Hierarchical Kernel Spectral Clustering

Carlos Alzate and Johan A. K. Suykens

Department of Electrical Engineering ESAT-SCD-SISTA,
Katholieke Universiteit Leuven,
B-3001 Leuven, Belgium

Abstract

A methodology to reveal the underlying cluster hierarchy in kernel spectral clustering problems is presented. The formulation fits in a constrained optimization framework where the primal problem is expressed in terms of high-dimensional feature maps and the dual problem is expressed in terms of kernel evaluations. Kernel spectral clustering can be seen a weighted form of kernel PCA where projections onto the eigenvectors constitute the clustering model. The kernel function acts here as the pairwise similarity function in classical spectral clustering. The clustering model can be evaluated at any data point allowing out-of-sample extensions which are useful for model selection in a learning setting. The parameters of the clustering problem can be found by optimizing the Fisher criterion on estimated eigenvectors for out-of-sample data. These eigenvectors display a localized structure when the clusters are well-formed. The Fisher criterion can also be used to reveal the hierarchical structure of the data. Simulations with toy data and image segmentation problems show the benefits of the proposed approach.

Keywords: hierarchical clustering, spectral clustering, kernel methods, out-of-sample extensions

1. Introduction

Spectral clustering methods have been successfully applied on a variety of applications. These methods correspond to family of unsupervised learning algorithms to find groups of data points that are similar (Shi and Malik, 2000; Ng et al., 2002; von Luxburg, 2007). The clustering information can be obtained from the eigenvectors with large eigenvalue of a matrix derived from pairwise similarities of the data points. These eigenvectors become a new representation of the data, where the clusters form a localized structure. Finding the final grouping from the eigenvectors is typically done by applying simple clustering techniques such as *k*-means. However, other specialized methods exists (Ng et al., 2002; Ding and He, 2004; Bach and Jordan, 2006). Applying *k*-means over the eigenvectors generally works because of the localized structure of the eigenspace. In general, the results obtained by classical spectral clustering formulations are only for the training data, i.e., the data points used to build the similarity matrix. Extending the cluster membership to unseen points (also called out-of-sample points) is unclear in this case, although approximations such as the Nyström method have been discussed (Fowlkes et al., 2004; Zhang et al., 2008). Another issue with classical spectral clustering is related to the determination of the similarity function and its parameters. The clustering parameters are typically obtained using heuristics, due mainly to the lack of underlying model.

Another view on spectral clustering from an optimization perspective has been discussed in (Alzate and Suykens, 2010). The formulation called kernel spectral clustering can be seen as a weighted version of kernel PCA set in a constrained optimization setting. The formulation has primal and dual model representations typical of least squares support vector machines (LS-SVM) (Suykens et al., 2002). By having a clustering model, it is possible to predict the cluster membership of out-of-sample points leading to model selection in a learning framework, good generalization capabilities and reduced computation times. The clustering model is expressed as projections which are linear combinations of kernel evaluations with eigenvectors as the coefficients.

The clustering results obtained using the eigenvectors give a flat description of the data where the clusters are disjoint. In some applications, a more informative hierarchical representation of the data is desirable (Clauset et al., 2008). This is typically the case when

the data contain different natural groupings depending of the scale. In this paper, we propose a methodology to represent the results of kernel spectral clustering in a hierarchical way if the data are found to contain such structure. This hierarchical representation is based on the Fisher criterion on validation data which quantify the goodness of clustering. The methodology is based on the estimation of eigenvectors for out-of-sample data proposed in (Alzate and Suykens, 2011). The eigenvectors obtained during training and the estimated out-of-sample eigenvectors display a special structure in cases with few cluster overlap and well-chosen similarity parameters: data points in the same cluster appear very localized in the eigenvector space. The main idea is to find clustering parameters such that the Fisher criterion is high, indicating strong cluster structures. With these optimal parameters and the corresponding clustering results, a dendrogram is built showing the underlying hierarchy.

This paper is organized as follows: Section 2 summarizes kernel spectral clustering. Section 3 contains a review of the out-of-sample eigenvector method introduced in (Alzate and Suykens, 2011). In Section 4, we describe the model selection criterion based on the Fisher criterion. Section 5 contains the proposed hierarchical representation together with 2 algorithms. In Section 6, we present the experimental results and in Section 7 we give conclusions.

