FINDING A BETTER *k*: A psychophysical investigation of clustering

Joshua Lewis
Cognitive Science Department, UC San Diego

Why Measure Visual Grouping?

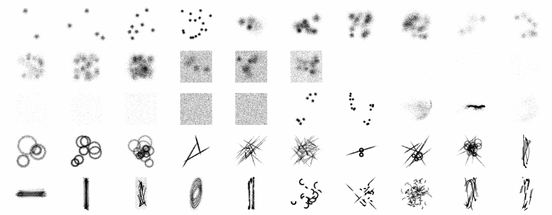
Choosing *k* has all the hallmarks of a difficult computer science problem. For most data there is no one right answer for what *k* should be and thus *k* is an inherently ambiguous quantity. Some of this ambiguity comes from multiple possible hierarchical interpretations of data–does one want to focus on the details or the broad trend? Some comes from decision criteria–is one counting Gaussians or counting disconnected groups of points? To make things even more difficult, *k* + 1 clusters will always fit data better than *k* clusters based on naïve measures.

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, after all. 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 (1), the authors state, "The results are surprisingly good... the algorithm reliably finds clusterings consistent with what a human would have chosen." However, these determinations are usually made based on author intuition, rather than experimental verification.

In this poster I describe research aimed at better understanding the visual grouping decisions humans make. I solicited thousands of grouping judgments on abstract point light displays (see below) from human subjects and compared the results to current methods in machine learning (top right panel). With a better understanding of human behavior, I modified an existing algorithm to fit the human data more closely (bottom center panel).

(1) Ng, Jordan & Weiss. On spectral clustering: analysis and an algorithm. *NIPS 2001.*

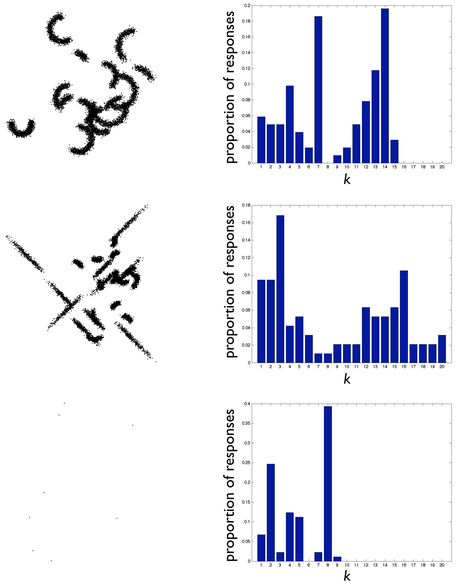
Stimuli



Stimuli were presented white-on-black, but are displayed here black-on-white for clarity.

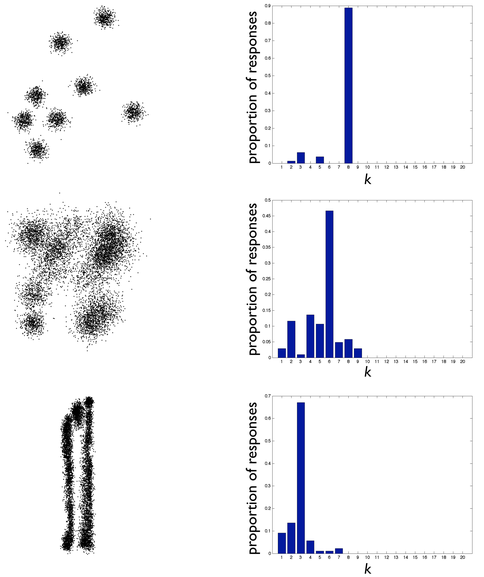
Humans Use At Least Two Distinct Strategies to Choose *k*

Many observed distributions are bimodal



For these displays, first try counting the number of arcs, lines or points, and then try counting unattached groups. Compare those counts to the observed human responses to the right.

