# Non-parametric Mixture Models for Clustering

Pavan Kumar Mallapragada, Rong Jin and Anil Jain

Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI - 48824

**Abstract.** Mixture models have been widely used for data clustering. However, commonly used mixture models are generally of a parametric form (e.g., mixture of Gaussian distributions or GMM), which significantly limits their capacity in fitting diverse multidimensional data distributions encountered in practice. We propose a non-parametric mixture model (NMM) for data clustering in order to detect clusters generated from arbitrary unknown distributions, using non-parametric kernel density estimates. The proposed model is non-parametric since the generative distribution of each data point depends only on the rest of the data points and the chosen kernel. A leave-one-out likelihood maximization is performed to estimate the parameters of the model. The NMM approach, when applied to cluster high dimensional text datasets significantly outperforms the state-of-the-art and classical approaches such as K-means, Gaussian Mixture Models, spectral clustering and linkage methods.

## 1 Introduction

Data clustering aims to partition a given set of $n$ objects represented either as points in a $d$ dimensional space or as an $n \times n$ similarity matrix. The lack of a universal definition of a cluster, and its task or data dependent nature has resulted in publication of a very large number of clustering algorithms, each with different assumptions about the cluster structure [1]. Broadly, the proposed approaches can be classified into *parametric* vs. *non parametric* approaches. Parametric approaches impose a structure on the data, where as non-parametric methods infer the underlying structure from the data itself.

Probabilistic finite mixture modeling [2,3] is one of the most popular parametric clustering methods. Several probabilistic models like Gaussian Mixture Model (GMM) [3] and Latent Dirichlet Allocation [4] have been shown to be successful in a wide variety of applications concerning the analysis of continuous and discrete data, respectively. Probabilistic models are advantageous since they provide principled ways to address issues like the number of clusters, missing feature values, etc. Parametric mixture models are effective only when the

The research was partially supported by ONR grant no. N000140710225, NSF grant no. IIS-0643494. Part of Anil Jain's research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology(R31-2008-000-10008-0).

underlying distribution of the data is either known, or can be closely approximated by the distribution assumed by the model. This is a major shortcoming since it is well known that clusters in real data are not always of the same shape and rarely follow a "nice" distribution like Gaussian [5]. In a general setting, each cluster may follow its own unknown distribution, which limits the performance of parametric mixture models. Similar shortcomings can be attributed to squared error based clustering algorithms such as K-means, which is one of the most popular clustering algorithms due to its ease of implementation and reasonable empirical performance [1].

The limitations of parametric mixture models can be overcome by the use of algorithms that exploit non-parametric density estimation methods. Several non-parametric clustering algorithms, for instance, Jarvis-Patrick [6], DBSCAN [7] and Mean-shift [8], have been proposed [1]. These methods first find a single kernel-density estimate of the entire data, and then detect clusters by identifying modes or regions of high density in the estimated density [8]. Despite their success, most of these approaches are not always successful in finding clusters in high-dimensional datasets, since it is difficult to define the neighborhood of a data point in a high-dimensional space when the available sample size is small [9]. For this reason, almost all non-parameteric density based algorithms have been applied only to low-dimensional clustering problems such as image segmentation [8,10]. Further, it is not possible to a priori specify the desired number of clusters in these methods.

In this paper, we assume that each cluster is generated by its own density function that is unknown. The density function of each cluster may be arbitrary and multimodal and hence it is modeled using a non-parametric kernel density estimate. The overall data is modeled as a mixture of the individual cluster density functions. Since the proposed approach, unlike other non-parametric algorithms (e.g., Spectral clustering), constructs an explicit probabilistic model for each cluster, it can naturally handle out-of-sample[2] clustering by computing the posterior probabilities for new data points. In summary, we emphasize that:

- The proposed approach is a non-parametric probabilistic model for data clustering, and offers several advantages compared to non-probabilistic models since (a) it allows for probabilistic assignments of data points to different clusters, unlike other non-parametric models (b) it can effectively explore probabilistic tools such as Dirichlet process and Gaussian process for non-parametric priors, and (c) the model naturally supports out of sample cluster assignments, unlike other non-parametric models.
- Contrary to most existing mixture models, the proposed approach does not make any explicit assumption about the parametric form of the underlying density function, and can model clusters following arbitrary densities.

