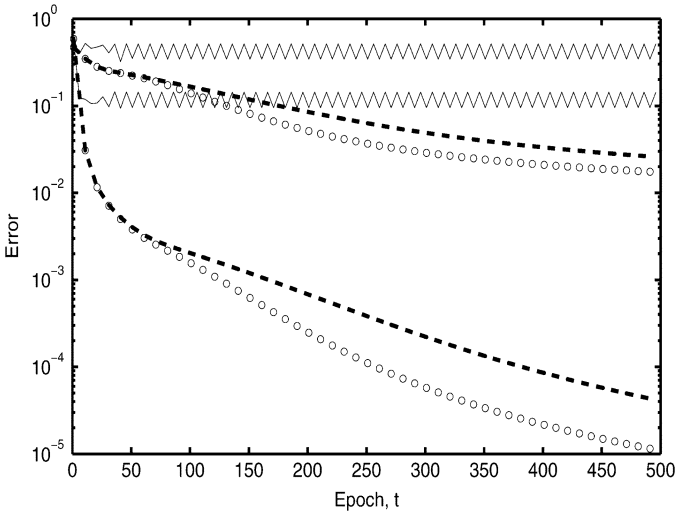


Fig. 3. The learning curves for the "Nonnegative PCA" algorithm. The dashed, circle and solid curves are drawn for the algorithm with constant learning rate η_1, adaptive learning rate $\eta(k)$ and constant learning rate η_2 respectively, showing nonnegative reconstruction error (lower curve), and crosstalk error (upper curve).

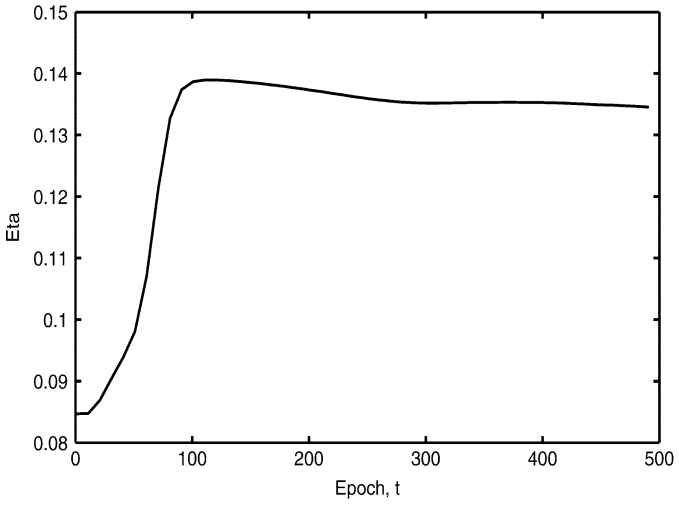


Fig. 4. The curve of the adaptive learning rate.

V. Conclusion

We have derived convergence conditions on the learning rate and the initial weight vector of the discrete-time "Nonnegative PCA" algorithm developed in [3]. A rigorous mathematical proof is given which does not change the discrete-time algorithms to the corresponding differential equations. The convergence condition on the learning rate helps alleviate the guesswork that accompanies the problem of choosing suitable learning rate in practical computation. Computer simulation results confirm our theoretical results and show the efficiency and effectiveness of our convergence theory.

References

[1] A. Cichocki and S.-I. Amar, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[3] E. Oja and M. D. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Comput.*, vol. 16, no. 9, pp. 1811–1825, 2004.

[4] E. Oja and M. D. Plumbley, "Blind separation of positive sources using nonnegative PCA," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, Nara, Japan, Apr. 2003, pp. 11–16.

[5] M. D. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 177–180, 2002.

[6] M. D. Plumbley and E. Oja, "A 'Nonnegative PCA' algorithm for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 66–76, 2004.

[7] M. D. Plumbley, "Lie group methods for optimization with orthogonality constraints," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, Sep. 2004, pp. 1245–1252.

[8] Z. Yi, K. K. Tan, and T. H. Lee, "Multistablity analysis for recurrent neural networks with unsaturating piecewise linear transfer functions," *Neural Comput.*, vol. 15, pp. 639–662, 2003.

[9] Q. Zhang, "On the discrete time dynamics of a PCA learning algorithm," *Neurocomputing*, vol. 55, pp. 761–769, 2003.

[10] P. J. Zuffiria, "On the discrete time dynamics of the basic Hebbian neural network node," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1342–1352, 2002.

Clustering via Kernel Decomposition

A. Szymkowiak-Have, Mark A. Girolami, and Jan Larsen

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

***Index Terms*—Aggregated Markov model, kernel decomposition, kernel principal component analysis (KPCA), spectral clustering.**

I. Introduction

The spectral clustering methods [1]–[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].

Manuscript received September 15, 2004; revised July 11, 2005.

