

A deterministic annealing approach to clustering AIRS data

Alexandre Guillaume, Amy
Braverman and Alexander Ruzmaikin

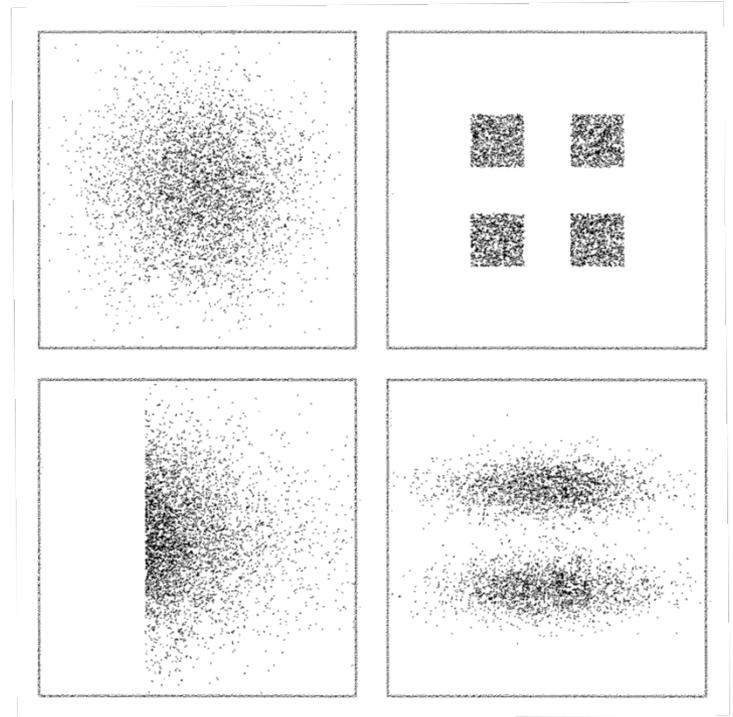
AIRS Science Team Meeting, Tuesday April 24th

Jet Propulsion Laboratory
California Institute of Technology



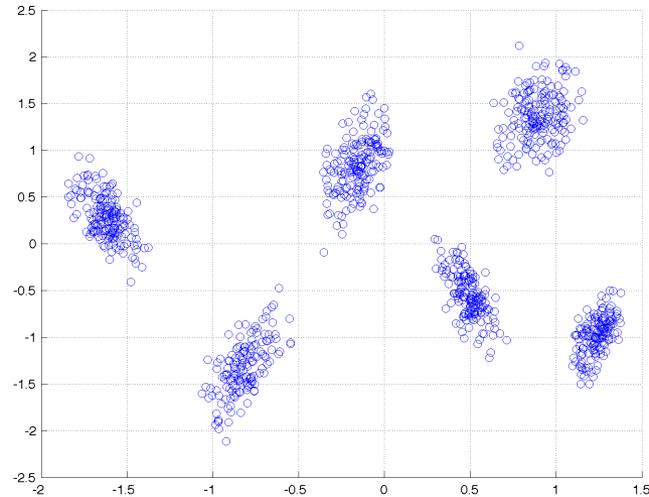
Why?

... we will examine the validity of means and standard deviations as a basis for climate data products. We will explore the conditions under which these two simple statistics are inadequate summaries of the underlying empirical probability distributions by contrasting them with a nonparametric method called Deterministic Annealing technique



These four data sets have identical statistics up to second-order, that is the same mean and covariance.

Clusters



“Clusters may be described as continuous regions of (d-dimensional) space containing a relatively high *density* of points, *separated* from other such regions by regions containing a relatively low *density* of points”

Roughly speaking, clustering procedures yield a data description in terms of clusters or groups of data points that possess strong *internal similarities*.

K-means

Given an initial set of centers, the K-means algorithm alternates the two steps:

- For each center we identify the subset of training points (its cluster) that is *closer* to it than any other center
- The means of each feature for the data points in each cluster are computed, and this mean vector becomes the new center for that cluster

These two steps are iterated until convergence. Typically the initial centers are *randomly* chosen observations from the training data.

