Clustering via Kernel Decomposition

A. Szymkowiak-Have*, M.A.Girolami †, Jan Larsen*
*Informatics and Mathematical Modelling
Technical University of Denmark, Building 321
DK-2800 Lyngby, Denmark
Phone: +45 4525 3899,3923
Fax: +45 4587 2599
E-mail: asz,jl@imm.dtu.dk
Web: isp.imm.dtu.dk
†Department of Computing Science
University of Glasgow, UK
Phone: +44 141 330 8628
Fax: +44 141 330 8627
E-mail: girolami@dcs.gla.ac.uk

***Abstract*— Spectral clustering methods were proposed recently which rely on the eigenvalue decomposition of an affinity matrix. In this work the affinity matrix is created from the elements of a non-parametric density estimator and then decomposed to obtain posterior probabilities of class membership. Hyperparameters are selected using standard cross-validation methods.**

***Index Terms*— spectral clustering, kernel decomposition, aggregated Markov model, kernel principal component analysis**

I. Introduction

The spectral clustering methods [1], [2], [3] are attractive in the case of complex data sets, which possess for example manifold structures, when the classical models such as K-means often fail in the correct estimation. The proposed models in the literature are, however, incomplete, since they do not offer methods for the estimation of the model hyperparameters which have to be manually tuned [1]. The need arises to construct a self-contained model, which would not only provide accurate clustering but also which would estimate both the model complexity and all the necessary parameters for estimation. The additional advantage can also be provided by the probabilistic outcome, where the confidence in the point assignment to the clusters is given.

The kernel principal component analysis (KPCA) [4] decomposition of a Gram matrix has been shown to be a particularly elegant method for extracting nonlinear features from multivariate data. KPCA has been shown to be a discrete analogue of the Nyström approximation to obtaining the eigenfunctions of a process from a finite sample [5]. Building on this observation the relationship between KPCA and non-parametric orthogonal series density estimation was highlighted in [6], and the relation with spectral clustering has recently been investigated in [7]. The basis functions obtained from KPCA can be viewed as the finite sample estimates of the truncated orthogonal series [6], however, a problem common to orthogonal series density estimation is that the strict non-negativity required of a probability density is not guaranteed when employing these finite order sequences to make point estimates [8], this is of course also observed with the KPCA decomposition [6].

To further explore the relationship between the decomposition of a Gram matrix, the basis functions obtained from KPCA and density estimation, a matrix decomposition which maintains the positivity of point probability density estimates is desirable. In this paper we show that such a decomposition can be obtained in a straightforward manner and we observe useful similarities between such a decomposition and spectral clustering methods [1], [2], [3].

The following sections consider the non-parametric estimation of a probability density from a finite sample [8] and relates this to the identification of class structure within the density from the sample. Two kernel functions, the choice of which dependent on the data type and dimensionality, are proposed. The derivation of the generalization error is also presented, which enables the determination of the model parameters and model complexity [9], [10]. The experiments are performed on artificial data sets as well as on more realistic collections.

II. Density Estimation and Decomposition of the Gram Matrix

Consider the estimation of an unknown probability density function $p(\mathbf{x})$ from a finite sample of N points $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ where $\mathbf{x} \in \mathcal{R}^d$. The sample drawn from the density can be employed to estimate the density in a non-parametric form by using a Parzen window estimator (refer to [8], [9], [10] for a review of such non-parametric density estimation methods) such that the estimate is given by

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{K}_h(\mathbf{x}, \mathbf{x}_n) \tag{1}$$

where $\mathcal{K}_h(\mathbf{x}_i, \mathbf{x}_j)$ denotes the kernel function of width h, between points $\mathbf{x}_i$ and $\mathbf{x}_j$, which itself satisfies the requirements of a density function [8]. It is important to note that

the pairwise kernel function values $\mathcal{K}_h(\mathbf{x}_i, \mathbf{x}_j)$ provide the necessary information regarding the sample estimate of the underlying probability density function $p(\mathbf{x})$. Therefore the kernel or Gram matrix constructed from a sample of points (and a kernel function which itself is a density) provides the necessary information to faithfully reconstruct the estimated density from the pairwise kernel interactions in the sample.

For applications of unsupervised kernel methods such as KPCA the selection of the kernel parameter, in the case of the Gaussian kernel h, is often problematic. However, noting that the kernel matrix can be viewed as defining the sample density estimate, then methods such as leave-one-out cross-validation can be employed in obtaining an appropriate value of the kernel width parameter. We shall return to this point in the following sections.

