(Statistical) Clustering: Spectral Clustering, Mixture Models, Nonparametric Clustering

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 18, 2016

What did we think about last time?

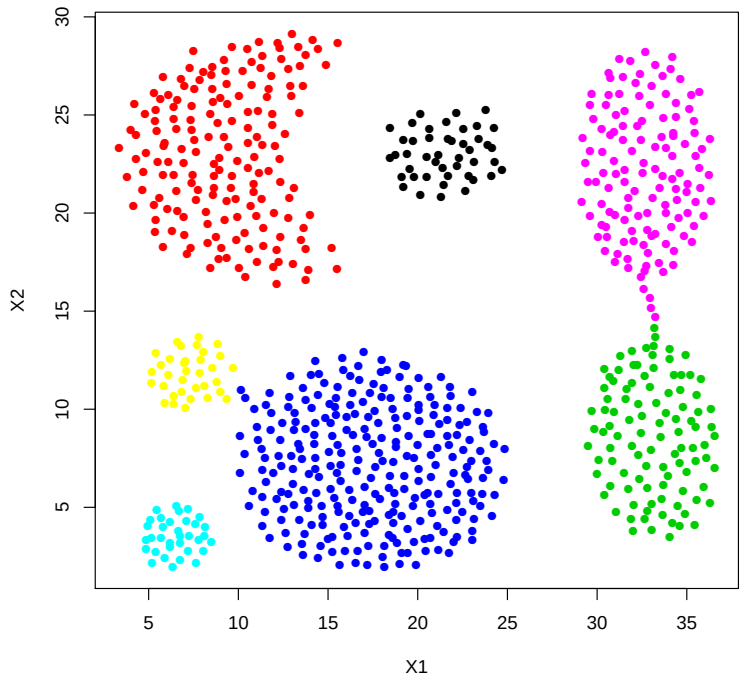
- *K-means:* partitioning obs into spherical clusters; algorithm iterates between choosing optimal assignments and centers
- Can have variation depending on starting centers; use elbow plot (or other consensus criteria) to choose K
- *Spherical K-Means:*
  - Normalize the distances, use cosine dissimilarity
  - Like projecting to surface of sphere, using Euclidean distance
- *Document Clustering*
  - Can turn documents into quantitative variables using TF-IDF
  - Cluster the Document-Term matrix using Sph.K-Means
  - Can select the most important words
  - Also visualize using word clouds

Now we'll

- check out some statistical clustering tools

Spectral Clustering

Interested in using the connected structure of the (dis)similarity matrix to find clusters; how can we "walk" from point to point?



Think about walking in (high-density) neighborhoods

- Where would be easy to walk? More difficult?
- How would our walks change if we change the size of the neighborhood?