Limitations:

- The convergence to a global optimum is not guaranteed
- The results depend on the initial (random) partition
- The user has to provide the desired number of clusters
- It does not work well for data that differ in size, density or are non-globular

Deterministic annealing

Distortion:

$$D = \sum_x \sum_y p(x, y) d(x, y)$$

Entropy:

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y)$$

Lagrangian:

$$F = D - TH$$

$$p(x, y) = p(x) p(y | x)$$

$$T = \alpha T$$

$$\alpha < 1$$



Minimizing F with respect to the association probabilities:

$$p(y | x) = \frac{\exp\left(-\frac{d(x, y)}{T}\right)}{\sum_y \exp\left(-\frac{d(x, y)}{T}\right)}$$

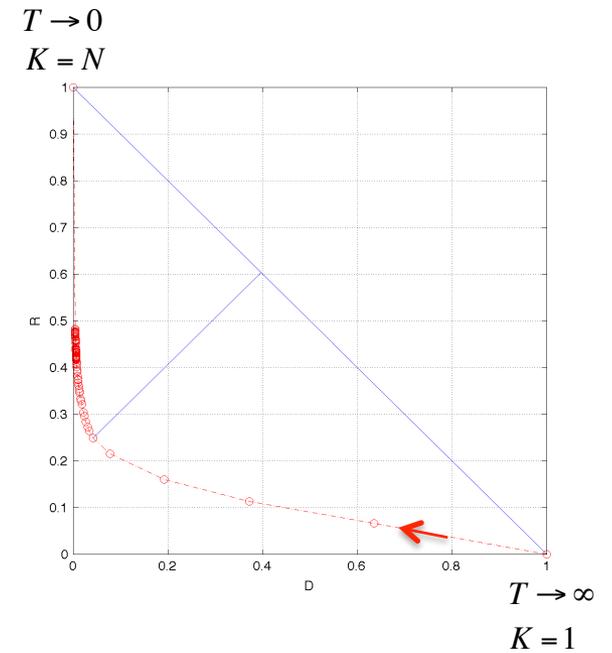
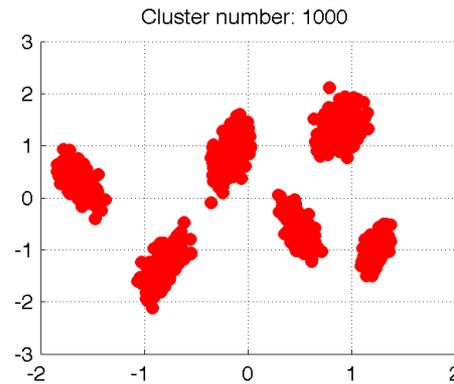
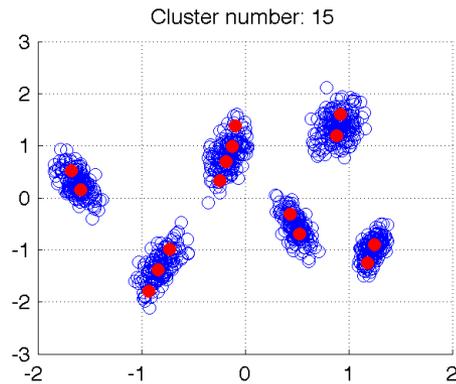
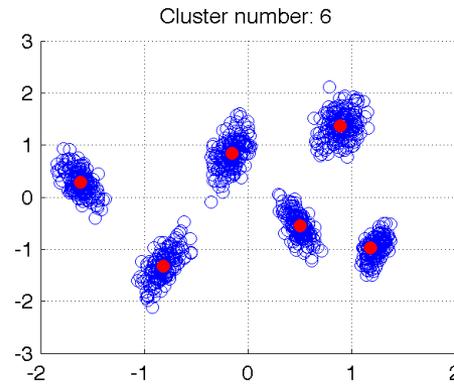
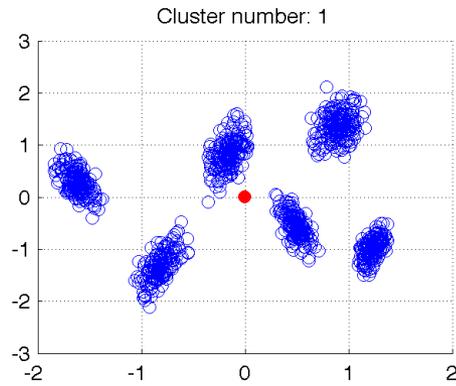
Minimize F^* with respect to cluster locations $\{y\}$:

$$y = \sum_x p(y | x) x$$

This algorithm is deterministic because it minimizes the cost function F directly rather than via stochastic simulation (random sampling).