In less ambiguous cases there is substantial agreement

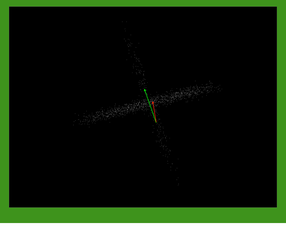


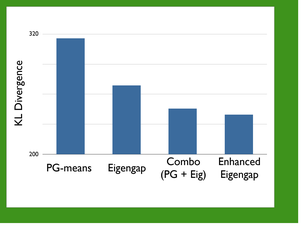
Here, both methods would likely produce similar results, and as we can see, there is much less variability in the observed human responses.

Integrating Additional Density Information into *k*-choosing Algorithms May Increase Performance

A typical setup for spectral clustering involves constructing an affinity matrix between samples based on their Euclidean distance. Humans use more than just distance to find partitions–they also consider changes in density to be a cluster indicator. I modified the Eigengap algorithm (see upper right panel) to use an affinity matrix based on a combination of distance information and local density differences between samples (with local density measured as the mean distance from a sample to its nearest neighbors). In this formulation, the sample at the base of the two arrows in the figure to the right will have a higher affinity with the point indicated by the green arrow due to their similar local densities and a lower affinity with the point indicated by the red arrow, when compared to corresponding values in a typical affinity matrix.

The sum KL Divergence between algorithm output and human responses over all fifty data sets is plotted on the far right. The version of Eigengap that is enhanced with additional density information outperforms all other methods.





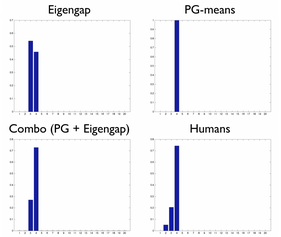
Density & Model Strategies in ML

Based on observed data and post-session interviews, humans follow two broad strategies when choosing *k*, density strategies and model fitting strategies. These strategies have ready analogues in the machine learning literature, and I tested two algorithms (one representing each strategy) to see how well their results agreed with the human data.

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 attempted to formalize these strategies, notably DBSCAN. A more recent algorithm, referred to in this poster as the Eigengap algorithm (1), brings similar strategies for finding k under the spectral clustering umbrella and uses the eigensystem of an affinity matrix between samples to find *k*.

Several model fitting strategies based on finding a mixture of Gaussians that best describes the data have been proposed in the past, such as X-means and G-means. I tested a recent variant called PG-means (2), which uses EM to fit *k* Gaussian centroids to the data and then applies the Kolmogorov-Smirnov test to multiple random one-dimensional projections of both model and data. The algorithm is initialized with *k* = 1, which is iteratively increased by 1 until terminating when a sufficiently likely *k* is found.

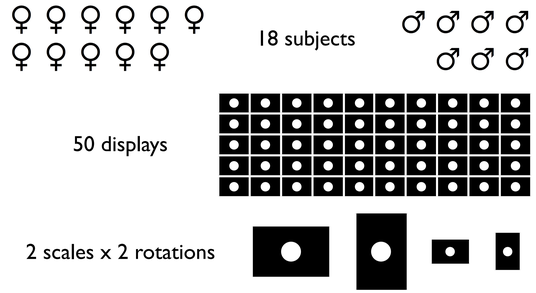
Results from both strategies, an unweighted combination of the two results, and human responses for a simple mixture of Gaussians are shown to the right. Combining both density and model output most closely matches the human results. This result holds over the entire data set as well (see KL Divergence plot in bottom center section).



(1) Azran & Gharamani. Spectral methods for automatic multiscale data clustering. *CVPR 2006.*

(2) Feng & Hamerly. PG-means: Learning the number of clusters in data. *NIPS 2006.*

Method



On each trial, subjects were asked to indicate the number of groups they perceived in the stimulus. Subjects were encouraged to give more than one answer if appropriate.

Special thanks to Virginia de Sa, Gedeon Deák, Marta Kutas, Lawrence Saul and Sanjoy Dasgupta for feedback on this work. This work was supported by an NSF IGERT Fellowship (Vision and Learning in Humans and Machines. PIs: Virginia de Sa and Garrison Cottrell).