[1] Although spectral clustering and linkage methods can be viewed as non-parametric methods, they are not discussed since they are not probabilistic models.

[2] A clustering algorithm can perform *out-of-sample* clustering if it can assign a cluster label to a data point unseen during the learning phase.

We show the performance of the proposed clustering algorithm on high-dimensional text datasets. Experiments demonstrate that, compared to several widely used clustering algorithms such as K-means and Spectral clustering, the proposed algorithm performs significantly better when data is of high dimensionality and is embedded in a low dimensional manifold.

## 2 Non-parametric mixture model

### 2.1 Model description

Let $\mathcal{D} = \{x_1, \dots, x_n\}$ be a collection of $n$ data points to be clustered, where each $x_i \in \mathbb{R}^d$ is a vector of $d$ dimensions. Let $G$ be the number of clusters. We aim to fit the data points in $\mathcal{D}$ by a non-parametric mixture model. Let $\kappa(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the kernel function for density estimation. We further assume that the kernel function is stationary, i.e., $\kappa(x_i, x_j) = \kappa_s(x_i - x_j)$, where $\int \kappa_s(x)dx = 1$. We denote by the matrix $K = [\kappa(x_i, x_j)]_{n\times n} \in \mathbb{R}_+^{n\times n}$ the pairwise kernel similarity for data points in $\mathcal{D}$.

Let $\{c_g\}, g = 1, \dots, G$ be the set of $G$ clusters that forms a partition of $\mathcal{D}$. We specify the conditional density function $p_g(x|c_g, \mathcal{D})$ for each cluster $c_g$ as follows:

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

where $|c_g|$ is the number of samples in cluster $c_g$, and $\sum_g |c_g| = n$. The unconditional (on clusters) density $p(x|\mathcal{D})$ is then written as

$$p(x|\mathcal{D}) = \sum_{g=1}^{G} \pi_g p_g(x|c_g, \mathcal{D}) \tag{2}$$

where $\pi_g = P(c_g)$ is the mixture coefficient for cluster $c_g$. We generalize the cluster conditional density $p(x|c_g, \mathcal{D})$ in (1) by considering soft cluster participation, i.e each data point $x_i$ contributes $q_i^g \in [0, 1]$ to the kernel density estimate of the cluster $c_g$.

$$p_g(x|c_g, \mathcal{D}) = \sum_{i=1}^{n} q_i^g \kappa(x_i, x), \text{where} \sum_{i=1}^{n} q_i^g = 1. \tag{3}$$

We refer to $q^g = (q_1^g, \dots, q_n^g)$ as the *profile vector* for cluster $c_g$, and $Q = (q^1, \dots, q^G)$ as the *profile matrix*. The objective of our clustering model is to learn the profile matrix $Q$ for data set $\mathcal{D}$. We emphasize that due to the normalization step, i.e., $\sum_{j=1}^n q_j^g = 1$, $q_j^g$ can no longer be interpreted as the probability of assigning $x_j$ to cluster $c_g$. Instead, it only indicates the relative importance of $x_j$ to the density function for cluster $c_g$. The density function in (3) is also referred to as the density estimate in “dual form” [11].