A. Szymkowiak-Have and J. Larsen are with the Informatics and Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark (e-mail: asz@imm.dtu.dk; jl@imm.dtu.dk).

M. A. Girolami is with the Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K. (e-mail: girolami@dcs.gla.ac.uk).

Digital Object Identifier 10.1109/TNN.2005.860840

Building on this observation the relationship between KPCA and nonparametric 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 nonnegativity 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]–[3].

The following sections consider the nonparametric 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 $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x} \in \mathcal{R}^d$. The sample drawn from the density can be employed to estimate the density in a nonparametric form by using a Parzen window estimator (refer to [8]–[10] for a review of such nonparametric 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, 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

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

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

such that each sample point is equally probable *a priori*, $P(\mathbf{x}_n) = N^{-1}$, i.e., data are 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

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

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 $P(\mathbf{x}_n) = 1/N$.

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

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

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

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

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 $\hat{P}(\mathbf{x}_i|\mathbf{x}_j) = \sum_{c=1}^{C} P(\mathbf{x}_i|c)P(c|\mathbf{x}_j), \ \forall \quad i, j = 1, \ldots, N$ we see that this is identical to the aggregate Markov model originally proposed by Saul and Periera [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 distance based criterion such as squared error

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \left\{ \breve{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j) - \left\{ \sum_{c=1}^{C} P(\mathbf{x}_i|c)P(c|\mathbf{x}_j) \right\} \right\}^2 \tag{7}$$

can be locally optimized 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 publication which have proposed efficient methods for obtaining such constrained matrix decompositions [12], [13] and as such the nonnegative 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. The multiplicative methods detailed in [12] have been employed in the experiments reported in this paper. 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 $P(c|\mathbf{z})$.

$$\hat{P}(c|\mathbf{z}) = \sum_{n=1}^{N} P(c|\mathbf{x}_n)\hat{P}(\mathbf{x}_n|\mathbf{z}) \tag{8}$$

$$= \sum_{n=1}^{N} P(c|\mathbf{x}_n)\breve{\mathcal{K}}(\mathbf{z},\mathbf{x}_n) \tag{9}$$

$$= \sum_{n=1}^{N} P(c|\mathbf{x}_n)\frac{\mathcal{K}(\mathbf{z},\mathbf{x}_n)}{\sum_{n'=1}^{N}\mathcal{K}(\mathbf{z},\mathbf{x}_{n'})}. \tag{10}$$

This can be viewed as a form of "kernel"-based nonnegative 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 nonparametric 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 nonparametric 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*, e.g., number of classes, a generalization error[1] based on the out-of-sample negative predictive likelihood, is defined as follows:

$$\mathcal{L}_{out} = N_{out}^{-1}\sum_{n=1}^{N_{out}} \log\{p(\mathbf{z}_n)\} \tag{11}$$

where N_{out} denotes the number of out-of-sample points. The out-of-sample likelihood (11) is derived from the decomposition in the following manner

$$p(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} p(\mathbf{z}|\mathbf{x}_n) \tag{12}$$

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

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 follows:

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

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

$$\mathcal{K}_h(\mathbf{x},\mathbf{x}_n) = (2\pi)^{-d/2}h^{-d}\exp\left\{-\frac{1}{2h^2}\|\mathbf{x}-\mathbf{x}_n\|^2\right\}. \tag{15}$$

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.

$$\mathcal{K}(\mathbf{x},\mathbf{x}_n) = \frac{\mathbf{x}^T\mathbf{x}_n}{\|\mathbf{x}\|\cdot\|\mathbf{x}_n\|}. \tag{16}$$

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 nondensity 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 nonparametric 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[2] 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 are described here.

1) **Linear structure.** Data set consist of five two-dimensional (2-D) Gaussian distributed clusters with a spherical covariance structure, shown in the left plot of Fig. 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 Fig. 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-D. This is given as an example of complex nonlinear data on which methods such as K-means will fail.

[1]In this context the generalization error is defined as a negative predictive log-likelihood.