2. Kernel Spectral Clustering

The kernel spectral clustering framework introduced in (Alzate and Suykens, 2010) puts spectral clustering in a constrained optimization setting with primal and dual model representations. The primal problem is formulated in terms of mappings to a high-dimensional feature space typical of support vector machine formulations. The dual problem is expressed as an eigenvalue decomposition of a matrix related to the random walks Laplacian. Thus, the clustering model can be expressed in the primal and in the dual space allowing extensions to unseen points (also called out-of-sample extensions). These extensions can be used for predictive purposes, model selection and reducing the computation times by appropriate subsampling schemes. The kernel spectral clustering framework is summarized as follows.

2.1. Primal and Dual formulations

Consider a set of training data points $\mathcal{D} = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, the objective of clustering is to find a partitioning of the dataset into $k > 1$ groups $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ such that data points assigned to the same cluster are more similar to each other than data points assigned to different clusters. The clustering model in the primal is formulated as a set of n_e projections of the training data points into a high-dimensional feature space $\mathcal{F}$ of dimension d_h:

$$e_i^{(l)} = w^{(l)^T}\varphi(x_i) + b_l, i = 1, \ldots, N, l = 1, \ldots, n_e, \quad (1)$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ is the mapping to $\mathcal{F}$ and b_l are the bias terms. The projections $e_i^{(l)}$ can be written in vector form as $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]^T$. Each $e^{(l)}$ represents the latent variable of a binary clustering given by $\text{sign}(e^{(l)})$. The set of n_e binary cluster decisions are then combined in a later stage into the final k groups. The primal problem is formulated as:

$$\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2}\sum_{l=1}^{n_e} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{n_e} \gamma_l e^{(l)^T} V e^{(l)} \quad (2)$$
$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N, l = 1, \ldots, n_e,$$

where $\gamma_l \in \mathbb{R}^+$ are regularization constants, $\Phi = [\varphi(x_1)^T, \ldots, \varphi(x_N)^T]$ is the $N \times d_h$ training feature matrix and $V = \text{diag}([v_1, \ldots, v_N])$, $v_i \in \mathbb{R}^+$ are user-defined weights. This optimization problem can be interpreted as a weighted version of kernel PCA since we would like to maximize a weighted L_2 loss function of the projected variables $e^{(l)}$ while trying to keep the norm of the primal projection vectors $w^{(l)}$ small. The dual problem is formalized in the following Lemma.

Lemma 1. *(Alzate and Suykens, 2010) Given a training dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, a positive definite diagonal weighting matrix V and a positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the Karush-Kuhn-Tucker (KKT) conditions of the Lagrangian of (2) give the eigenvalue problem:*

$$VM_V\Omega\alpha^{(l)} = \lambda_l\alpha^{(l)}, l = 1, \ldots, n_e \quad (3)$$

where $M_V = I_N - 1_N 1_N^T V/(1_N^T V 1_N)$ is a weighted centering matrix, I_N is the $N \times N$ identity matrix, 1_N is an N-dimensional vector of ones, $\Omega \in \mathbb{R}^{N\times N}$ is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j)$, $\lambda_l = N/\gamma_l$ are the ordered eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{n_e}$ and $\alpha^{(l)}$ are the corresponding eigenvectors.

Proof. The Lagrangian of (2) is

$$\mathcal{L}(w^{(l)}, e^{(l)}, b_l; \alpha^{(l)}) = \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1}\gamma_l e^{(l)^T} V e^{(l)} + \sum_{l=1}^{k-1}\alpha^{(l)^T}(e^{(l)} - \Phi w^{(l)} - b_l 1_N)$$

with Karush-Kuhn-Tucker (KKT) optimality conditions given by:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \rightarrow w^{(l)} = \Phi^T \alpha^{(l)} \\ \dfrac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \rightarrow \alpha^{(l)} = \dfrac{\gamma_l}{N} V e^{(l)} \\ \dfrac{\partial \mathcal{L}}{\partial b_l} = 0 \rightarrow 1_N^T \alpha^{(l)} = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \rightarrow e^{(l)} = \Phi w^{(l)} + b_l 1_N, \end{cases} \tag{4}$$

eliminating the primal variables $w^{(l)}$ and expressing the KKTs in terms of the dual variables $\alpha^{(l)}$ leads to the eigenvalue problem (3). □