**Algorithm 1** $[Q, \mathit{\Gamma}] = \mathrm{NonParametricMixtureFit}(\mathcal{D}, G, \lambda, \sigma)$

```
Input: Dataset $\mathcal{D}$, no. of clusters $G$, parameters $\lambda$ and $\sigma$
Output: Cluster labels $\mathit{\Gamma}$ and the profile matrix $Q$
Compute the kernel matrix $K$ for the points in $\mathcal{D}$ with bandwidth $\sigma$. Normalize $K$ such that $\sum_j K_{ij} = 1$.
Set the iteration $t \leftarrow 0$.
Initialize $Q^{(t)} \leftarrow Q_0$, such that $Q_0 \succcurlyeq 0$, $Q_0^T 1_n = 1_G$.
repeat
    $t \leftarrow t + 1$;
    Compute the $\gamma_i^g$ using Eq (9)
    By fixing the values of $\gamma_i^g$, obtain $Q^{(t)}$ by minimizing Eq (7).
    $\mathit{\Delta} Q \leftarrow Q^{(t)} - Q^{(t-1)}$.
until $||\mathit{\Delta} Q||_2^2 \leq \epsilon$, ($\epsilon$ is pre-set to a desired precision)
return $Q, \mathit{\Gamma}$
```

### 2.2 Estimation of profile matrix $Q$

To estimate the profile matrix $Q$, we follow the idea of maximum likelihood, i.e., find the matrix $Q$ by solving the optimization problem $\max_Q \sum_{i=1}^n \log p(x_i|\mathcal{D})$. One major problem with this approach is that, when estimating $p(x_i|\mathcal{D})$, $x_i$ is already an observed data point in $\mathcal{D}$ that is used to construct the density function $P(x_i|\mathcal{D})$. As a result, simply maximizing the likelihood of data may lead to an overestimation of the parameter $Q$, a problem that is often referred to as overfitting in machine learning [12]. We resolve this problem by replacing $p(x_i|\mathcal{D})$ with its leave-one-out (LOO) estimate [13].

Let $p_i(x_i|c_g, \mathcal{D}_{-i})$ be the LOO conditional probability for each held out sample $x_i$, conditioned on the clusters and the rest of the data:

$$p_i(x_i|c_g, \mathcal{D}_{-i}) = \frac{1}{\sum_{j=1}^n (1 - \delta_{j,i}) q_j^g} \sum_{j=1}^n (1 - \delta_{j,i}) q_j^g K_{i,j}, \tag{4}$$

where $D_{-i} = \mathcal{D} \backslash \{x_i\}$ denotes the subset of $\mathcal{D}$ that excludes sample $x_i$. Using the LOO cluster conditional probability $p_i(x_i|c_g, \mathcal{D}_{-i})$, we further define the LOO unconditional (on cluster) density for each held out sample $x_i$ as follows:

$$p_i(x_i|\mathcal{D}_{-i}) = \sum_{g=1}^G \gamma_i^g p_i(x_i|c_g, D_{-i}). \tag{5}$$

where $\gamma_i^g = P(c_g|\mathcal{D}_{-i})$, and $\sum_g \gamma_i^g = 1, \forall i = 1, \ldots, n$. Note that unlike the mixture model in (2) where the same set of mixture coefficients $\{\pi_g\}_{g=1}^G$ is used for any $x_i$, the mixture coefficients $\{\gamma_i^g\}_{g=1}^G$ depend on sample $x_i$, due to the leave-one-out estimation. We denote by $\gamma_i = (\gamma_i^1, \cdots, \gamma_i^G)$ and $\mathit{\Gamma} = (\gamma_1, \ldots, \gamma_n)^\top \in \mathbb{R}_+^{n \times G}$.

To improve the robustness of estimation, we introduce a Gaussian prior for profile matrix $Q$, i.e.,

$$p(Q) \propto \exp\left(-\lambda \sum_i \sum_g [q_i^g]^2\right), \tag{6}$$

where $\lambda$ is a hyperparameter that will be determined empirically. For notational convenience, we set $K_{i,i} = 0$ in Eq (4). Now, using the condition $\sum_{i=1}^n q_i^g = 1$, the LOO log-likelihood of data, denoted by $\ell_{LOO}(\mathcal{D}; Q, \varGamma)$, can be expressed as follows

$$\begin{aligned}\ell_{LOO}(\mathcal{D}; Q, \varGamma) &= \log p(Q) + \sum_{i=1}^n \log p_i(x_i|\mathcal{D}_{-i}) \\ &= -\lambda \sum_{i=1}^n \sum_{g=1}^G (q_i^g)^2 + \sum_{i=1}^n \log\left(\sum_g \gamma_i^g \frac{\sum_{j=1}^n K_{i,j} q_j^g}{1 - q_i^g}\right). \end{aligned} \tag{7}$$