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

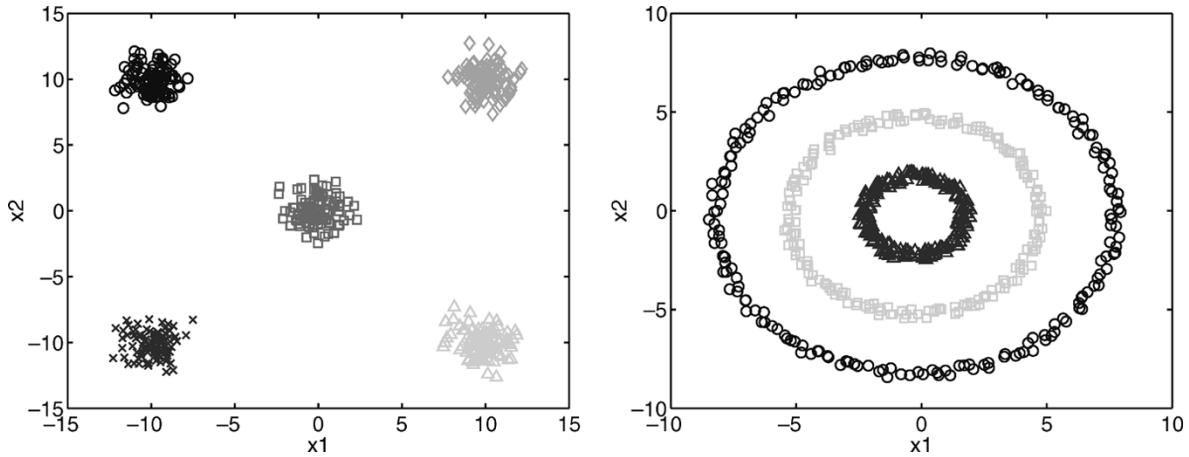


Fig. 1. The scatter plots of the artificial data for five Gaussian distributed clusters (left figure) and three cluster ring formations (right panel).

3) **Email collection.** The Email data set[3] consist of emails grouped into three categories: *conference, job,* and *spam*, used earlier in [16]–[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[4] and other high and low frequency words, stemming[5] 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[6] consist originally of 20 categories each containing approximately 1000 newsgroup documents. In the performed experiments, four 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 were stored in. The data were processed in a similar way as that presented above. In preprocessing, two documents were removed[7]. After preprocessing, the data consist 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 (15) is used. With discrete data sets (Emails and Newsgroups), the cosine inner-product (16) is applied.

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

[4]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

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

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

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

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 (11) and (13).

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 Fig. 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 Fig. 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 the Appendix .

For Gaussian clusters, the cluster posterior $p(c|\mathbf{z})$ is presented on Fig. 3. Perfect decision surfaces can be observed. For comparison, on Fig. 4, the components of the traditional kernel PCA are presented. Here, both the positive and the negative values are observed, which makes it difficult to determine the optimum decision surface.

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 Fig. 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 (Fig. 5, left plot) is equal $h = 0.065$ and the minimum in generalization error is obtained for three clusters. As in the Gaussian clusters example, a smaller model of two clusters is also probable[8].

The cluster posterior for Rings data set and the kernel PCA components are presented in Figs. 6 and 7, respectively. Also in this case perfect (0/1) decision surfaces are observed (Fig. 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.

The generalization error for Email collection is shown in left plot of Fig. 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

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

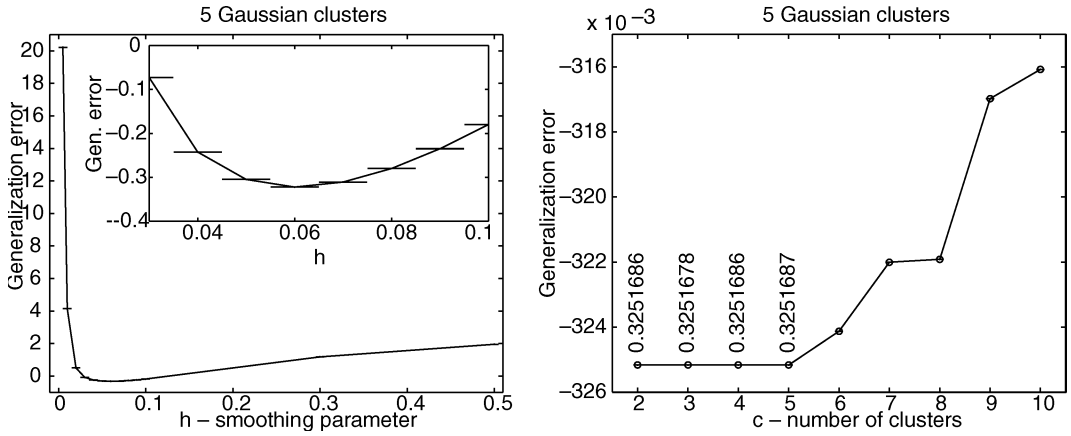


Fig. 2. The generalization error as a function of smoothing parameter h (left panel) for five Gaussian distributed clusters. The optimum choice is $h = 0.06$. The right figure presents, for optimum smoothing parameter, the generalization error as a function of number of clusters. Here, any cluster number 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.