A. Kernel Decomposition

The density estimate can be decomposed in the following probabilistic manner as

$$\hat{p}(\mathbf{x}) = \sum_{n=1}^{N} p(\mathbf{x}, \mathbf{x}_n) \quad (2)$$

$$= \sum_{n=1}^{N} p(\mathbf{x}|\mathbf{x}_n) P(\mathbf{x}_n) \quad (3)$$

such that each sample point is equally probable *a priori*, $P(\mathbf{x}_n) = N^{-1}$, i.e. data is assumed *independent and identically distributed (i.i.d)*. The kernel operation, such that the kernel is itself a density function, can then be seen to be the above conditional density $p(\mathbf{x}|\mathbf{x}_n) = \mathcal{K}_h(\mathbf{x}, \mathbf{x}_n)$.

The sample of N points drawn from the underlying density forms a set and as such we can define a probability space over the N points. A discrete posterior probability can be defined for a point $\mathbf{x}$ (either in or out of sample) given each of the N sample points

$$\hat{P}(\mathbf{x}_n|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{x}_n)P(\mathbf{x}_n)}{\sum_{n'=1}^{N} p(\mathbf{x}|\mathbf{x}_{n'})P(\mathbf{x}_{n'})}$$

$$= \frac{\mathcal{K}(\mathbf{x}, \mathbf{x}_n)}{\sum_{n'=1}^{N} \mathcal{K}(\mathbf{x}, \mathbf{x}_{n'})} \equiv \breve{\mathcal{K}}(\mathbf{x}, \mathbf{x}_n) \quad (4)$$

such that $\sum_{n=1}^{N} \hat{P}(\mathbf{x}_n|\mathbf{x}) = 1$, $\hat{P}(\mathbf{x}_n|\mathbf{x}) \geq 0 \ \ \forall\ n$ and each $P(\mathbf{x}_n) = \frac{1}{N}$.

Now if it is assumed that there is an underlying, hidden class/cluster structure in the density then the sample posterior probability can be decomposed by introducing a discrete class variable such that

$$\hat{P}(\mathbf{x}_n|\mathbf{x}) = \sum_{c=1}^{C} P(\mathbf{x}_n, c|\mathbf{x}) = \sum_{c=1}^{C} P(\mathbf{x}_n|c, \mathbf{x}) P(c|\mathbf{x}) \quad (5)$$

and noting that the sample points have been drawn i.i.d from the respective C classes forming the distribution such that points are independent given the class variable i.e.$\mathbf{x}_n \perp \mathbf{x} \mid c$, then

$$\hat{P}(\mathbf{x}_n|\mathbf{x}) = \sum_{c=1}^{C} P(\mathbf{x}_n, c|\mathbf{x}) = \sum_{c=1}^{C} P(\mathbf{x}_n|c) P(c|\mathbf{x}) \quad (6)$$

with constraints $\sum_{n=1}^{N} P(\mathbf{x}_n|c) = 1$ and $\sum_{c=1}^{C} P(c|\mathbf{x}) = 1$.

Considering the decomposition of the posterior sample probabilities for each point in the available sample $\hat{P}(\mathbf{x}_i|\mathbf{x}_j) = \sum_{c=1}^{C} P(\mathbf{x}_i|c)P(c|\mathbf{x}_j),\ \forall \quad i, j = 1, \cdots, N$ we see that this is identical to the aggregate Markov model originally proposed in [11], where the matrix of posteriors (elements of the normalized kernel matrix) can be now be viewed as an estimated state transition matrix for a first order Markov process. This decomposition then provides class posterior probabilities $P(c|\mathbf{x}_n)$ which can be employed for clustering purposes.

A divergence based criterion such as cross-entropy

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \breve{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j) \log \left\{ \sum_{c=1}^{C} P(\mathbf{x}_i|c)P(c|\mathbf{x}_j) \right\} \quad (7)$$

or distance based criterion such as squared error

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \left\{ \breve{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j) - \left\{ \sum_{c=1}^{C} P(\mathbf{x}_i|c)P(c|\mathbf{x}_j) \right\} \right\}^2 \quad (8)$$

subject to the constraints that each $P(\mathbf{x}_i|c)$ and $P(c|\mathbf{x}_j)$ are strictly positive and $\sum_{n=1}^{N} P(\mathbf{x}_n|c) = 1$, $\sum_{c=1}^{C} P(c|\mathbf{x}) = 1$. Due to these constraints which ensure that the decomposition provides interpretable probabilities then a matrix factorization which enforces positivity of the elements in the decomposition is required. There has been a number of recent publications which have proposed efficient methods for obtaining such constrained matrix decompositions [12], [13] and as such the non-negative matrix multiplicative update equations (NMF) [12], [13] or equivalently the iterative algorithm which performs Probabilistic Latent Semantic Analysis (PLSA) [14] can be employed in optimizing the above criteria subject to the required constraints.

