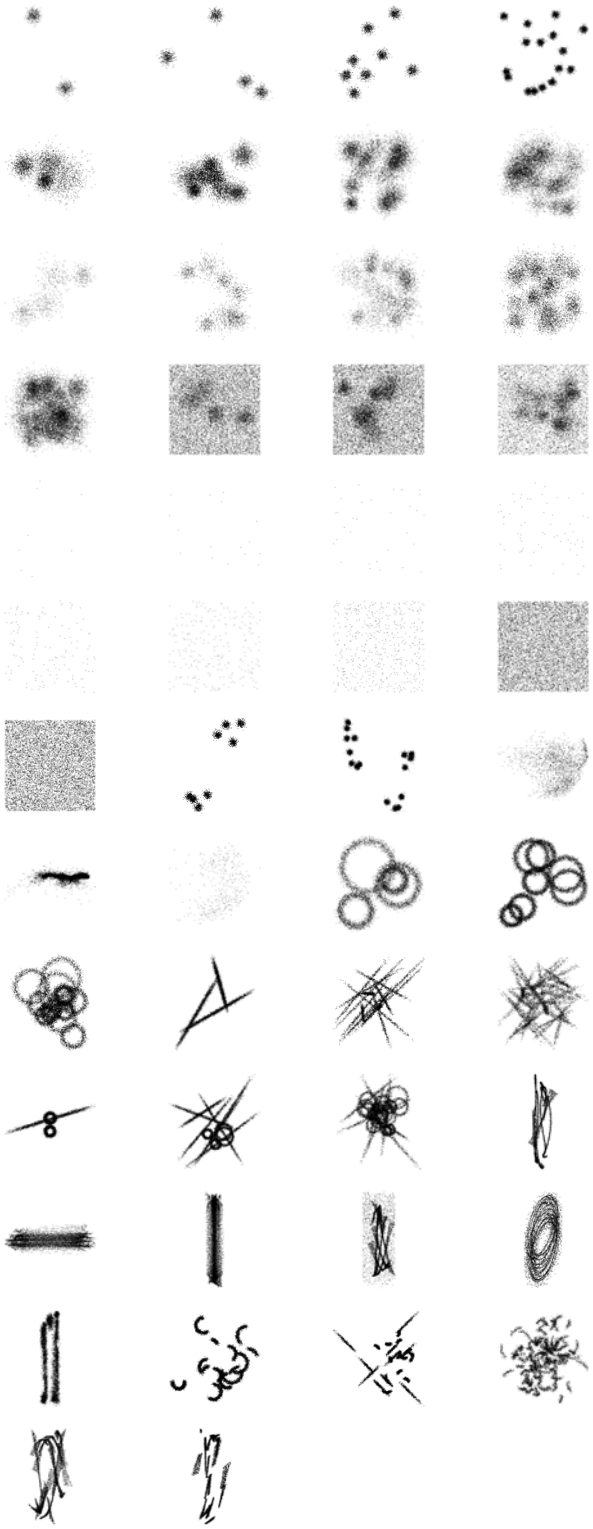

Figure 2: The stimuli.

ing shapes or blobs ($N = 11$). Many of those subjects report using both strategies ($N = 9$). The latter strategy can be interpreted as a model fitting strategy, where subjects see a collection (mixture) of objects (e.g. arcs or Gaussians) and then explicitly count the number of those objects regardless of overlapping density. Rarer strategies include grouping by shape orientation ($N = 1$), and grouping by shape type (if there are both circles and lines in a display, there are two groups, $N = 1$). Finally, one subject explicitly mentions a hierarchical strategy, where he or she searches for small clusters first, and then groups them into larger clusters.

Subjects cite two main sources of difficulty: displays containing very few data points ($N = 9$) and displays with lots of (often overlapping) shapes ($N = 12$). A few subjects consider displays with random noise to be difficult ($N = 3$).

