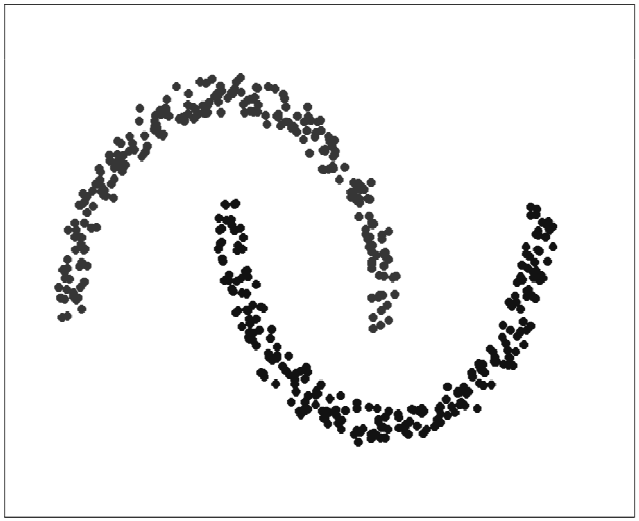
Non-parametric mixtures for data clustering

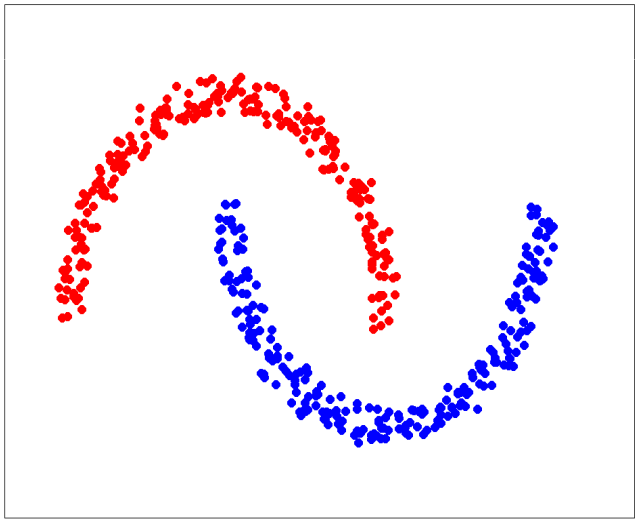
Pavan Mallapragada, Rong Jin and Anil Jain
Michigan State University

Data Clustering

- Given a set of *n* *d*-dimensional points
- Goal is to partition them in to G clusters.



2-dimensional input data



Clustered data

Clustering Methods

Method/Family	Algorithms	Non-parametric	Out-of-sample	Outputs	Specify Number of Clusters
Squared Error	K-means, ISODATA	N	Y	Labels	Y
Parametric Mixture Models	Gaussian Mixture Models	N	Y	Probabilities	Y
Spectral Algorithms	Normalized Cut	Y	N	Labels	Y
Hierarchical Clustering	Single Link	Y	N	Labels	N
Information Theoretic	Information Bottleneck	Y	N	Labels	Y
Density Based	DBSCAN, Meanshift	Y	Some	Labels	N
Proposed NMM		Y	Y	Probabilities	Y

Mixture Models for Clustering

- Mixture models represent the density at a point *x* as a mixture of *G* component densities. Let *g = 1,…,G* denote the index of the component c_g,

$$p(x) = \sum_{g=1}^{G} P(c_g)\, p(x \mid c_g)$$

- Each component density represents one cluster.

E.g., Gaussian Mixture Model (GMM).

$$p(x \mid \Theta) = \sum_{g=1}^{G} \alpha_g N(x \mid \theta_g)$$

$$\Theta = \{\theta_1, \ldots, \theta_G\} \qquad \theta_g = \{\mu_g, \Sigma_g\} \qquad \sum_{g=1}^{G} \alpha_g = 1$$

- Perform well provided data follows the assumed model.