If the normalized Gram matrix is defined as $\mathbf{G} = \{\hat{P}(\mathbf{x}_i, \mathbf{x}_j)\} = \breve{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j)$ then the decomposition of that matrix with NMF [12], [13] or PLSA [14] algorithms will yield $\mathbf{G} = \mathbf{WH}$ such that $\mathbf{W} = \{P(\mathbf{x}_i|c)\}$ and $\mathbf{H} = \{P(c|\mathbf{x}_j)\}$ are understood as the required probabilities which satisfy the previously defined stochastic constraints.

B. Clustering with the Kernel Decomposition

Having obtained the elements $P(\mathbf{x}_i|c)$ and $P(c|\mathbf{x}_j)$ of the decomposed matrix employing NMF or PLSA, the class posteriors $P(c|\mathbf{x}_j)$ will indicate the class structure of the samples. We are now in a position to assign newly observed *out-of-sample* points to a particular class. If we observe a new point $\mathbf{z}$ in addition to the sample then the estimated decomposition components can, in conjunction with the kernel, provide the required class posterior $P(c|\mathbf{z})$.

$$\hat{P}(c|\mathbf{z}) = \sum_{n=1}^{N} P(c|\mathbf{x}_n)\hat{P}(\mathbf{x}_n|\mathbf{z}) \quad (9)$$

$$= \sum_{n=1}^{N} P(c|\mathbf{x}_n)\breve{\mathcal{K}}(\mathbf{z}, \mathbf{x}_n) \quad (10)$$

$$= \sum_{n=1}^{N} P(c|\mathbf{x}_n) \frac{\mathcal{K}(\mathbf{z}, \mathbf{x}_n)}{\sum_{n'=1}^{N} \mathcal{K}(\mathbf{z}, \mathbf{x}_{n'})} \quad (11)$$

This can be viewed as a form of 'kernel' based non-negative matrix factorization where the 'basis' functions $P(c|\mathbf{x}_n)$ define the class structure of the estimated density.

For the case of a Gaussian (radial basis function) kernel this interpretation of a kernel based clustering motivates the definition of the kernel smoothing parameter by means of out-of sample predictive likelihood and as such cross-validation can be employed in estimating the kernel width parameter. In addition the problem of choosing the number of possible classes, a problem common to all non-parametric clustering methods such as spectral clustering [1],[15] can now be addressed using theoretically sound model selection methods such as cross-validation. This overcomes the lack of an objective means of selecting the smoothing parameter in most other forms of spectral clustering [1],[15] as the proposed method first defines a non-parametric density estimate, and then the inherent class structure is identified by the basis decomposition of the normalized kernel in the form of class conditional posterior probabilities. This highlights another advantage, over partitioning based methods [1], [15], of this view on kernel based clustering in that projection coefficients are provided enabling new or previously unobserved points to be allocated to clusters. Thus projection of the normalized kernel function of a new point onto the class-conditional basis functions will yield the posterior probability of class membership for the new point.

In attempting to identify the *model order*, eg. number of classes, a generalization error based on the out-of-sample negative predictive likelihood, is defined as follow

$$\mathcal{L}_{out} = N_{out}^{-1} \sum_{n=1}^{N_{out}} \log \{p(\mathbf{z}_n)\} \tag{12}$$

where N_{out} denotes the number of out-of-sample points. The out-of-sample likelihood (12) is derived from the decomposition in the following manner

$$p(\mathbf{z}) = \frac{1}{N} \sum_{n=1}^{N} p(\mathbf{z}|\mathbf{x}_n) \tag{13}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} p(\mathbf{z}|c) P(c|\mathbf{x}_n) \tag{14}$$

The $p(\mathbf{z}|c)$ can be decomposed given the finite sample such that $p(\mathbf{z}|c) = \sum_{l=1}^{N} p(\mathbf{z}|\mathbf{x}_l) P(\mathbf{x}_l|c)$ where $p(\mathbf{z}|\mathbf{x}_l) = \mathcal{K}(\mathbf{z}|\mathbf{x}_l)$. So the unconditional density estimate of an out of sample point given the current kernel decomposition which assumes a specific class structure in the data can be computed as the following.

$$p(\mathbf{z}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{N} \sum_{c=1}^{C} \mathcal{K}(\mathbf{z}|\mathbf{x}_l) P(\mathbf{x}_l|c) P(c|\mathbf{x}_n), \tag{15}$$

where $P(\mathbf{x}_l|c) = \mathbf{W}$ and $P(c|\mathbf{x}_n) = \mathbf{H}$ are estimated parameters.