We find echoes of these subjective measures in the choices of k that humans make. Insofar as the distribution over k is less peaked (has higher entropy) for a particular data set, one might interpret that data set as more difficult. Conforming with interview responses indicating that small sample sizes cause difficulty, we can see in Fig. 3 that entropy decreases as sample size increases for displays of uniform noise.

In concordance with interview responses indicating two primary strategies, we find several examples of bimodal responses for displays where these two strategies would diverge. Some examples are shown in Fig. 4.

In all of the mixture of Gaussian cases, humans perform very consistently. In cases where the Gaussians have low variance and well separated means, almost all subjects indicate the correct number of Gaussians. Where the Gaussians have high variance and close means, humans generally agree on a tight range of values for k that corresponds to the number of "blobs" in the display. See Fig 5 for some examples of these results.

Strategies

Based on the observations discussed above, humans follow at least two broad strategies when choosing k, density strategies and model fitting strategies. In this section, two algorithms from recent work in the field that represent these two strategies will be briefly described and their performance compared to the human data.

Density Strategies

Density strategies discover clusters by looking for regions of low density between groups of points, following density within groups to find all the points that belong to them, and attempting to ignore low density noise. Several algorithms have endeavored to formalize these strategies, notably [7].

A more recent algorithm [8], which this paper will refer to as the Eigengap algorithm, brings similar strategies for finding k under the spectral clustering umbrella. The Eigengap algorithm treats each data point as a node on a graph, and then performs a random walk between the nodes, with the probability of transitioning between any two nodes weighted by the distance between them. If two nodes are close together then the probability of transitioning from one to the other will be high and if two nodes are far apart then the probability of transitioning from one to the other will be low. Thus, if a

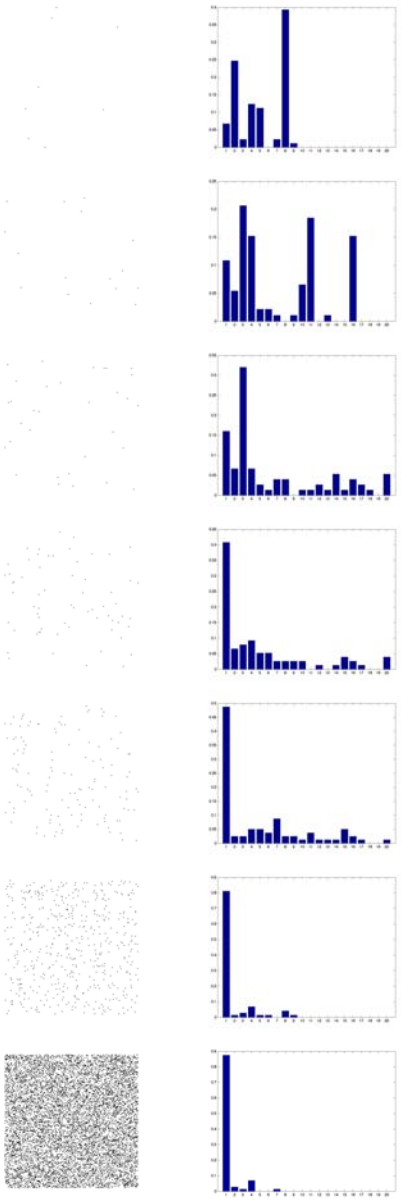

Figure 3: Responses to uniform noise. Displays with few samples present significant ambiguity. As sample size increases, entropy decreases.