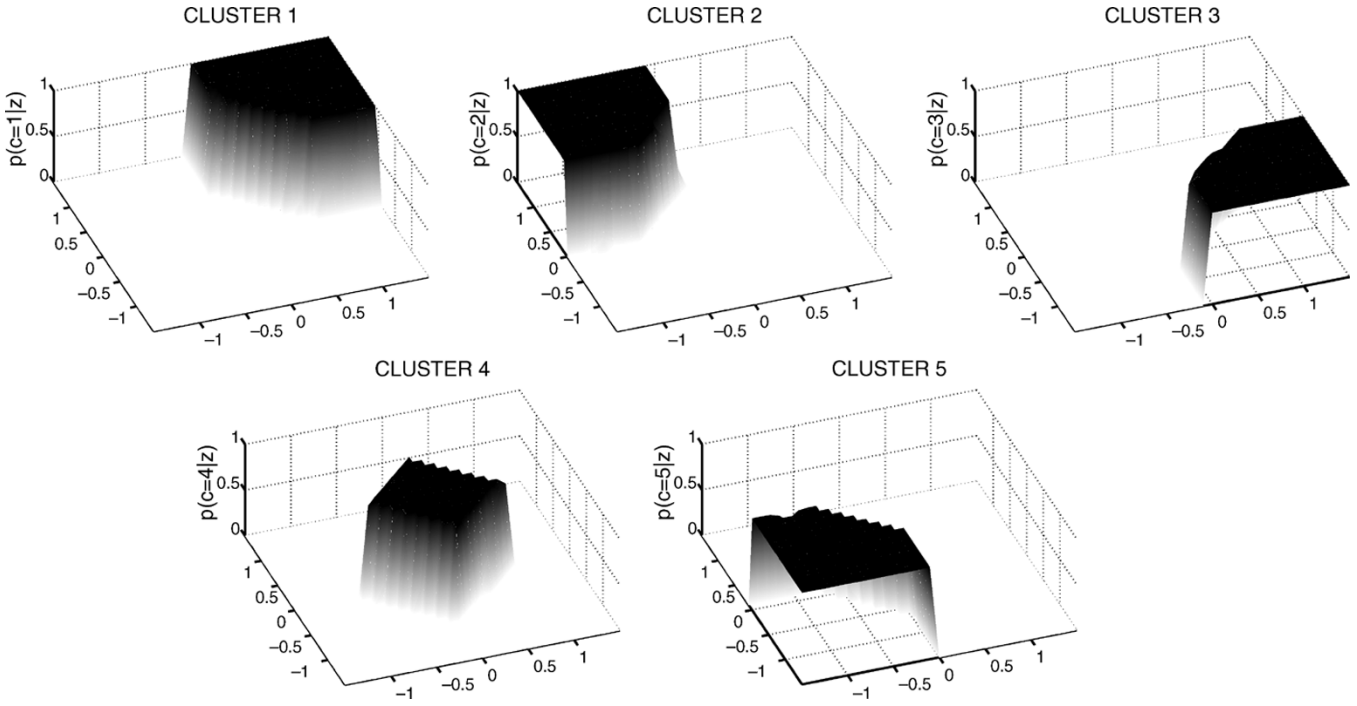


Fig. 3. The cluster posterior values $p(c|\mathbf{z})$ obtained from the aggregate Markov model for Gaussian clusters. The decision surfaces are positive. The separation in this case is perfect.

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 three 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 Fig. 8, the confusion matrix[9] is presented. With respect to the labels, it can be concludedthat the *spam* emails are well separated (99.5%) and the overlapping between the *conference* and *job* emails is only slightly larger. In general, the data is well classified.

The generalization error for the Newsgroups collection is shown in the left plot of Fig. 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. Forty experiments was performed and Fig. 9 displays the mean value of the generalization error. The optimum model has four clusters in model using 20-connected graph, even though the differences around the minimum are small compared to the maximum values of the investigated generalization error. In the right plot of the Fig. 8 the confusion matrix is presented. With respect to the labels, it can be concluded that the data are well separated and classified. The data points are, however, more confused than in the case of the 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 an discrete data sets, using both the Gaussian kernel and inner-product the inves-