C. Kernels

For continuous data such that $\mathbf{x} \in \mathcal{R}^d$ a common choice of kernel, for both kernel PCA and density estimation, is the isotropic Gaussian kernel

$$\mathcal{K}_h(\mathbf{x}, \mathbf{x}_n) = (2\pi)^{-\frac{d}{2}} h^{-d} \exp\left\{ -\frac{1}{2h^2} ||\mathbf{x} - \mathbf{x}_n||^2 \right\} \tag{16}$$

Of course many other forms of kernel can be employed, though they may not themselves satisfy the requirements of being a density. For example in the case of vector space representations of text the standard similarity measure employed is the cosine inner-product.

$$\mathcal{K}(\mathbf{x}, \mathbf{x}_n) = \frac{\mathbf{x}^T \mathbf{x}_n}{||\mathbf{x}|| \cdot ||\mathbf{x}_n||}. \tag{17}$$

The decomposition of this cosine based Gram matrix directly will yield the required probabilities.

Although, the cosine inner-product does not satisfy itself the density requirements ($\int \mathcal{K}(\mathbf{x}, \mathbf{x}_n) = 1$) it can be applied in the presented model as long as the kernel integral is finite. This condition is satisfied when the data points are a priori normalized, e.g. to the unit sphere and the empty vectors are excluded from the data set. The density values obtained from such non-density kernels provide the incorrect generalization errors which are scaled by the unknown constant factor and therefore can be still used in estimation of the parameters.

This interpretation provides a means of spectral clustering which, in the case of continuous data, is linked directly to non-parametric density estimation and extends easily to discrete data such as for example text. We should also note that the aggregate Markov perspective allows us to take the random walk viewpoint as elaborated in [15] and so a K-connected graph[1] may be employed in defining the kernel similarity $\mathcal{K}_K(\mathbf{x}, \mathbf{x}_n)$. Similarly to the smoothing parameter and the number of clusters, the number of connected points in the graph can be also estimated from the generalization error.

The following experiments and subsequent analysis provide an objective assessment and comparison of a number of classical and recently proposed clustering methods.

III. Experiments

In the experiments we used, the four following data sets described below.

1) ***Linear structure.*** Data set consist of five 2-dimensional Gaussian distributed clusters with a spherical covariance structure, shown in the left plot of figure 1 (left plot). The clusters are linearly separable. This artificially created data is used for illustration of a simple clustering problem.
2) ***Manifold structure.*** Data set consist of three clusters as shown in the right plot of figure 1. Clusters are formed in the shape of rings all centered at the origin with radii 3, 5 and 8, respectively. The ring structure is a standard example used in spectral clustering, for example [1], [2]. The data is 2 dimensional. This is given as an example of complex nonlinear data on which methods such as K-means will fail.

[1]The K-connected graph is performed by remaining the dependencies between only K closest points.

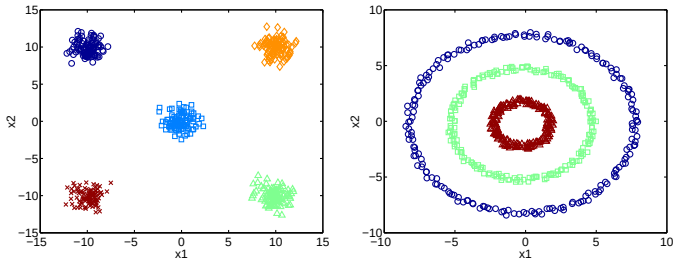


Fig. 1. The scatter plots of the artificial data for 5 Gaussian distributed clusters (left figure) and 3 cluster ring formations (right panel).

3) ***Email collection.*** The Email data set[2] consist of emails grouped into three categories: *conference*, *job* and *spam*, used earlier in [16], [17], [18]. The collection was hand-labeled. In order to process text data a *term-vector* is defined as the complete set of all the words existing in all the email documents. Then each email document is represented by a *histogram*: the frequency vector of occurrences of each of the word from a term-vector. The collection of such email histograms is denoted *the term-document matrix*. In order to achieve good performance, suitable preprocessing is performed. It includes removing stopwords[3] and other high and low frequency words, stemming[4] and normalizing histograms to unit $\mathcal{L}_2$-norm length. After preprocessing, the term-document matrix consist of 1405 email histograms described by 7798 terms. The data points are discrete and high dimensional and the categories are not linearly separable.
4) ***Newsgroups.*** The collection[5] consist originally of 20 categories each containing approximately 1000 newsgroup documents. In the performed experiments 4 categories (*computer graphics*, *motorcycles*, *baseball* and *Christian religion*) were selected each containing 200 instances. The labels of the collection are selected based on the catalogs names the data was stored in. The data was processed in a similar way as that presented above. In preprocessing 2 documents were removed[6]. After preprocessing the data consists of 798 newsgroup documents described in the space of 1368 terms.