Note that the regularization constants γ_l are related to the eigenvalues λ_l. This means that the value of γ_l is determined automatically by the solutions of the eigenvalue problem (3). Using the KKT optimality conditions, we can express the clustering model in terms of the dual variables $\alpha^{(l)}$:

$$e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_i, x_j) + b_l, \tag{5}$$

where the bias terms b_l are equal to:

$$b_l = -\frac{1}{1_N^T D^{-1} 1_N} 1_N^T D^{-1} \Omega \alpha^{(l)}, l = 1, \ldots, n_e. \tag{6}$$

2.2. Choosing V and n_e

The user-defined weighting matrix V and the number of additional projected variables n_e still need to be defined in order to obtain eigenvectors with discriminative clustering information. If $V = D^{-1}$ where $D = \text{diag}([d_1, \ldots, d_N]), d_i = \sum_{j=1} \Omega_{ij}$ is the so-called degree of x_i , then the dual problem (3) becomes:

$$D^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, l = 1, \ldots, n_e \tag{7}$$

where $M_D = I_N - 1_N 1_N^T D^{-1} / (1_N^T D^{-1} 1_N)$. In this case, the resulting matrix $D^{-1} M_D \Omega$ has spectral properties that are useful for clustering which are explained as follows. Consider first the random walks Laplacian $P = D^{-1} S$ where S is an $N \times N$ similarity matrix with ij-th entry $S_{ij} = s(x_i, x_j) \geq 0, i, j = 1, \ldots, N$. The spectral properties of P have been extensively studied in the context of clustering and graph partitioning (Chung, 1997; Meila and Shi, 2001; von Luxburg, 2007). The data points in $\mathcal{D}$ are represented as the nodes of an undirected graph. Similarities are computed for every pair of nodes in the graph and are represented as edge weights. The whole graph is entirely defined by its similarity (also called affinity) matrix S which is non-negative. If the graph contains $k > 1$ subgraphs with high intra-cluster similarity and low inter-cluster similarity, then there are k eigenvectors of P with eigenvalue close to 1 and they display a special structure that can be used to find the underlying grouping. When the subgraphs inter-similarity is equal to zero, these k eigenvectors are *piecewise constant* on the partitioning and the corresponding eigenvalues are equal to 1. One of these eigenvectors is equal to 1_N. In other words, data points belonging to the same cluster are represented with exactly the same value in the eigenvectors. This property can be interpreted as a transformation from the original input space to a k-dimensional space spanned by the eigenvectors where the clusters are more evident and easier to identify (Meila and Shi, 2001; Deuflhard et al., 2000).

Now, we establish the link between the spectral properties of P and the spectral properties of $D^{-1} M_D \Omega$. The kernel matrix Ω acts here as the similarity matrix S and the kernel function $K(x, z)$ as the similarity function $s(x, z)$. We restrict ourselves to kernel functions leading to non-negative values (e.g., the well-known RBF kernel). Let us assume that $\mathcal{D}$ contains k clusters, the inter-cluster similarity is zero and that $\alpha^{(l)}$ are eigenvectors of P. Expanding (7) leads to:

$$D^{-1} \Omega \alpha^{(l)} - \frac{1}{1_N^T D^{-1} 1_N} D^{-1} 1_N 1_N^T D^{-1} \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}$$

which can be further simplified to:

$$\alpha^{(l)} - \frac{1}{1_N^T D^{-1} 1_N} D^{-1} 1_N 1_N^T \alpha^{(l)} = \lambda_l \alpha^{(l)}, l = 1, \ldots, n_e.$$

Thus, a piecewise constant eigenvector $\alpha^{(l)}$ of P is also an eigenvector of $D^{-1} M_D \Omega$ with eigenvalue 1 if it has zero mean. Due to one of the KKTs, all eigenvectors of $D^{-1} M_D \Omega$ have zero mean. Combining these two results leads to the link between the spectrum of P and the spectrum of $D^{-1} M_D \Omega$: the latter matrix has $k - 1$ piecewise constant eigenvectors[1] with zero mean corresponding to the eigenvalue 1. With this in mind, we set $n_e = k - 1$.