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

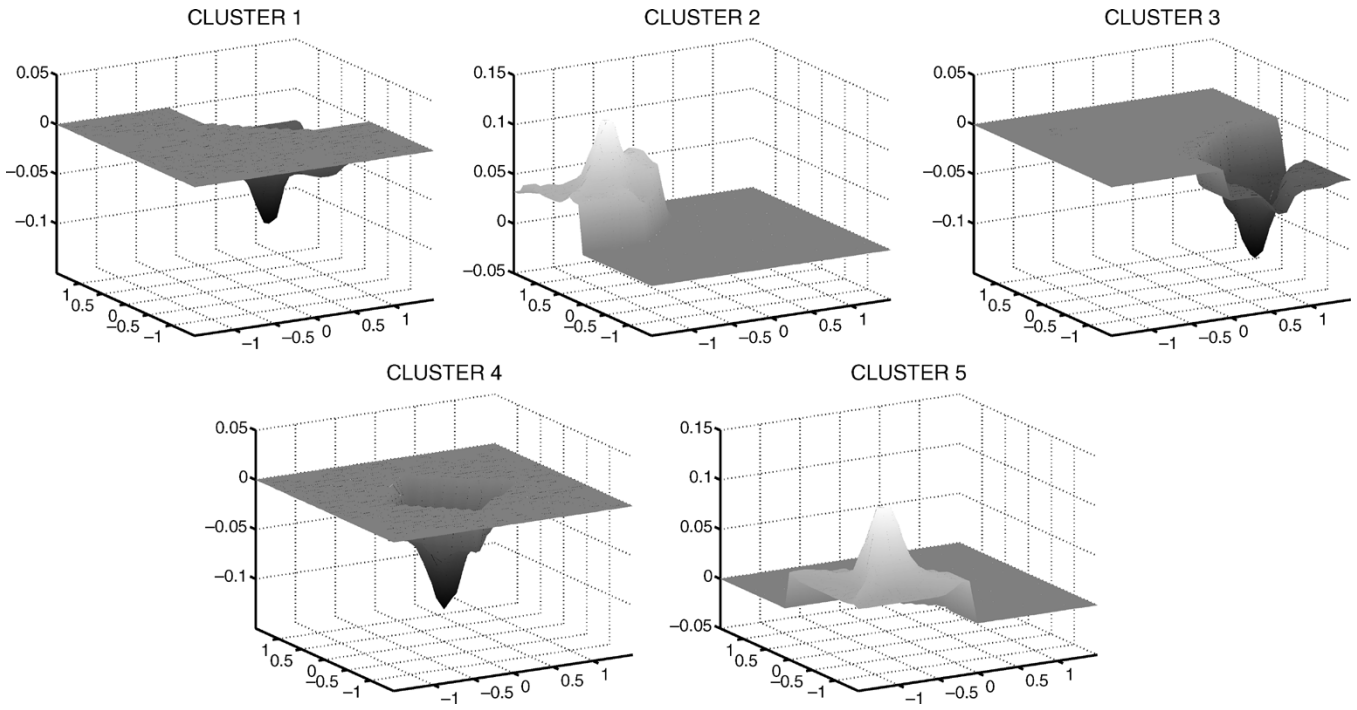


Fig. 4. The components of the traditional kernel PCA model for Gaussian clusters. The decision surfaces are both positive and negative.

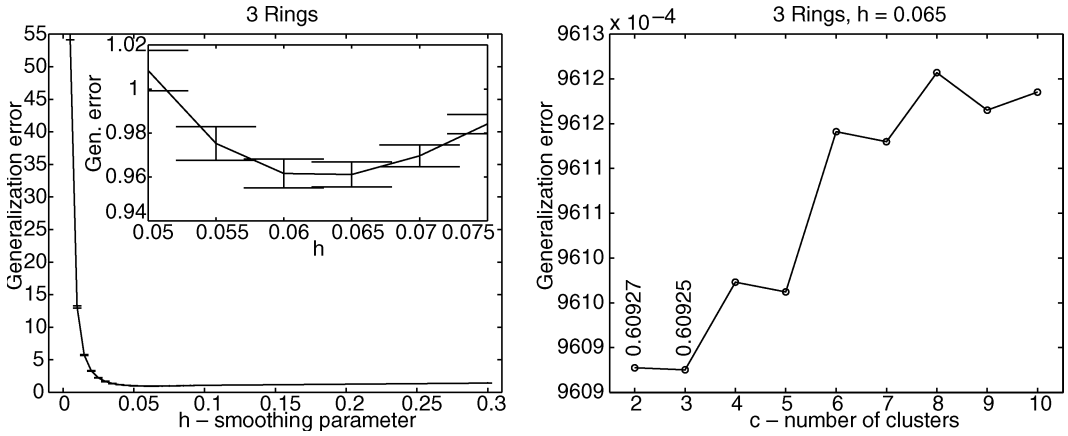


Fig. 5. The generalization error as a function of smoothing parameter h for three clusters formed in the shape of rings (left panel). The optimum choice is $h = 0.065$. On the right figure the generalization error as a function of number of clusters is shown for the optimum choice of smoothing parameter.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

tigation of the overall performance in classification[10] 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 misclassification error. However, the spectral clustering model was less sensitive to the choice of smoothing parameter h. The benefit of the proposed method is that a full set of posterior probabilities of cluster membership can be obtained.

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. The clustering is then a two-stage process, first, a nonparametric kernel density is obtained with the width parameter being selected using cross-validation. 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 oversmoothed 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 nonexistent 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 to other spectral clustering methods is that the computational complexity may be larger due to the nonlinear optimization required. In the experiments reported, the matrix decomposition is initialized with the Gram matrix eigenvectors which improved convergence speed and eventual solutions.