In the case of continuous space collections (Gaussian and Rings clusters) data vectors were normalized with its maximum value so, they fall in the range between 0 and 1. This step is necessary when the features describing data points have significantly different values and ranges. The normalization to the unit $\mathcal{L}_2$-norm length was applied for the Email and Newsgroup collections.

For Gaussian and Rings clusters the isotropic Gaussian kernel equation (16) is used. With discrete data sets (Emails and Newsgroups) the cosine inner-product equation (17) is applied.

[2]The Email database ia available at http://isp.imm.dtu.dk/staff/anna

[3]Stopword are high frequency words that are helping to build the sentence, e.g. conjunctions, pronouns, prepositions etc. A list of 584 stopwords is used in the experiments

[4]Stemming denotes the process of merging words with typical endings into the common stem. For example for English language the endings like e.g. *-ed*, *-ing*, *-s* are considered.

[5]The Newsgroups collection is available at e.g. http://kdd.ics.uci.edu/

[6]Reduction in term space (stopwords removing, stemming, etc.) resulted with empty documents, which were removed from the data set.

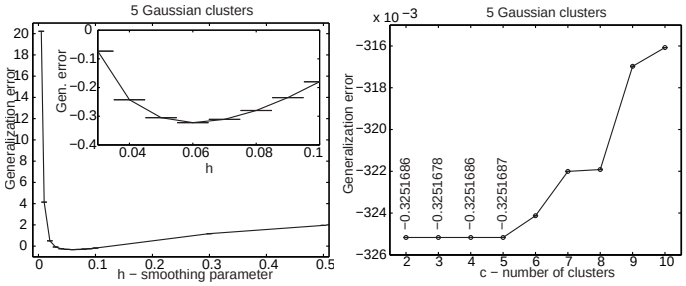


Fig. 2. The generalization error as a function of smoothing parameter h (left panel) for 5 Gaussian distributed clusters. The optimum choice is $h = 0.06$. The right figure presents, for optimum smoothing parameter, the generalization error as a function of number of clusters. Here, any cluster number below or equal 5 may give the minimum error for which the error values are shown above the points. The optimum choice is a maximum model, i.e. $K = 5$ (see the explanation in the text). The error-bars show $\pm$ standard error of the mean value.

The Gaussian clusters example is a simple linear separation problem. The model was trained using 500 randomly generated samples, and generalization error computed from 2500 validation set samples. The aggregated Markov model, as a probabilistic framework, allows the new data points, not included in the training set, to be uniquely mapped in the model. It is possible to select optimum model parameters: h, K in K-connected graph in discrete data sets and the optimum number of clusters c by minimizing the generalization error defined by the equations (12) and (14).

In 20 experiments, different training sets were generated, and the final error is an average over 20 outcomes of the algorithm on the same validation set. The left plot of figure 2 presents the dependency of the generalization error as a function of the kernel smoothing parameter h, averaged for all the model orders. The minimum is obtained for $h = 0.06$. For that optimum h the model complexity c is then investigated (right plot of figure 2). Here, the minimal error is obtained for all 2, 3, 4 and 5 clusters and as the optimal solution 5 clusters are chosen, which is explained in appendix .

For Gaussian clusters the cluster posterior $p(c|\mathbf{z})$ is presented on figure 3. Perfect decision surfaces can be observed. For comparison, on figure 4, the components of the traditional

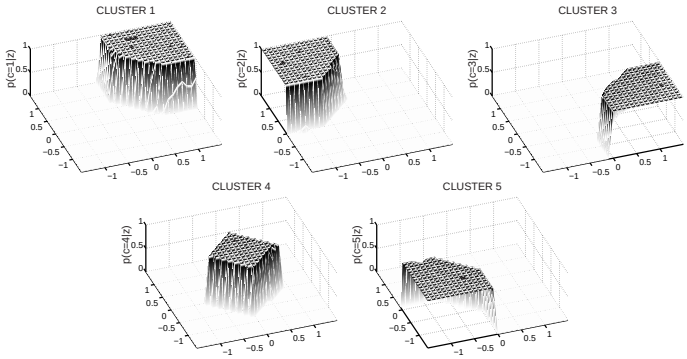


Fig. 3. The cluster posterior values $p(c|\mathbf{z})$ obtained from the aggregate Markov model for Gaussian clusters. The decision surfaces are positive. The separation in this case is perfect.

kernel PCA are presented. Here, both the positive and the negative values are observed, which makes it difficult to determine the optimum decision surface.