The parameters in the above simplified model are $\gamma_i^g$ and $q_i^g$, for $i = 1, \cdots, n$ and $g = 1, \cdots, G$. They are estimated by maximizing the LOO log-likelihood $\ell_{LOO}(\mathcal{D}; Q, \varGamma)$. The optimal values of $Q$ and $\varGamma$ can be obtained by solving the following optimization problem:

$$\{Q^*, \varGamma^*\} = \arg\max_{Q,\varGamma} \ell_{LOO}(\mathcal{D}; Q, \varGamma) \tag{8}$$

The optimization procedure is described in the following section.

### 2.3 Optimization methodology

To determine the optimal values of $\varGamma$ and $Q$ that maximize the log-likelihood in Eq (8), we apply an alternating optimization strategy [14]. At each iteration, we first optimize $\varGamma$ with fixed $Q$, and then optimize $Q$ with fixed $\varGamma$, as summarized below. For a fixed $Q$, the LOO log-likelihood of a sample $x_i$ is maximized when

$$\gamma_i^g = \delta(g, \arg\max_{g'} p_i(x_i|c_{g'}, \mathcal{D}_{-i})). \tag{9}$$

The variable $\gamma_i^g$ is closely related to the posterior distribution $\Pr(c_g|x_i)$, and therefore can be interpreted as the cluster label of the $i$-th sample, i.e., $\gamma_i^g = 1$ if $x_i \in c_g$ and 0, otherwise.

It is difficult to directly optimize the log-likelihood in Eq (7) with respect to $Q$. Therefore, we minimize a convex variational upper bound on the negative log-likelihood for efficient inference. At each iteration, we maintain a touch point between the bound and the negative log-likelihood function, which guarantees convergence to at least a local minima [15]. The procedure for finding $Q$ and $\varGamma$ that maximize the log-likelihood in Eq (7) is summarized in Algorithm 1. Upon convergence, the value of $\gamma_i$ determines the cluster label for $x_i$.

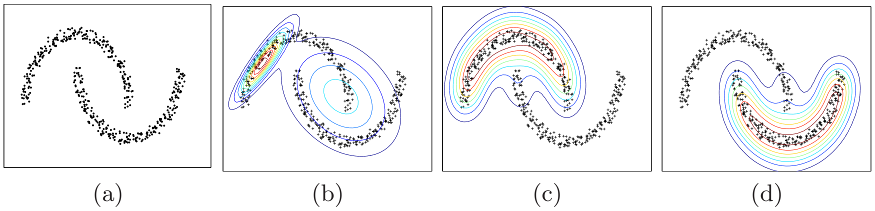

(a) (b) (c) (d)

**Fig. 1.** Illustration of the non-parametric mixture approach and Gaussian mixture model (GMM) on the "two-moon" dataset. (a) Input data with two clusters. (b) Gaussian mixture model with two components. (c) and (d) the iso-contour plot of non-paramateric estimates of the class conditional densities for each cluster. The warmer the color, the higher the probability.

### 2.4 Implementation details

Normalization is one of the key issues in kernel density estimation. Conventionally, the kernel function is normalized over the entire domain of the data, $\kappa_\sigma(\mathbf{x}) = (\pi\sigma)^{-d} \exp\left(-||\mathbf{x}||^2/2\sigma^2\right)$. However, the $\sigma^{-d}$ term may be close to 0 ($\sigma < 1$) or be very large ($\sigma > 1$). This may cause serious numerical problems in density estimation for high-dimensional data (large values of $d$) with small sample size. To overcome this problem, we normalize the kernel matrix such that each row sums to 1, i.e. $\sum_j K_{i,j} = 1$. This nullifies the effect of dimension on the estimation process, and therefore is useful in handling sparse datasets.