[10]measured by the miss-classification error

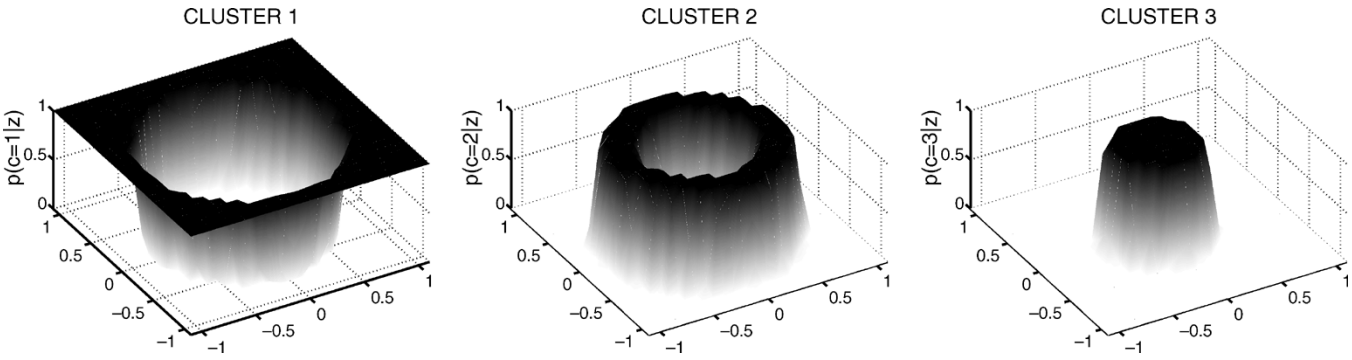


Fig. 6. The cluster posterior values $p(c|\mathbf{z})$ obtained from the aggregate Markov model for Rings. The decision surfaces are positive. The separation in this case is perfect.

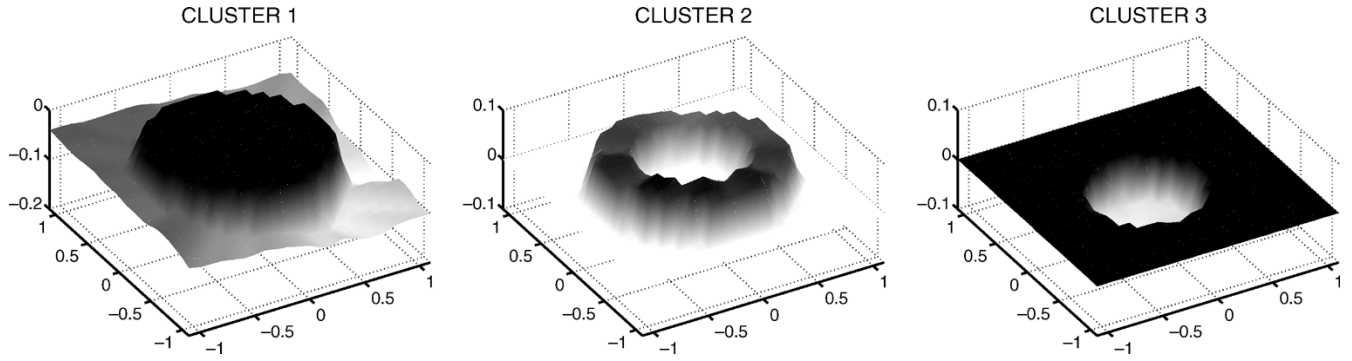
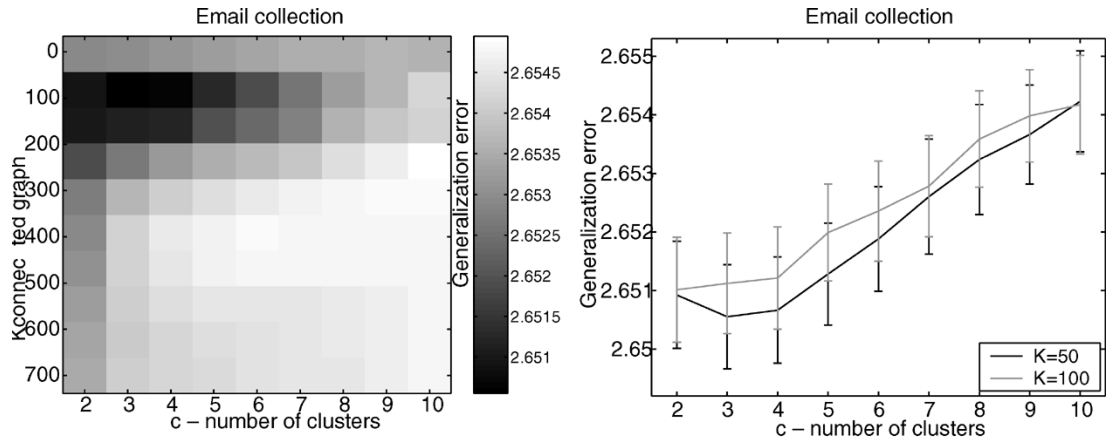
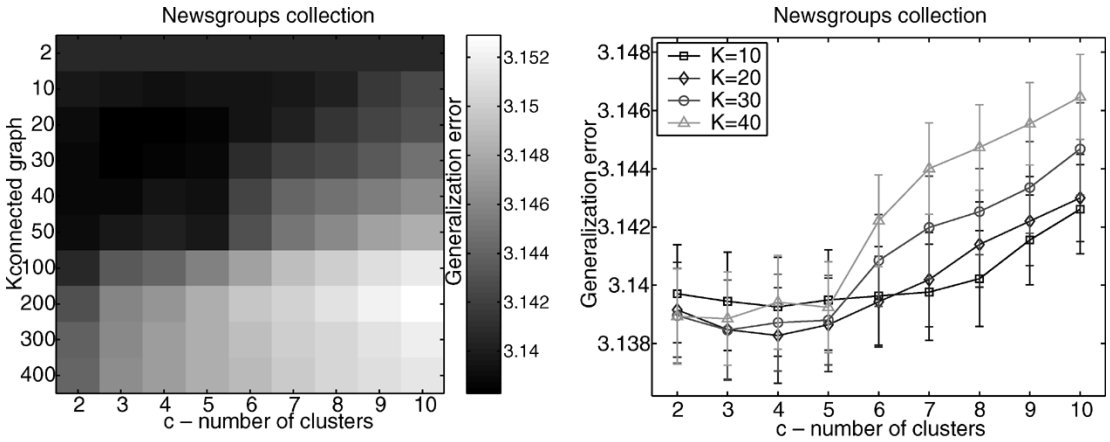