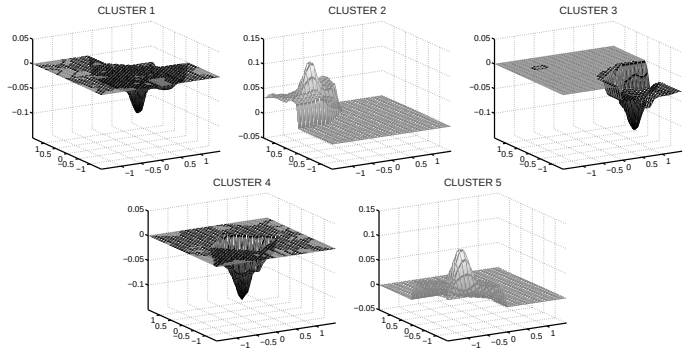


Fig. 4. The components of the traditional Kernel PCA model for Gaussian clusters. The decision surfaces are both positive and negative.

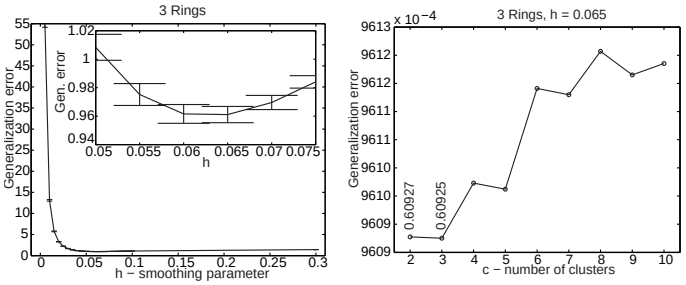


Fig. 5. The generalization error as a function of smoothing parameter h for 3 clusters formed in the shape of rings (left panel). The optimum choice is $h = 0.065$. On the right figure the generalization error as a function of number of clusters is shown for the optimum choice of smoothing parameter. The error bars shows the standard error of the mean value.

The Ring data is a highly nonlinear clustering problem. In the experiments, 600 examples were used for training, for generalization 3000 validation set samples were generated and the experiments were repeated 40 times, with different training sets. The generalization error, shown on figures 5, is an average over errors obtained in each of the 40 runs on the same validation set and for all the model orders c. The optimum smoothing parameter (figure 5, left plot) is equal $h = 0.065$ and the minimum in generalization error is obtained for 3 clusters. As in the Gaussian clusters example, a smaller model of 2 clusters is also probable[7].

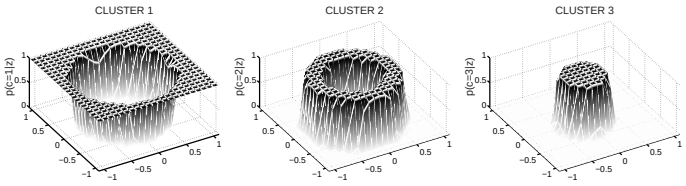


Fig. 6. The cluster posterior values $p(c|\mathbf{z})$ obtained from the aggregate Markov model for Rings. The decision surfaces are positive. The separation in this case is perfect.

The cluster posterior for Rings data set and the kernel PCA components are presented in figures 6 and 7, respectively. Also

[7]The generalization error is similar for both 2 and 3 numbers of clusters.

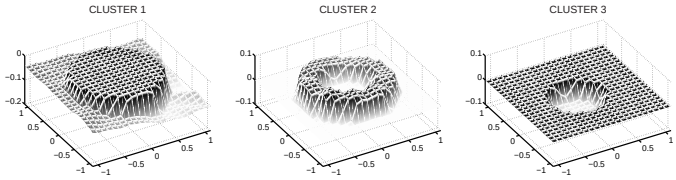
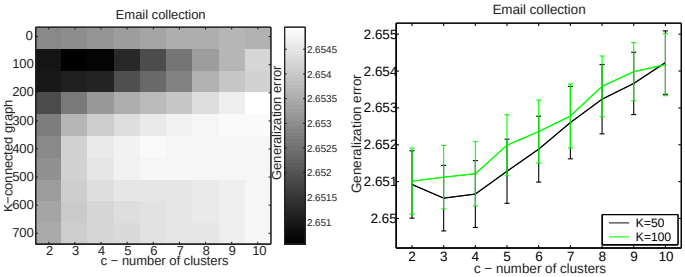


Fig. 7. The components of the traditional Kernel PCA model for Rings structure. The decision surfaces are both positive and negative.

in this case perfect (0/1) decision surfaces are observed (figure 6), which are the outcome of the aggregated Markov model. The components of kernel PCA present, as in the previous case, the separation possibility but with more ambiguity for selection the decision surface.