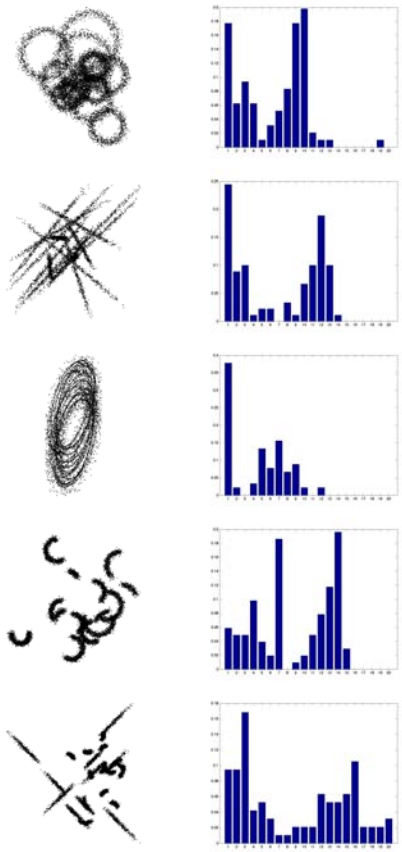

Figure 4: A sample of displays that elicited bimodal responses from subjects.

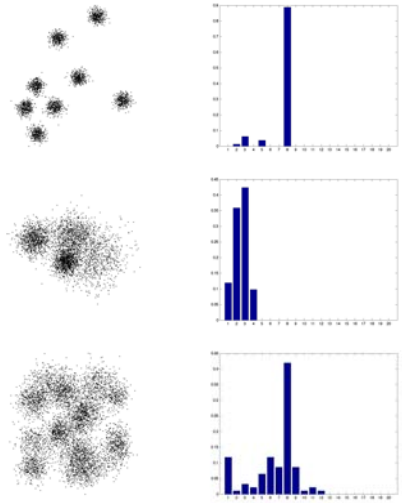

Figure 5: Human responses are generally consistent for mixture of Gaussians data sets.

group of points is separated by a large distance from the rest of the data, a random walk will be unlikely to transition across that gap. In this case, all the points within the group will have a high probability of ending up on other points in the group and little probability of ending up outside the group.

A matrix, P, representing the probability of any point ending up at any other point in the data set will therefore be block diagonal if there are distinct groups within the data set that are separated by sufficient distance. This block diagonal structure is represented by the n largest eigenvalues of P, and eigenvalues greater than the nth will generally be much smaller than the first n eigenvalues. By finding the largest difference between neighboring eigenvalues sorted in descending order, one can find a useful estimate of the number of groups in the data. For example, if the difference between the third and

fourth eigenvalues is 0.4 and that distance is greater than the distance between all other adjacent eigenvalues, then there are likely to be three groups in the data.

As the random walk progresses the Eigengap algorithm naturally finds groups of coarser and coarser structure. Over an increasing number of steps, a random walk will become more and more likely to cross over low density sections of the data set, and thus two groups that initially might be separated will over time merge and lower values for k will be discovered. In this way the Eigengap algorithm can respond well to hierarchical data given a sufficiently long random walk.

The implementation of the Eigengap algorithm in this paper uses a small tweak as compared to [8]. Given a data set with N points, the authors of [8] suggest searching over N possible values for σ, a parameter used in generating the transition probability matrix, between the minimum and maximum pairwise distances in the data set. The algorithm used in this paper searched over 10 possible values for σ in order to drastically reduce computation time while still investigating a reasonable range of values. Also, given the large (over 10,000) number of points in some of the data sets, a sparse implementation of the Eigengap algorithm was used, with pairwise distances only calculated between nearest neighbors (and the number of nearest neighbors equal to one percent of the total number of points in the data set).

Model Fitting Strategies

Several model fitting strategies based on an assumption of mixture of Gaussian distributed data have been proposed in the past [4] [5]. This section describes a recent variant called PG-means [6]. PG-means searches for Gaussian clusters in a data set using an iterative process. The algorithm is initialized with $k = 1$ and it attempts to find an appropriate centroid and covariance matrix for a single Gaussian cluster given the data using the Expectation-Maximization (EM) algorithm. PG-means then randomly projects the data set and the Gaussian model down to one dimension n times (we used $n = 10$). The Kolmogorov-Smirnov (KS) test is applied to each projection and if every KS test indicates a sufficiently good fit (as measured by a parameter α that was set to 0.001) then the current value for k is accepted. Otherwise, k is incremented by one and the entire process is repeated.