Fig. 7. The components of the traditional kernel PCA model for Rings structure. The decision surfaces are both positive and negative.



	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 ($\mathrm{K} = 50$ and $\mathrm{K} = 100$) are shown on the right plot. The optimal model is the choice of $K = 50$ (50-connected graph) with three 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.



	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 four 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.

APPENDIX

In the 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, let us consider the structure of three linear separable clusters. The generalization error 13 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, three 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.

REFERENCES

[1] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, 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. (2000) "On clusterings: good, bad and spectral," Tech. Rep.. Yale Univ., CS Dep.. [Online]. Available: citeseer.nj.nec.com/495 691.html

[4] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, 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. Cambridge, MA: MIT Press, 2000, vol. 13, pp. 682–688.

[6] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Comput.*, 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," Tech. Rep.," Univ. Montréal, Dép. d'informatique et recherche opérationnelle, 2003.

[8] A. J. Izenman, "Recent developments in nonparametric density estimation," *J. Amer. Statist. Assoc.*, vol. 86, pp. 205–224, 1991.

[9] B. Silverman, "Density estimation for statistics and data analysis," *Monographs on Statist. Appl. Probabil.*, 1986.

[10] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, pp. 568–576, 1991.

[11] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," in *Proc. 2nd Conf. Empirical Methods in Natural Language Processing*, C. Cardie and R. Weischedel, Eds., Somerset, NJ, 1997, pp. 81–89.

[12] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 556–562.

[13] D. Donoho and V. Stodden, "When does nonnegative matrix factorization give a correct decomposition into parts," in *Advances in Neural Information Processing Systems*, 2003.

[14] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Ann. ACM Conf. Res. Develop. Information Retrieval*, Berkeley, CA, Aug. 1999, pp. 50–57.

[15] M. Meila and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, 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 Computat. Statist. Data Anal.*, vol. 38, pp. 517–532, 2002.
[17] J. Larsen, A. Szymkowiak-Have, and L. Hansen, "Probabilistic hierarchical clustering with labeled and unlabeled data," *Int. J. Knowledge-Based Intell. Eng. Syst.*, vol. 6, no. 1, pp. 56–62, 2002.
[18] A. Szymkowiak, J. Larsen, and L. Hansen, "Hierarchical clustering for datamining," in *Proc. KES-2001 5th Int. Conf. Knowledge-Based Intelligent Info. Eng. Syst. Allied Technol.*, 2001, pp. 261–265.

On Stability of Recurrent Neural Networks—An Approach From Volterra Integro-Differential Equations

Pingzhou Liu and Qing-Long Han

***Abstract*—The uniform asymptotic stability of recurrent neural networks (RNNs) with distributed delay is analyzed by comparing RNNs to linear Volterra integro-differential systems under Lipschitz continuity of activation functions. The stability criteria obtained have unified and extended many existing results on RNNs.**