Spectral Clustering

Interested in finding clusters based on "connectivity" of data

Example Algorithm

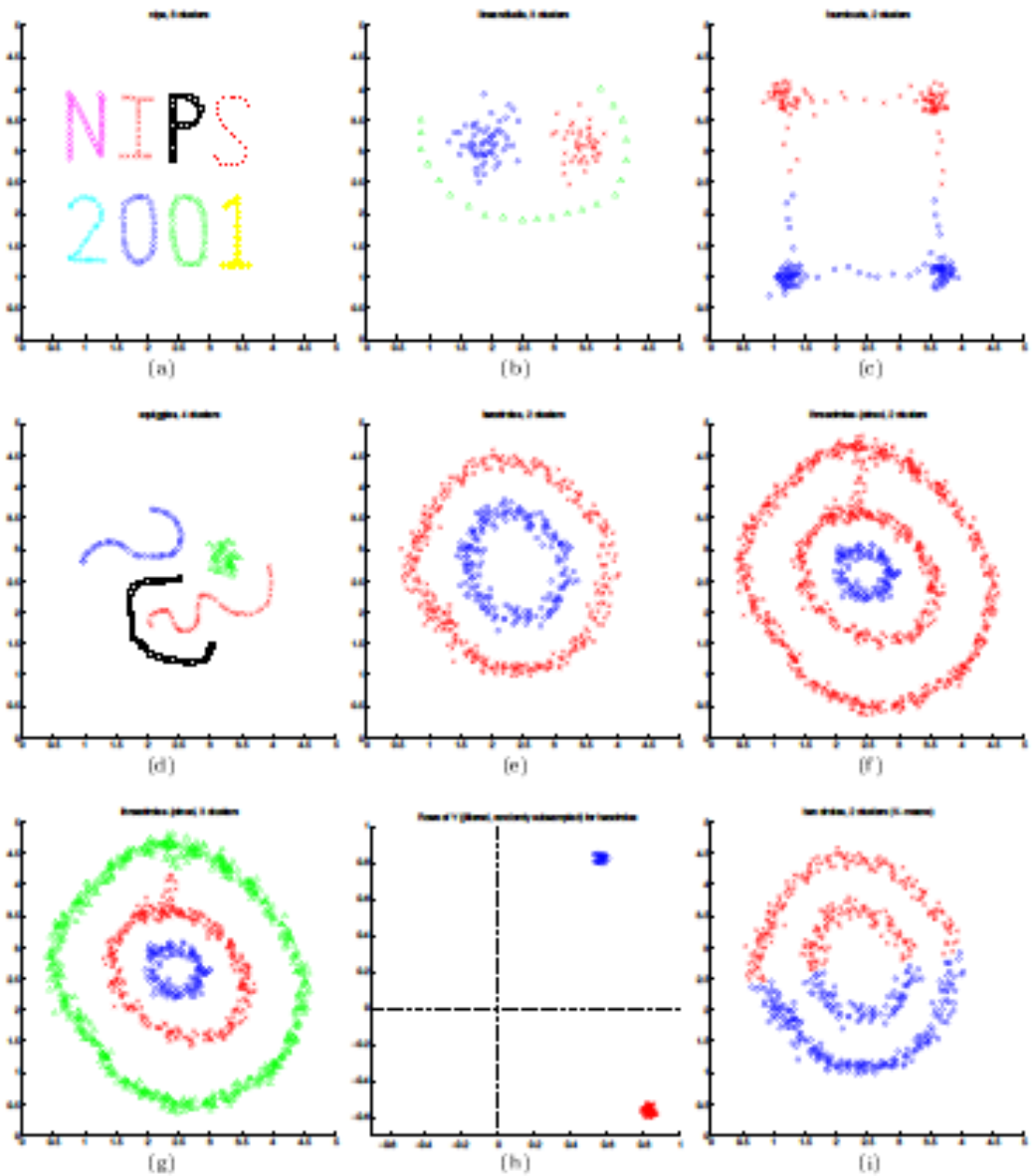
- Calculate the affinity A_{ij} for all pairs of points

$$A_{ij} = e^{(-\|x_i - x_j\|^2 / 2\sigma^2)}$$

- Calculate D, diagonal matrix with elements = row sums of S_{ij}
  Construct $L = D^{-1/2} A D^{-1/2}$
- Find the eigenvector/value decomposition of L;
  select the k largest eigenvectors
- Create new data set Y (eigenvectors in columns);
  normalize the rows to have unit length
- Cluster Y using k-means