The heuristic used in spectral clustering [16] to select the value for $\sigma$ is also effective in estimating kernel width. Empirical results show that the clustering performance is not very sensitive to the choice of the kernel width $\sigma$. The parameter $\lambda$ also is not very critical and is chosen to be sufficiently small; in all of our experiments we choose $\lambda = 10^{-4}$, which results in mild smoothing of the $q_i^g$ values, and avoids any numerical instability in the algorithm due to the logarithm. The number of variables to be solved for is of $O(nG)$, similar to that of spectral clustering. On the other hand, Gaussian mixture models solve for $O(d^2)$ number of variables which is large, especially for high-dimensional sparse datasets (specifically when $(n+1)G < \left(dG + \frac{d(d+1)}{2}\right)$, as shown in Table 1).

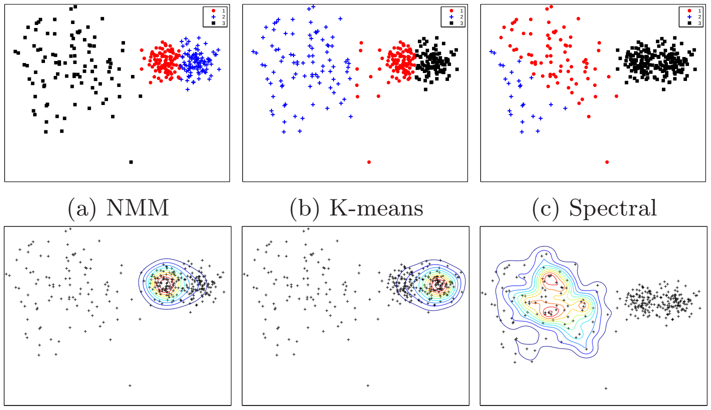
## 3 Results and discussion

The proposed non-parametric mixture fitting algorithm is evaluated on text datasets derived from the 20-newsgroups[3] dataset [17].

### 3.1 Baseline methods

The proposed non-parametric mixture algorithm is compared with three classes of well known clustering algorithms: (a) K-means and Gaussian mixture model

[3] http://people.csail.mit.edu/jrennie/20Newsgroups/



(a) NMM (b) K-means (c) Spectral

(d) (e) (f)

**Fig. 2.** Illustration of the non-parametric mixture approach, K-means and spectral clustering on the example dataset from [18]. Input data contains 100 points each from three spherical two-dimensional Gaussian clusters with means (0,0), (6,0) and (8,0) and variances $4I_2$,$0.4I_2$ and $0.4I_2$ respectively. Spectral clustering and NMM use $\sigma = 0.95$. (a) NMM (b) K-means (c) Spectral clustering. Plots (d)-(f) show the cluster-conditional densities estimated by the proposed NMM.

(GMM) with diagonal and full covariance matrices, (b) one kernel-based algorithm: NJW spectral clustering [19], and (c) three non-parametric hierarchical clustering algorithms: Single Link, Complete Link and Average Link. For (a) and (c), we use the implementations from the Matlab's Statistics Toolbox. For the linkage based methods, the number of clusters is externally specified. We chose the state-of-the-art spectral clustering algorithm implementation based on [19]. Each algorithm is run 10 times and the mean performance value is reported in Table 1, with the best performance shown in bold face. Comparison with Mean-shift, or related algorithms is difficult as the datasets are high-dimensional and further, it is not possible to specify the number of clusters in these algorithms. Since the number of dimensions is greater than the number of data points, GMM is not succesful for this data.

At each run, the proposed algorithm, K-means and Spectral clustering were initialized with 5 different starting points; only the best performance is reported. Due to the space limitation, we only show the best performance among the three hierarchical linkage based algorithms, without specifying which algorithm achieved it.