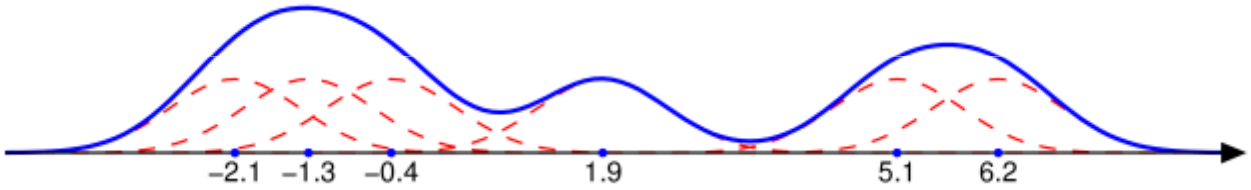
Mixture models for Clustering

- Real data is rarely Gaussian (or any parametric density)
- Each cluster may come from a different and arbitrary probability distribution.
- Clusters may be multimodal

Non parametric density estimation

- Given a set of data points $D = \{x_1, x_2, \ldots, x_n\}$, the probability density at a point x is given by

$$\hat{p}(x) = \frac{1}{n\sigma^d} \sum_{i=1}^{n} k(x, x_i; \sigma)$$



- $k(x, x')$ is a kernel function; σ is a smoothness parameter, called as bandwidth
- A kernel is a probability density in itself.
- A stationary kernel $k(x, x') = \kappa(\| x - x' \|)$

Can model arbitrary distributions.

Non-parametric Mixture models (NMM)

- Clustering as the estimation of a mixture of non-parametric densities.
- Each component density is a kernel density estimate
- Recall that the mixture density is given by

$$p(x) = \sum_{g=1}^{G} P(c_g)\, p(x \mid c_g)$$

Component densities are kernel density estimates.

$$p_g(x|c_g, \mathcal{D}) = \frac{1}{|c_g|} \sum_{x_i \in c_g} \kappa(x, x_i)$$

Goal

- Find the clusters c_g by maximizing the data likelihood

$$p(D) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \sum_{g=1}^{G} P(c_g)\, p(x \mid c_g)$$

- NP Hard Problem!
- An efficient approximate algorithm is developed for finding the clusters.

Weighted density estimate

Kernel density estimate

$$p_g(x|c_g, \mathcal{D}) = \frac{1}{|c_g|} \sum_{x_i \in c_g} \kappa(x, x_i)$$

Relaxation

Weighted Kernel density estimate

$$p_g(x|c_g, \mathcal{D}) = \sum_{i=1}^{n} q_i^g \kappa(x_i, x) \qquad \sum_{j=1}^{n} q_j^g = 1$$

- q^i_g, $i = 1,\ldots,n$; $g = 1,\ldots,G$ are the weights of contribution of point x_i to the density of cluster c_g
- Collect q^i_g into an $n \times G$ matrix Q; Q is called the profile matrix.

A Leave-one-out likelihood maximization scheme is adopted to estimate the profiles.

Leave-one-out parameter estimation

- Define corresponding leave-one-out densities

- Leave-one-out (LOO) Conditional Probability

$$p_i(x_i|c_g, \mathcal{D}_{-i}) = \frac{1}{\sum_{j=1}^{n}(1-\delta_{j,i})q_j^g} \sum_{j=1}^{n}(1-\delta_{j,i})q_j^g K_{i,j}$$