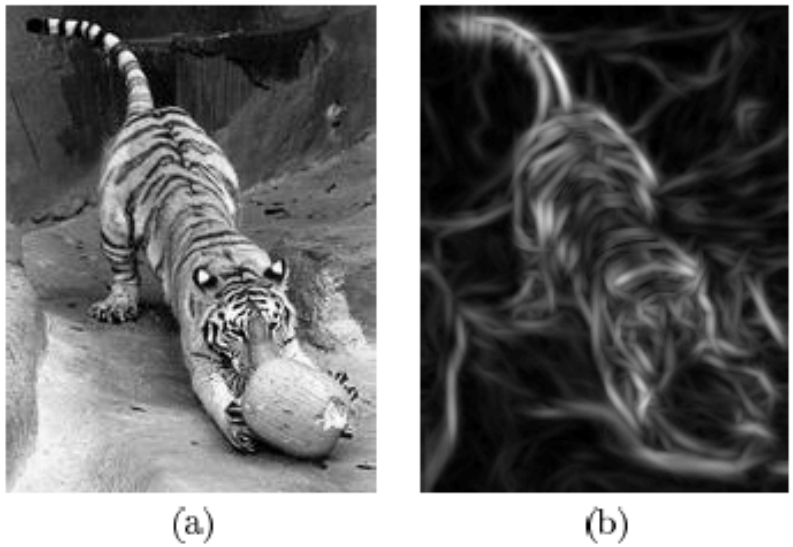
Essentially, we're creating a transition matrix, using the eigenvectors to project observations into a different space, clustering the observations there

Examples (Ng, Jordan, Weiss)



Examples (Meila, Shi)/Variations

- Very commonly used with image segmentation
- Normalized Cut algorithms; Markov Walk algorithms



Model-Based Clustering/Mixture Models

Now adopting the statistical clustering approach:

Assume data are sample from a population with underlying density; estimate density, use its features to determine clustering structure

In model-based clustering, we assume the population is a weighted combination of the true groups

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$$

where $\sum_k \pi_k = 1$ and $0 \leq \pi_k \leq 1$

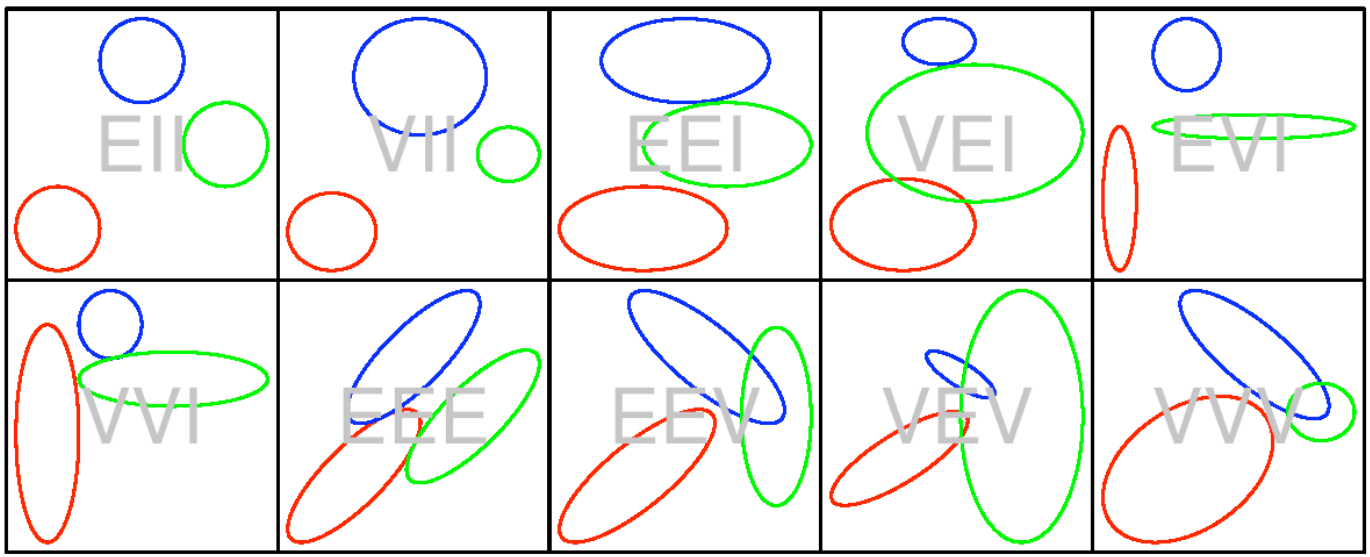
Most common to assume that the components are Gaussian; however all sorts of variations exist:

- skewed normal, t-distributions, beta, inverse hyperbolics, you name it (if it can be estimated)
- add noise component, group for outliers, contaminated or truncated distributions

Estimating the Density

The estimation procedure uses an EM algorithm where it searches for the "best fit" to the data

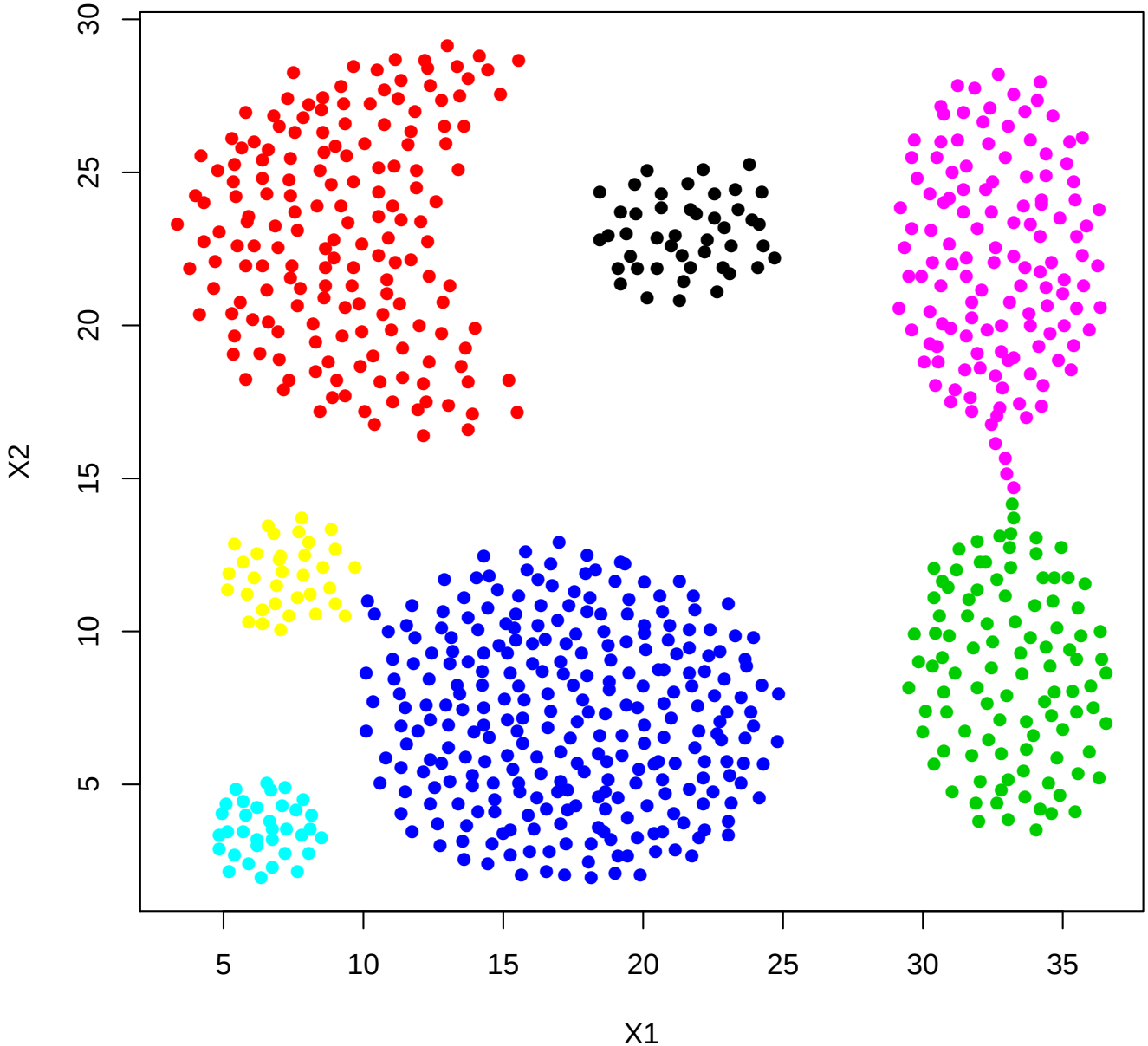
What is a "best fit"?