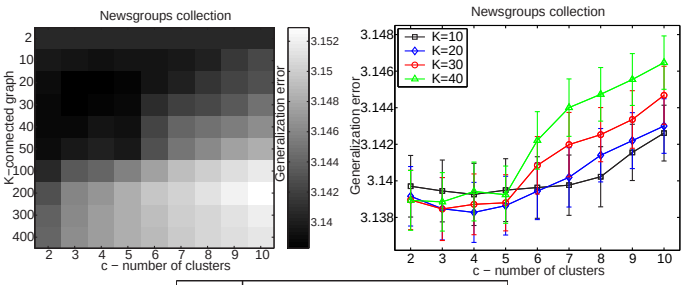
	CONF	JOB	SPAM
1	1.5	1.6	99.5
2	9.7	97.6	0.3
3	88.8	0.8	0.3

Fig. 8. Left upper panel presents the mean generalization error as a function of both the cluster number and the k - cutoff threshold in the k-connected graph for Emails collection. For clarity in made decision the selected cut off thresholds (K=50 and K=100) are shown on the right plot. The optimal model is the choice of $K = 50$ (50-connected graph) with 3 clusters. Lower figure presents the confusion matrix for labeling produced by the selected optimal model and the original labeling. Only the small confusion can be observed.

The generalization error for Email collection is shown in left plot of figure 8. The mean values are presented averaged from 20 random choices of the training and the test set. For training 702 samples are reserved and the rest of 703 examples is used in calculation of the generalization error. Since, used kernel is the cosine inner-product, the K-connected graph is applied to set the threshold on the Gram matrix and remain the dependency only between the closest samples. For Email collection, the minimal generalization error is obtained when using 50-connected graph with the model complexity of 3 clusters. In this example, since the data categories are overlapping, the smaller models are not favored as it was in the case of well separated data as Rings and Gaussian data sets. In the right plot of the figure 8 the confusion matrix[8] is presented. With respect to the labels it can be concluded that the *spam* emails are well separated (99.5%) and the overlapping between

[8]The confusion matrix contains information about actual and predicted classification done by the classification system.

the *conference* and *job* emails is only slightly larger. In general, the data is well classified.



	Comp.	Motor	Baseb.	Christ.
1	12.9	95.1	6.4	4.0
2	1.2	2.0	0.9	91.1
3	83.5	1.0	3.6	5.0
4	2.4	2.0	89.1	0.0

Fig. 9. Left upper panel presents the mean generalization error as a function of both the cluster number and the k - cutoff threshold in the k-connected graph for Newsgroups collection. The error for selected K (10 20 30 40) is shown on the right plot. The optimal model complexity is 4 clusters when using 20-connected graph. Lower figure presents the confusion matrix for labeling produced by the selected optimal model and the original labeling. Only the small confusion can be observed.

The generalization error for the Newsgroups collection is shown in the left plot of figure 9. For the training, 400 samples randomly selected from the set was used and the rest of the collection (398 examples) was designated for generalization error. 40 experiments was performed and figure 9 displays the mean value of the generalization error. The optimum model has 4 clusters in model using 20-connected graph, even though the differences around the minimum are small compered to the maximum values of the investigated generalization error. In the right plot of the figure 8 the confusion matrix is presented. With respect to the labels it can be concluded that the data is well separated and classified. The data points are, however, more confused than in the case od email collection. In average 10% of each cluster is misclassified.

In order to perform the comparison of the aggregated Markov model with the classical spectral clustering method as presented in [1] another experiment was performed, the results of which are not presented in this paper due to space constraints. For both, continuous and discrete data sets, using both the Gaussian kernel and inner-product an investigation of the overall performance in classification[9] was made. It was found, that both aggregated Markov model and the spectral clustering model for selected model parameters did equally well in the sense of miss-classification error. However, the spectral clustering model was less sensitive to the choice of smoothing parameter h.

IV. Discussion

The aggregated Markov model provides a probabilistic clustering and the generalization error formula can be derived leading to the possibility of selecting model order and parameters. These virtues were not offered by the classical spectral clustering methods like [1] and [15].