- LOO Unconditional density

$$p_i(x_i|\mathcal{D}_{-i}) = \sum_{g=1}^{G} \gamma_i^g p_i(x_i|c_g, D_{-i})$$

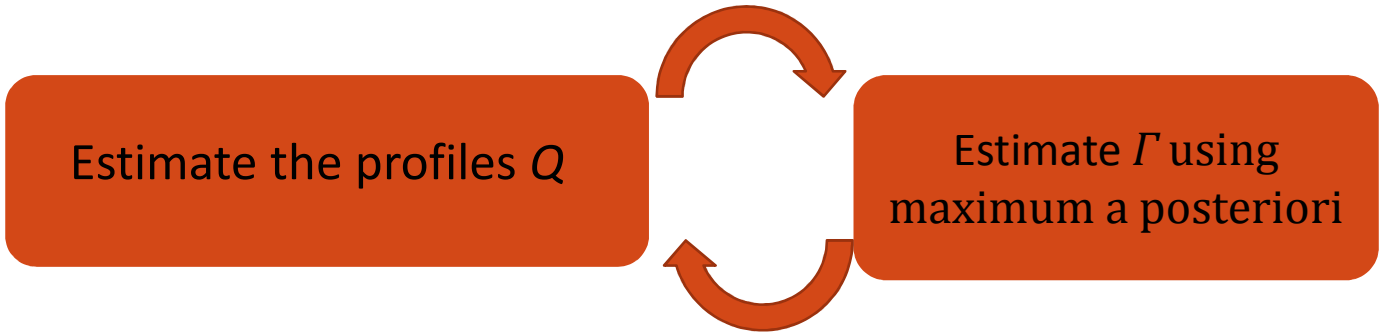
Each leave one out density has its own mixing parameters.
These provide the cluster labels.

Alternating Optimization

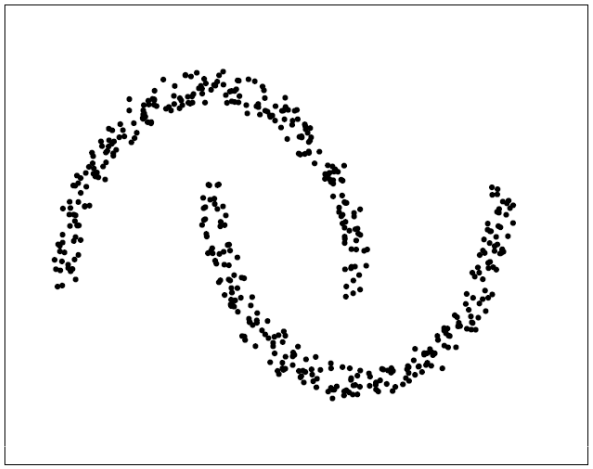
- LOO Likelihood

$$\begin{aligned}\ell_{LOO}(\mathcal{D};Q,\Gamma) &= \log p(Q) + \sum_{i=1}^{n} \log p_i(x_i|\mathcal{D}_{-i}) \\ &= -\lambda \sum_{i=1}^{n}\sum_{g=1}^{G} (q_i^g)^2 + \sum_{i=1}^{n} \log\left(\sum_g \gamma_i^g \frac{\sum_{j=1}^{n} K_{i,j} q_j^g}{1-q_i^g}\right)\end{aligned}$$

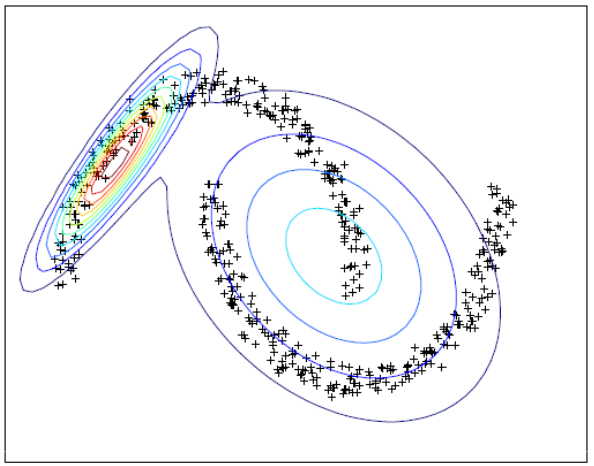
The unknown variables *Q* and *Γ* can be obtained by maximizing the LOO likelihood