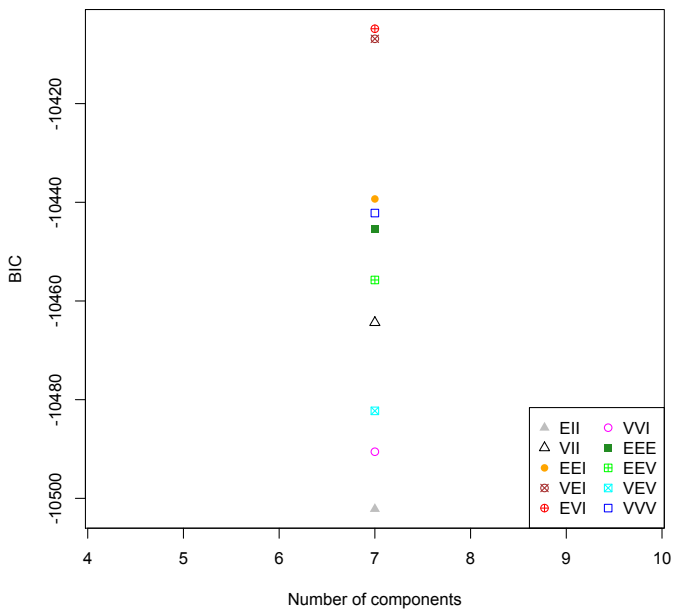
Chooses the best K, the best type of Gaussians (or others) using the Bayesian Information Criterion:

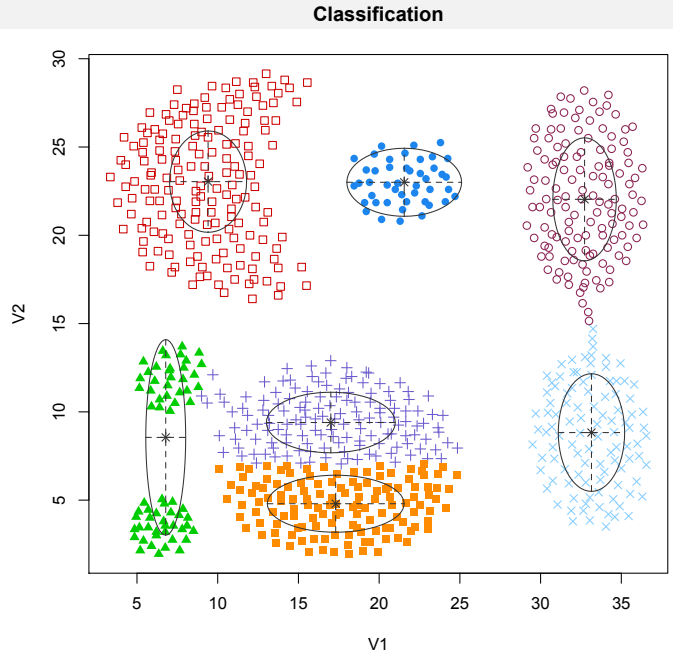
$$BIC = 2 \cdot \log(L(x|\theta)) - \log(n) \cdot p$$