### 3.2 Synthetic Datasets

The proposed algorithm aims at identifying clusters of arbitrary shapes, while estimating their conditional density. Figure 1 illustrates the performance of NMM

**Table 1.** Mean pairwise $F_1$ value for different clustering algorithms over 10 runs of each algorithm on eight high-dimensional text datasets. The kernel width is chosen as the $5^{th}$ percentile of the pairwise Euclidean distances for Kernel based algorithms. The best performance for each dataset is shown in bold. The name of the dataset, number of samples (n), dimensions (d), and the number of target clusters (G) are shown in the first 4 columns, respectively. The last column shows the best $F_1$ value achieved by Single (S), Complete (C) and Average (A) link algorithms.

| | | | | Proposed | K-means | NJW-Spec | Linkage |
|---|---|---|---|---|---|---|---|
| Dataset | n | d | G | | | | max(S,C,A) |
| cmu-different-1000 | 2975 | 7657 | 3 | **95.86** | 87.74 | 94.37 | 40.31 |
| cmu-similar-1000 | 2789 | 6665 | 3 | **67.04** | 49.86 | 45.16 | 37.28 |
| cmu-same-1000 | 2906 | 4248 | 3 | **73.79** | 49.40 | 48.04 | 30.01 |
| cmu-different-100 | 300 | 3251 | 3 | **95.27** | 79.22 | 87.47 | 75.74 |
| cmu-similar-100 | 288 | 3225 | 3 | **50.89** | 40.10 | 38.35 | 43.82 |
| cmu-same-100 | 295 | 1864 | 3 | **48.97** | 44.85 | 46.99 | 41.79 |
| cmu-classic300 | 300 | 2372 | 3 | 85.32 | **86.32** | 86.02 | 80.61 |
| cmu-classic400 | 400 | 2897 | 3 | **61.26** | 60.13 | 51.01 | 53.31 |

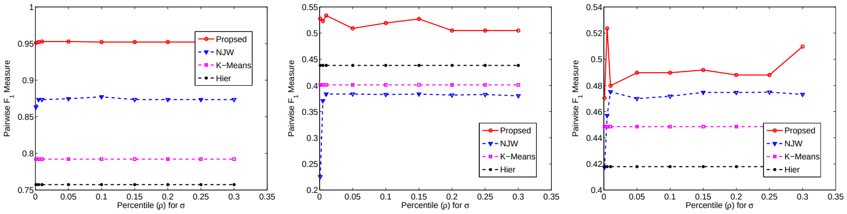
on a dataset not suitable for GMM. Figure 1(a) shows the input data. Figure 1(b) is shown to contrast the proposed non-parametric mixture approach against the parametric Gaussian mixture model (GMM) with the number of mixture components set to two. Figures 1(c) and (d) show the class conditional densities for each of the two clusters. The proposed algorithm is able to recover the underlying clusters, as well as estimate the associated conditional densities, which is not possible for GMM as shown in Figure 1(b).

Figure 2 illustrates the performance of the proposed algorithm on a dataset that is known to be difficult for spectral clustering [18]. Both K-means and spectral clustering fail to recover the clusters due to the difference in the variance of the spherical clusters. The proposed algorithm however, is purely local, in that the cluster label of a point is affected only by the cluster labels of neighboring points. The clusters, therefore, are recovered nearly perfectly by the proposed algorithm as shown in Figure 2(a) and the cluster conditional densities are shown in Figures 2(d)-(f).

### 3.3 Text Datasets

We use eight high dimensional text datasets to show the efficacy of the algorithm. These datasets are popularly used in document clustering [20].