***Index Terms*—Delay, recurrent neural networks (RNNs), stability, Volterra integro-differential systems.**

I. Introduction

In this letter, we consider the following recurrent neural network (RNN) model

$$\begin{cases} \frac{dx(t)}{dt} = -Dx(t) + B\int_a^t K(t-s)F(x(s))ds \\ x(s) = \phi(s), \quad x \in (a, t_0), \quad t_0 \geq 0 \end{cases} \tag{1}$$

which is the generalization of the most extensively studied model

$$\frac{dy}{dt} = -Dy + BF(y) + Du. \quad y(t_0) = y_0 \tag{2}$$

by introducing distributed delay and translating equilibrium to the origin, where $x, y \in R^n$ are the state vectors, $D = \mathrm{diag}(d_1, d_2, \ldots, d_n) \in R^{n\times n}$ is a constant diagonal matrix with $d_i > 0, B = [b_{ij}] \in R^{n\times n}$ is a constant connection weight matrix, $u \in R^n$ is a constant input vector, the delay kernel $K(\cdot) = [k_{ij}(\cdot)] \in L^1(R^+)$. $F(\cdot) = \mathrm{col}(f_1(\cdot), f_2(\cdot), \ldots, f_n(\cdot))$ is a vector-value activation function from R^n to R^n and is assumed to be of class $\mathcal{GL}$ or $\mathcal{L}$ and $F(0) = 0$. If $F \in \mathcal{GL}$, then there exist ℓ_i, such that $\forall x, y \in R$ and $x \neq y$

$$0 \leq \frac{f_i(x) - f_i(y)}{x - y} \leq \ell_i, \quad i = 1, 2, \ldots, n. \tag{3}$$

If $F \in \mathcal{L}$, then there exist constants ℓ_i, such that $\forall x, y \in R$

$$|f_i(x) - f_i(y)| \leq \ell_i |x - y|, \quad i = 1, 2, \ldots, n. \tag{4}$$

In the following we only consider the case of $a = 0$ or $a = -\infty$. If $a = 0$, the system has finite memory and one needs to deal with "uniform" stability. If $a = -\infty$, the system has infinite memory. In practice situations, the distant past usually has less influence compared to the recent behavior of the state. The case of $a = -\infty$, which has drawn the most concern in the neural networks research, is just a mathematical simplification.

In this letter, we will use the well-known results about linear Volterra integro-differential equations and the nonlinearity nature of Lipschitz continuity (3) or (4), to study the global uniform asymptotic stability of (1). As we put different kinds of delays under one umbrella—distributed delay, it also provides a way to approximately consider the delay dependency of the neural networks by choosing an appropriate and easy-to-handle delay kernel.

II. Volterra Integro-Differential Equations

Let $C(a, \infty)(a = 0 \text{ or } a = -\infty)$ denote the set of all continuous functions φ: $(a, \infty) \to R^n$ such that, for any $t \in R^1$, the semi-norm

$$\|\varphi\|_t = \sup\{|\varphi(s)| : a < s \leq t\}$$

is finite. Let $B(\cdot) \in L^1(R^+)$ be a real matrix function.

Consider the following linear Volterra integro-differential equation [9], [10]

$$\frac{dx(t)}{dt} = Ax(t) + \int_a^t B(t-s)x(s)\,ds \tag{5}$$

for $t \geq t_0$ with $x(t) = \varphi(t)$, where $x(t) = \mathrm{col}(x_1(t), x_2(t), \ldots, x_n(t)) \in R^n$ with Euclidean norm $|x| = (\sum_{i=1}^n x_i^2)^{1/2}$ and A is a real constant matrix, $t \geq t_0$ and $x(t) = \phi(t)$ on $a \leq t \leq t_0$. The solution of (5) with initial values (t_0, ϕ) will be denoted by $x(t, t_0, \phi)$.

Notice that for $a = -\infty$, (5) is autonomous. It follows that one needs only to consider the case of $a = -\infty$ with initial time $t_0 = 0$. Moreover, stability and uniform stability are equivalent.

Let

$$\beta_{ij} = \int_0^\infty b_{ij}(t)dt, \quad \beta_{ij}^+ = \int_0^\infty |b_{ij}(t)|dt \quad \text{and}$$
$$R_i = \sum_{j=1}^n (|a_{ij}| + {\beta_{ij}}^+).$$

Theorem 1: [9]: Let $B(\cdot) \in L^1$ and

$$|a_{ii} + \beta_{ii}||a_{kk} + \beta_{kk}| > \left(\sum_{j\neq i}|a_{ij} + \beta_{ij}|\right)\left(\sum_{j\neq k}|a_{kj} + \beta_{kj}|\right) \tag{6}$$

Manuscript received March 21, 2004; revised July 25, 2005. This work was supported in part by the Faculty of Informatics and Communication, Central Queensland University, for the 2003–2004 Research Project "A Volterra integro-differential equation approach to stability for a class of recurrent neural networks."

P. Liu is with the Faculty of Informatics and Communication, Central Queensland University, Rockhampton, Australia (e-mail: q.han@cqu.edu.au), on leave from Flinders University, South Australia, Australia, and Shanxi Normal University, China.

Q.-L. Han is with the Faculty of Informatics and Communication, Central Queensland University, Rockhampton, Australia.

Digital Object Identifier 10.1109/TNN.2005.860859