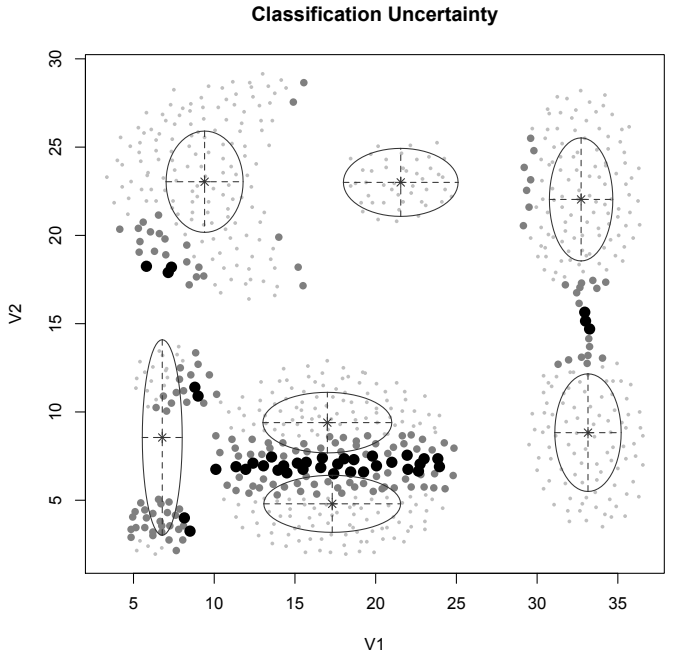
Back to Our Example

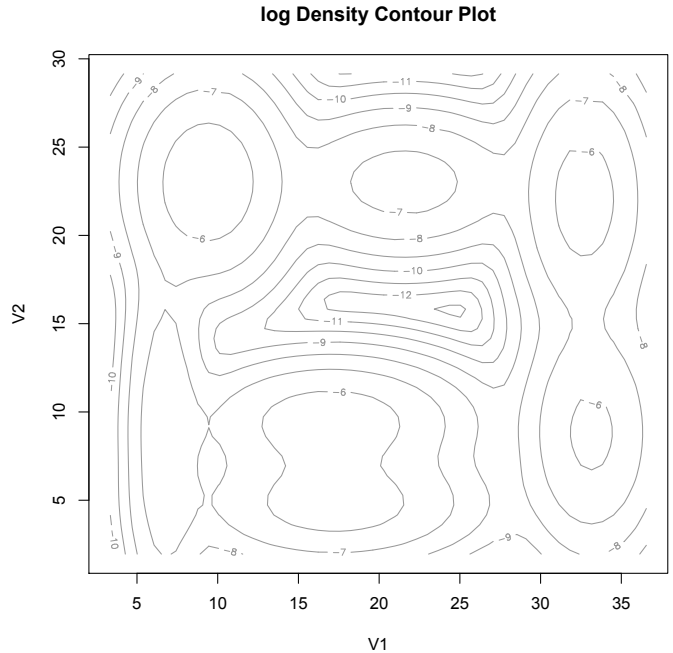


Asked for 7 (Gaussian) Clusters

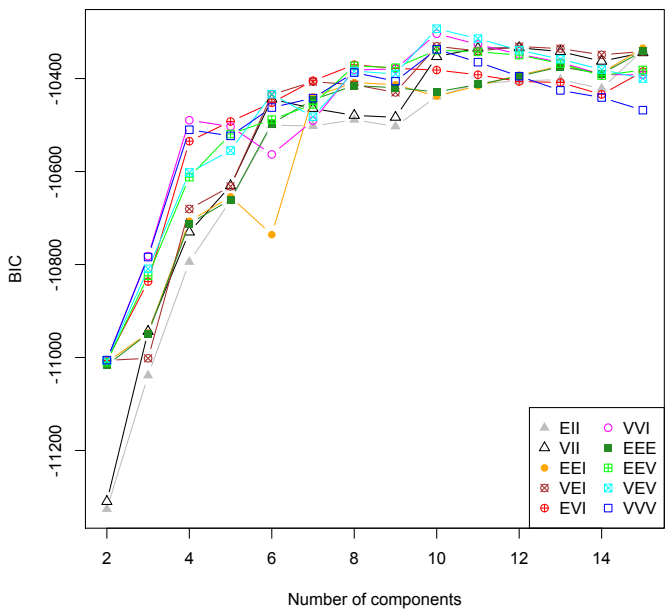


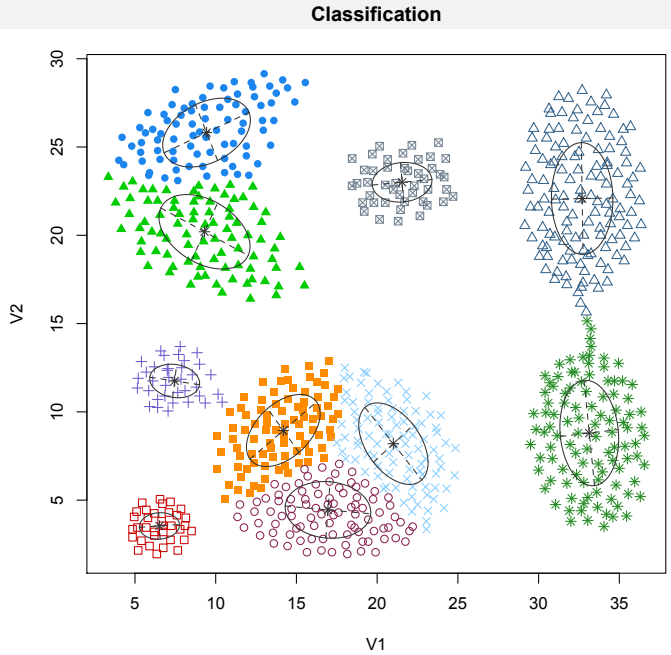