If PG-means did not find an answer less than $k = 20$, the algorithm was halted and its response considered to be $k = 1$. Note that unlike the Eigengap algorithm, PG-means will only give one possible value for k.

Comparison with Human Data

To broadly compare Eigengap and PG-means performance with human performance, both the human results and the algorithmic results are interpreted as probability distributions over k. The sum Kullback-Leibler (KL) divergence is then calculated between the human results and both Eigengap and PG-means over all 50 data sets. The human results are considered the true distribution and the algorithmic results are considered the model distributions for purposes of calculating KL.

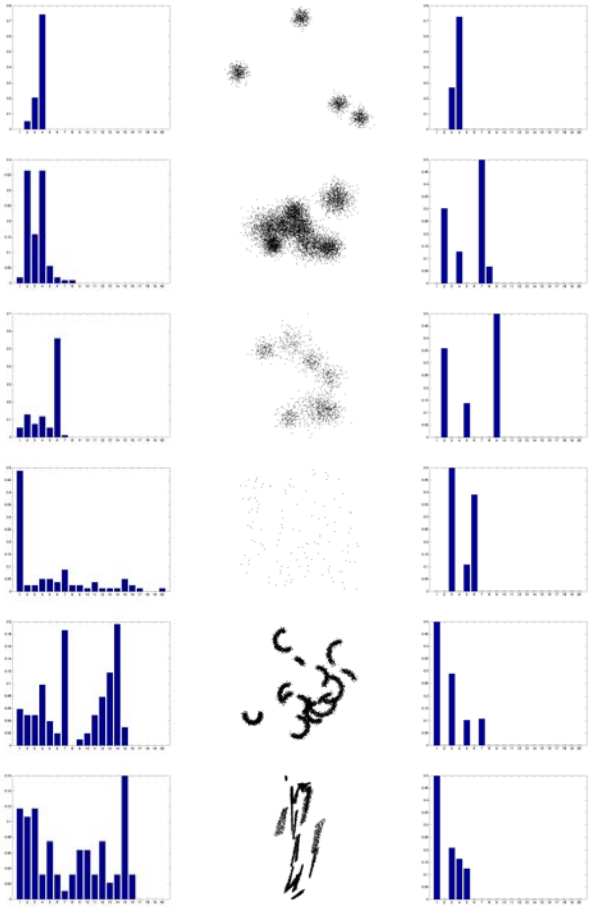

Figure 6: Sample human (left) versus combined Eigengap and PG-means (right) probability distributions over k.

Unsurprisingly, given its ability to return multiple values of k and discover hierarchical organization, Eigengap outperforms PG-means with a sum KL divergence of 269.1 compared to 316.2 for PG-means. A simple unweighted combination of the two, however, performs better than either algorithm on its own with a sum KL divergence of 245.8 (an improvement of 8.7 percent over the Eigengap algorithm). See Fig. 6 for some sample comparisons of this combined result to human responses.

New Density Strategies

While the Eigengap algorithm performs its function of following density well, this paper proposes two novel strategies that use density in other ways. Both of the techniques proposed are based on leveraging higher order density information than traditional pairwise distances or affinities.

First, we are developing an algorithm that intelligently culls uninformative samples from a dataset in order to increase the accuracy and decrease the computational complexity of k-choosing algorithms. These uninformative sam-

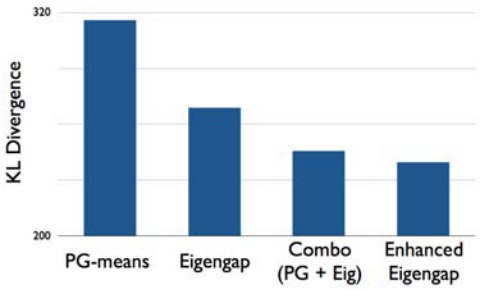

Figure 7: The KL divergence scores for PG-means, Eigengap, their combination, and the enhanced version of Eigengap. (Note that the Y-axis is scaled to make the distinctions more visible.)