Comparison with Mixture Models

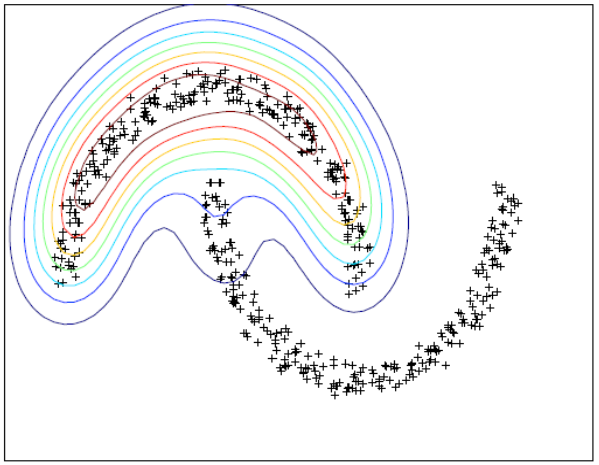


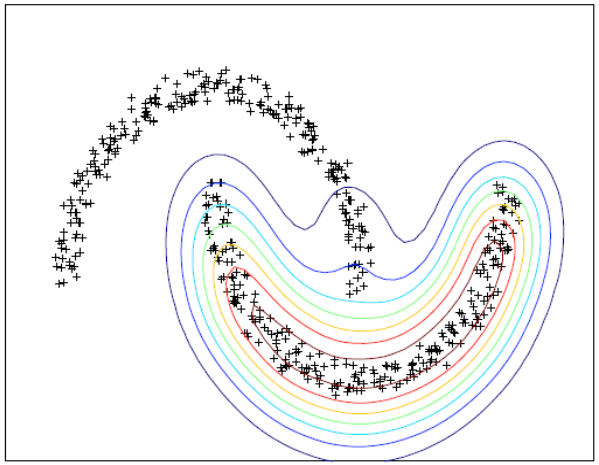
Input 2-cluster data



GMM with 2 components

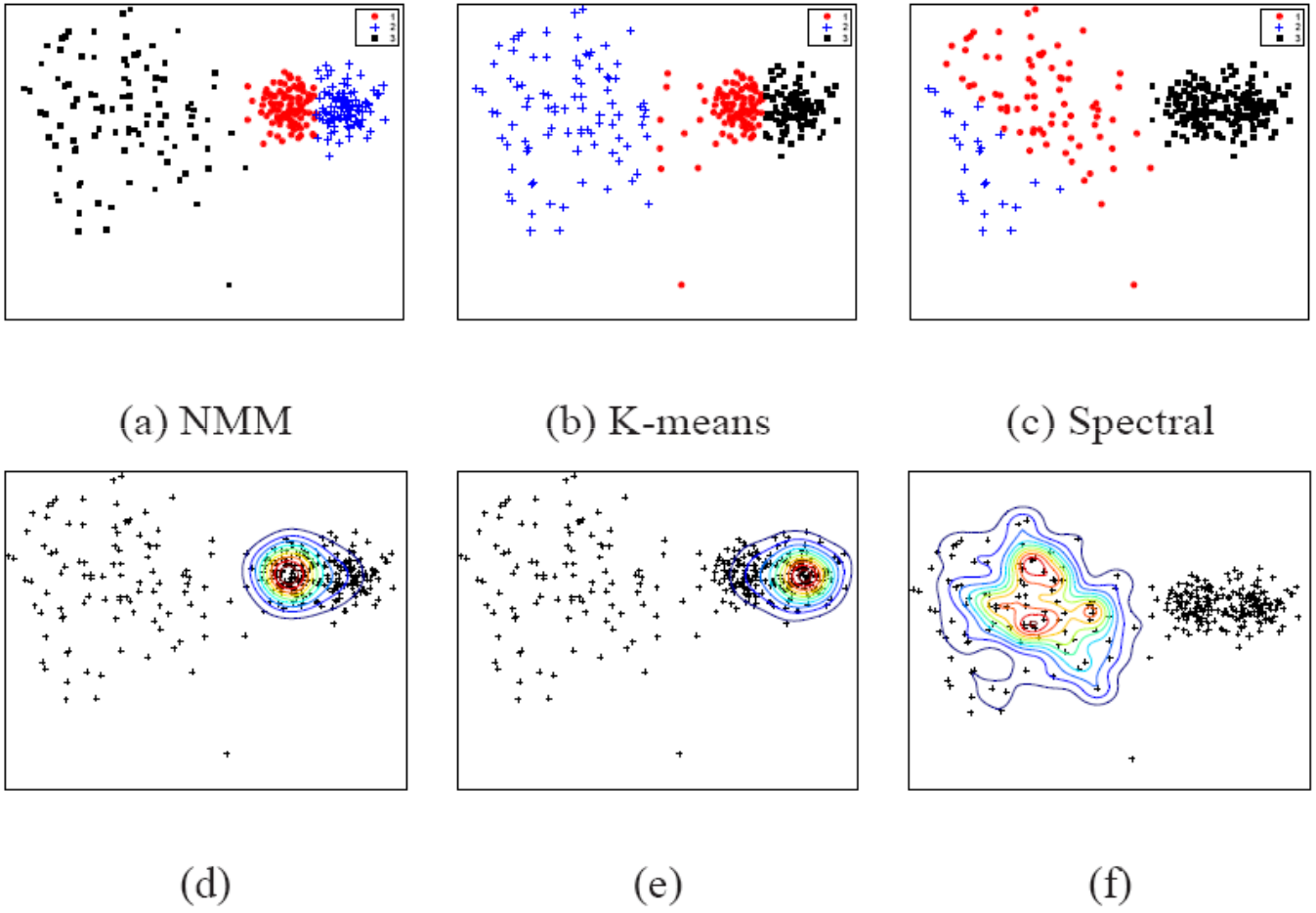
Desired Density for cluster 1





Desired Density for cluster 2

Comparison with Spectral Clustering



Three Gaussian clusters with means (0,0), (6,0) and (8,0) and variances, 4*I, 0.4I and 0.4I* from left to right*; kernel bandwidth is selected as 5th percentile of the pairwise similarities; regularizer weight = 1e-4*

Experiments

- 8 high-dimensional text datasets
- Pairwise F_1 is used to evaluate the clustering performance
- Baselines
  - K-means
  - Spectral clustering
  - Linkage algorithms
- K-means and proposed algorithms are prone to local minima, and are initialized 5 different times each.
- Each algorithm is run 10 times and the mean performance is reported.

Performance on High-dimensional data

Dataset	n	d	G	NMM	K-means	Spectral	Linkage
Different-1000	2975	7657	3	**95.8**	87.8	94.4	40.3
Similar-1000	2789	6665	3	**67.1**	49.9	45.1	37.3
Same-1000	2906	4248	3	**73.8**	49.4	48.0	30.0
Different-100	300	3251	3	**95.3**	79.2	87.5	75.7
Similar-100	288	3225	3	**50.9**	40.1	38.4	43.8
Same-100	295	1864	3	**49.0**	44.8	47.0	41.8
Classic400	400	2897	4	**61.2**	60.1	51.1	53.3
4Newsgroups	3000	500	2	**76.8**	73.8	74.1	68.2

Since the dimensionality of the data is larger than the number of points, Gaussian Mixtures are not even applicable on these datasets.

Conclusions

- A non-parametric mixture model for data clustering is proposed.
  - Probabilistic model that can fit clusters with arbitrary densities.
- The proposed NPM algorithm out performs some of the well known clustering algorithms on document clustering.
- Works well for high-dimensional sparse data
- Future work includes automatic bandwidth selection, scalability and applications to other sparse domains (e.g. bioinformatics).