In the case of continuous data, it can be noted that the quality of the clustering is directly related to the quality of the density estimate. Once a density has been estimated the proposed clustering method attempts to find modes in the density. Also if the density is poorly estimated due to perhaps a window smoothing parameter which is too large then class structure may be over-smoothed and so modes may be lost, in other words essential class structure may not be identified by the clustering. The same argument applies to a smoothing parameter which is too small thus causing non-existent structure to be discovered. The same argument can be made for the connectedness of the underlying graph connecting the points under consideration.

The disadvantage of the proposed model, in comparison with considered classical spectral clustering methods, is the computational complexity which is larger. As the vectors initializing the Gram matrix decomposition the eigenvectors are used, what ensures faster convergence and better decomposition outcome. It is, however, not necessary, in the case of well separated, simple data sets.

Appendix

In case of the presented model the minimal values of the generalization error are observed for all the model complexities smaller or equal the correct complexity. It is noticed only in the case of well separated clusters which is the case of the presented examples. When perfect (0/1 valued) cluster posterior probability $p(c|\mathbf{z}_i)$ is observed, the probability of the sample $p(\mathbf{z}_i)$ is similar for both smaller and larger models. It is true, as long as the natural cluster separations are not split, i.e. as long as the sample has large (close to 1) probability of belonging to one of the clusters $p(c|\mathbf{z}_i) \approx 1$. As an example lets consider the structure of 3 linear separable clusters. The generalization error 14 depends on the out-of-sample kernel function $\mathcal{K}(\mathbf{z}|\mathbf{x}_l)$, which is constant for various values of the model parameter c and the result of the Gram matrix decomposition $P(\mathbf{x}_l|c)P(c|\mathbf{x}_n)$. Therefore, the level of the generalization error as a function of model complexity parameter c depends only on the result of the Gram matrix decomposition. In the presented case, for the correct, 3 cluster, scenario the class posterior takes the binary 0/1 values. When smaller number of clusters are considered, the out-of-sample class posterior values are still binary as in the presented model it is enough that the out-of-sample is close to any of the training samples in the clusters and not to all of them. For more complex models the class posterior is no longer binary, since the natural cluster structure is broken, i.e./ at least two clusters are placed close to each other and the point assignment is ambiguous. Therefore, the generalization error values are increased.

[9] measured by the miss-classification error

References

[1] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *In Advances in Neural Information Processing Systems*, vol. 14, 2001, pp. 849–856.

[2] F. R. Bach and M. I. Jordan, "Learning spectral clustering," in *Advances in Neural Information Processing Systems*, 2003.
[3] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," CS Department, Yale University, Tech. Rep., 2000. [Online]. Available: citeseer.nj.nec.com/495691.html
[4] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.
[5] C. William and M. Seeger, "Using the nystrom method to speed up kernel machines," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000, pp. 682–688.
[6] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, no. 3, pp. 669 – 688, 2002.
[7] Y. Bengio, P. Vincent, and J.-F. Paiement, "Learning eigenfuncions of similarity : Linking spectral clustering and kernel pca," Dpartement d'informatique et recherche oprationnelle, Universit de Montral, Tech. Rep., 2003.
[8] I. A.J, "Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, pp. 205–224, 1991.
[9] B. Silverman, "Density estimation for statistics and data analysis," *Monographs on Statistics and Applied Probability*, 1986.
[10] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, pp. 568–576, 1991.
[11] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," in *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, C. Cardie and R. Weischedel, Eds. Somerset, New Jersey: Association for Computational Linguistics, 1997, pp. 81–89.
[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.
[13] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts," in *Advances in Neural Information Processing Systems*, 2003.
[14] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, August 1999, pp. 50–57.
[15] M. Meila and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems*, 2000, pp. 873–879.
[16] J. Larsen, L. Hansen, A. Szymkowiak-Have, T. Christiansen, and T. Kolenda, "Webmining: Learning from the world wide web," *special issue of Computational Statistics and Data Analysis*, vol. 38, pp. 517–532, 2002.
[17] J. Larsen, A. Szymkowiak-Have, and L. Hansen, "Probabilistic hierarchical clustering with labeled and unlabeled data," *International Journal of Knowledge-Based Intelligent Engineering Systems*, vol. 6(1), pp. 56–62, 2002.
[18] A. Szymkowiak, J. Larsen, and L. Hansen, "Hierarchical clustering for datamining," in *Proceedings of KES-2001 Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies*, 2001, pp. 261–265.