Synthetic data

DA outputs:
 $\{y\}$, $\{p(y)\}$, $\{p(y|x)\}$, D and R

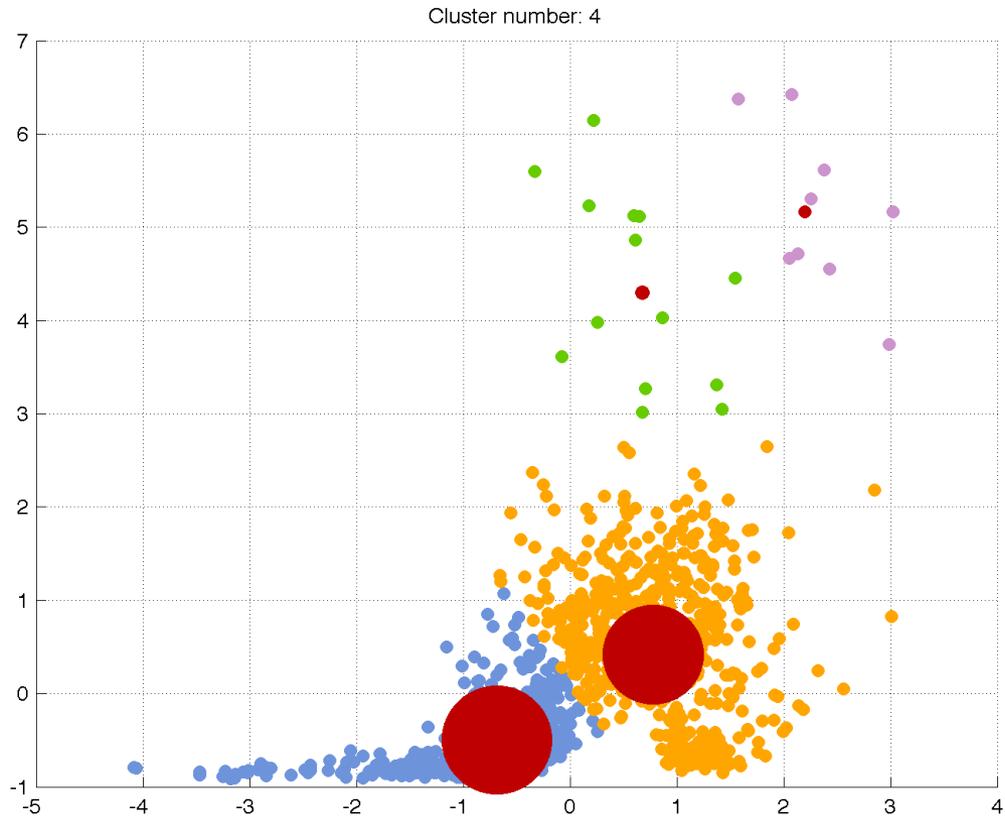


AIRS data

P = 407 hPa



AIRS data



Current clustering applications

AIRS Level 3 quantization products (L3Q) is a standard AIRS data product available from the Goddard DAAC for the entire length of the AIRS mission.

CloudSat: First algorithm performs *clustering* analysis to group individual cloud profile in to a *cluster*, then applies rules and classification methods to classify it into different cloud types. The cloud *clustering* analysis provides cloud horizontal and vertical extent features.

A K-means *clustering* algorithm was used to classify Tropical Rainfall Measuring Mission (TRMM) Precipitation Radar (PR) scenes within 18 square patches over the tropical (158S–158N) oceans. Three *cluster* centroids or “regimes” ... were sought.

...

Conclusion

The deterministic annealing algorithm has advantages (over other algorithms):

- It is deterministic: there are rules to calculate the different quantities. There is no need for random sampling or expensive stochastic trials.
- It depends on only one parameter. Moreover this dependence is very weak (non-existent) for most practical cases (relatively low number of clusters requested). *It's easy for the user.*
- It does not solely depend on the distance criterion



Deterministic annealing can be used for:

- compression while retaining some statistical properties of the data
- feature extraction