ples should correspond to the less dense or "bright" samples that humans tend to ignore. Consider a sample, $\vec{x_i}$, and the set of its κ (kappa) nearest neighbors, ν_i. Define the neighborhood variance of $\vec{x_i}$ as $\sigma_i^2 = \sum_{n \in \nu_i} ||\vec{x_i} - \vec{x_n}||^2 / \kappa$ and define the normalized neighborhood variance of $\vec{x_i}$ as $norm(\sigma_i^2) = \sigma_i^2 / \min_{n \in \nu_i}(\sigma_n^2)$. Remove all samples from the dataset whose $norm(\sigma_i^2)$ measure is above threshold. Both this threshold and κ can be reasonably set based on the number of samples in the dataset and the potential range of k. In a mixture of Gaussians setting, samples with high $norm(\sigma_i^2)$ will be far from the mean of their underlying distribution, and thus less prototypical than points with lower $norm(\sigma_i^2)$.

For an intuition on how this might help a density based approach to choosing k, consider a simple mixture of two Gaussians whose means are well separated but whose samples overlap in some part of the space. An Eigengap-style algorithm will be able to traverse points across both Gaussians quite easily due to this overlap, leaving little indication that there are two clear clusters. By culling all but a small number of points with the lowest normalized neighborhood variance this "bridge" between the two Gaussians is removed.

Second, we propose an extension to spectral clustering based on the observation that human subjects consider samples with large differences in neighborhood variance to be likely drawn from different clusters (though this is not the case when the variance changes smoothly across space). σ_i^2 can be interpreted as the projection of $\vec{x_i}$ into a one-dimensional space, and a new pairwise affinity matrix between samples can be created based on distances in this space. By adding this new affinity matrix as a second view to spectral clustering one might expect to obtain results more similar to human judgments. Early data from applying this technique to the algorithm in [8] are promising and support this expectation. A version of this technique improves Eigengap performance as compared to human performance by approximately 11 percent with a sum KL divergence of 239.7 (compared to 269.1 for the standard Eigengap algorithm). Interestingly, combining this modified version of Eigengap with PG-means nets only a 2.1 percent improvement over the modified Eigengap alone (compared to the 8.7 percent improvement when combining PG-means with standard Eigengap), indicating that sensitivity to density changes might be part of what drives model fitting strategies in humans. See Fig. 7 for a comparison across all algorithms.

Conclusion

Finding reasonable values for k is an important and difficult problem in unsupervised machine learning. As one can see from the samples in Fig. 6, current algorithms do well in certain situations and very poorly in others. By further investigating human performance and attempting to apply the insights garnered from such investigation, substantial progress can be made in developing new algorithms to tackle this thorny problem.

Acknowledgments

The author would like to thank Virginia de Sa, Gedeon Deák and Marta Kutas for valuable feedback on this project. This work was supported by NSF IGERT Grant #DGE-0333451 to GW Cottrell/VR de Sa.

References

[1] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Cambridge, MA, 2002.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[3] J.M. Santos and J. Marques de Sá. Human clustering on bi-dimensional data: An assessment. Technical Report 1, INEB Instituto de Engenharia Biomédica, Porto, Portugal, 2005.

[4] Dan Pelleg and Andrew Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proc. 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, San Francisco, CA, 2000.

[5] Greg Hamerly and Charles Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, volume 17, 2003.

[6] Yu Feng and Greg Hamerly. Pg-means: learning the number of clusters in data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 393–400. MIT Press, Cambridge, MA, 2007.

[7] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[8] A. Azran and Z. Ghahramani. Spectral methods for automatic multiscale data clustering. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:190–197, 17-22 June 2006.