Table 1 shows that the proposed non-parametric mixture (NMM) algorithm performs significantly better (paired t-test, 95% confidence) than the other clustering methods on all the high dimensional text datasets, except for `cmu-classic-300`, where its performance is slightly inferior to K-means. Since the datasets are high-dimensional, and non-spherical, the proposed approach outperforms K-means on



(a) Different-100 (b) Similar-100 (c) Same-100

**Fig. 3.** Performance of the non-parametric mixture model on three text datasets, with varying value of the percentile ($\rho$) for choosing the kernel bandwidth ($\sigma$). The proposed algorithm is compared with NJW (Spectral clustering), K-means and the best of three linkage based methods.

most of the datasets. Spectral clustering considers only the top $G-1$ eigenvectors for clustering a dataset into $G$ clusters; the superior performance of the proposed NMM can be attributed to its utilization of the complete kernel matrix without discarding any portion of it. These datasets could not be clustered by GMM (Gaussian mixture models) since they are prone to numerical estimation problems when the number of dimensions is larger than the number of samples.

### 3.4 Sensitivity to parameters:

There are two parameters in the non-parametric mixture clustering algorithm: the regularizer weight $\lambda$ and the kernel width $\sigma$. The parameter $\sigma$ is set to the $\rho^{th}$ percentile of the pairwise Euclidean distances among the data points. A useful range for $\rho$ is 5-10%, as suggested in [16]. Figure 3 shows the performance of the proposed algorithm in comparison to $K$-means, spectral clustering and hierarchical clustering on three text datasets. These plots show that there exists a wide range of kernel bandwidth values for which the proposed algorithm performs significantly better than the competing methods. For some datasets (e.g., Different-100 and Classic-400), the algorithm is more stable compared to that of other datasets. We observed that the algorithm is not sensitive to the value of $\lambda$, over the range $(10^{-4}, 10^{4})$. While the performance is the same for almost all the values of $\lambda$, the parameter $\lambda$ does play a role in determining the sparsity of the profile matrix. As $\lambda$ increases, the profile of data points between the clusters tends to get smoother. The key role of $\lambda$ is to provide numerical stability to the algorithm.

## 4 Conclusions and future work

We have proposed a non-parametric mixture model for data clustering. It is a probablistic model that clusters the data by fitting a kernel density estimate to each cluster. Experimental results show that the non-parametric mixture model based clustering outperforms some of the well known clustering algorithms on the task of document clustering, which can be characterized as high dimensional sparse data. The non-parametric mixture model opens up a wide range of possible theoretical analysis related to clustering, which is a part of our ongoing work. Automatic kernel bandwidth selection, scalability of the algorithm and application to other sparse data domains (e.g., bioinformatics) are possible extensions.

## References

1. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters **31** (2010) 651–666 1, 2
2. McLachlan, G.L., Peel, D.: Finite Mixture Models. Wiley (2000) 1
3. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. TPAMI **24** (2002) 381–396 1
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022 1
5. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988) 2
6. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Transactions on Computers **22** (1973) 9 2
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. KDD. (1996) 226–231 2
8. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (May 2002) 603–619 2
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006) 2
10. Andreetto, M., Zelnik Manor, L., Perona, P.: Non-parametric probabilistic image segmentation. In: Proceedings of the ICCV. (2007) 1–8 2
11. Shawe-Taylor, J., Dolia, A.N.: A framework for probability density estimation. In: Proc. AISTATS. (2007) 468–475 3
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2001) 4
13. Wand, M.P., Jones, M.C.: Kernel Smoothing (Monographs on Statistics and Applied Probability). Chapman & Hall/CRC (December 1994) 4
14. Csiszar, I., Tusnady, G.: Information geometry and alternating minimization procedures. Statistics and Decision (1984) 5
15. Jaakkola, T.S.: Tutorial on variational approximation methods. In: Advanced Mean Field Methods: Theory and Practice. (2000) 129–159 5
16. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 888–905 6, 9
17. Slonim, N., Tishby, N.: Agglomerative information bottleneck. In: Advances in NIPS. (2000) 6
18. Nadler, B., Galun, M.: Fundamental limitations of spectral clustering. In: NIPS 19, Citeseer (2007) 1017–1025 7, 8
19. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in NIPS, MIT Press (2001) 849–856 7
20. Banerjee, A., Langford, J.: An objective evaluation criterion for clustering. In: Proceedings of the KDD. (2004) 515–520 8