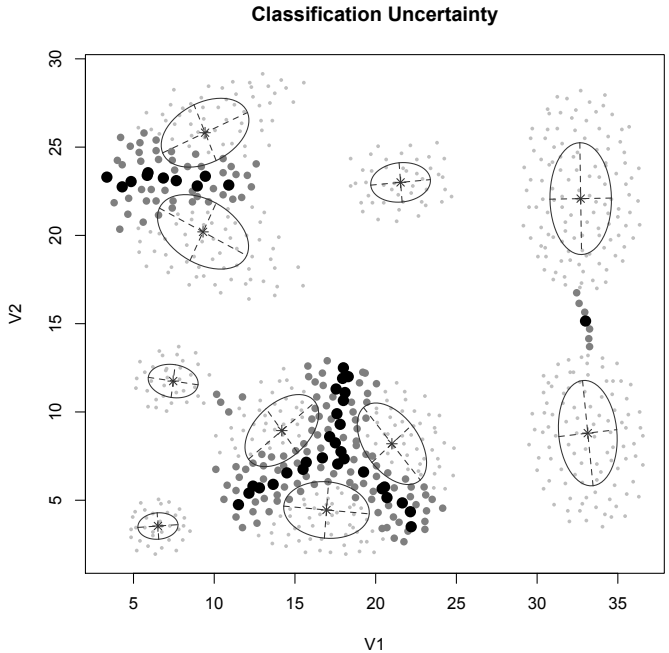


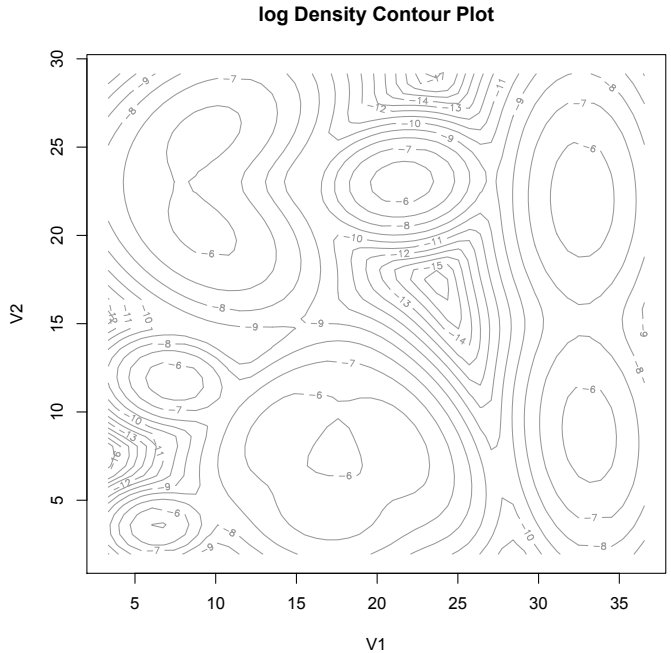


Searched over 2:15 Clusters



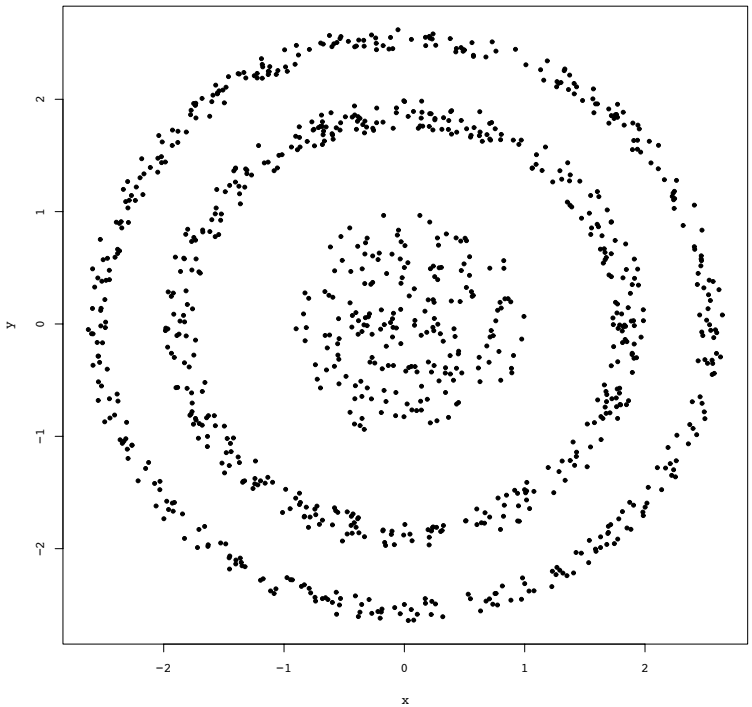


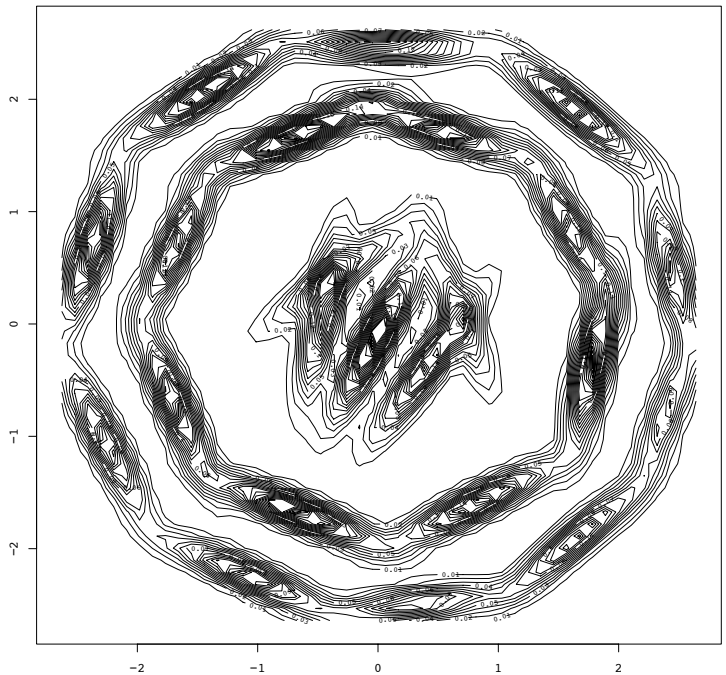




Overfitting MBC

If your assumed density shape is not a good match, MBC often picks too many components; one group = one cluster prob wrong





Scenario where you overfit on purpose and then post-process by merging components that "go together" (entropy, connectivity)