2.3. From Eigenvectors and Projections to Clusters

After the piecewise constant eigenvectors $\alpha^{(l)}$ and the projections $e^{(l)}$ have been computed, the set of $k - 1$ binary cluster decision vectors given by $\text{sign}(e^{(l)})$ can

[1] Note that, 1_N is an eigenvector of P but not of $D^{-1} M_D \Omega$, since it does not fulfill the third KKT condition in (4).

be combined into the final k clusters. Due to the fact that the eigenvectors have zero mean and the projections have the same sign pattern of the eigenvectors, data points in the same cluster will be mapped to the same orthant in the $(k-1)$-dimensional projection space.

Then, each training data point x_i has an associated binary encoding vector given by $\text{sign}([e_i^{(1)}, \ldots, e_i^{(k-1)}]^T), i = 1, \ldots, N$. Data points in the same cluster will have the same binary encoding vector and in the same way, different clusters will have a different encoding vector. A codebook can be formed during the training stage, containing the codewords that represent the different clusters. The cluster membership of the i-th data point can then be assigned by comparing its binary encoding vector with respect to all codewords in the codebook and selecting the cluster with minimal Hamming distance.

2.4. Approximately Piecewise Constant Eigenvectors

In practice, the inter-cluster similarities can be small but not exactly zero due to cluster overlap. This is also the case for commonly-used similarity functions such as the RBF kernel. From perturbation analysis (Kato, 1995; von Luxburg, 2007) we know that if the inter-cluster similarities are small and the eigengap $\lambda_k - \lambda_{k+1}$ is large, then the eigenvectors of P will be *approximately* piecewise constant and will still contain discriminative clustering information. Since decisions are taken depending on the sign pattern of the projections, the perturbation should be large enough to cause a change in the sign pattern. In this way, the cluster decisions are robust with respect to small perturbations of the projections due to the fact that the eigenvectors are approximately piecewise constant. Figure 1 shows an illustrative example of a clustering problem with two well-separated Gaussian clouds. The parameters of the kernel which acts as a similarity function have been chosen such that the inter-cluster similarities are very small (tending to zero) which leads to a piecewise constant eigenvector, and large enough to cause the eigenvector to be approximately piecewise constant. In both cases, the clustering results are the same since decisions are taken depending on the sign pattern of the projections which is the same as the eigenvectors.

2.5. Out-of-Sample Extensions

One of the main advantages of having a clustering model is the possibility to evaluate it at unseen data points. This out-of-sample extension becomes important for performing clustering in a learning setting ensuring good predictive capabilities, tuning the parameters of the model and reducing the computational burden

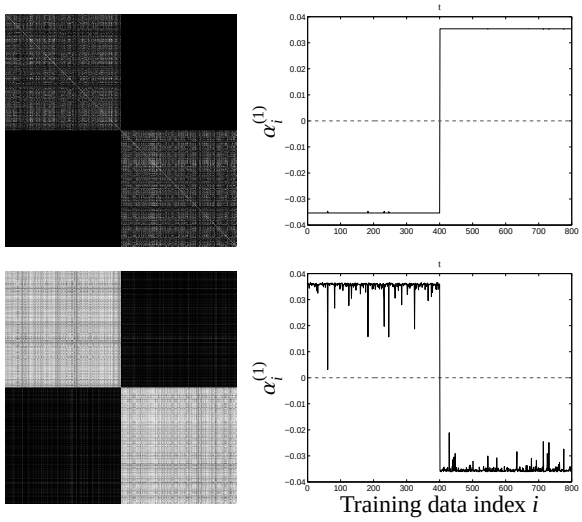

Figure 1: The top row shows the similarity (kernel) matrix with given kernel parameters such that the inter-cluster similarities are very small (maximum value 1.3×10^{-9}) and the resulting piecewise constant eigenvector. In the bottom row, the maximum value of the inter-cluster similarities is 0.76 causing the eigenvector to be approximately piecewise constant but still, leading to the same clustering results.

by appropriate subsampling. Given an arbitrary data point x, the latent variable of the clustering model becomes:

$$\hat{e}^{(l)}(x) = w^{(l)^T} \varphi(x) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_j, x) + b_l, \qquad (8)$$

$l = 1, \ldots, k-1$, with binary encoding vector given by $\text{sign}([e^{(1)}(x), \ldots, e^{(k-1)}(x)]^T)$. The cluster membership assignment goes in the same way as in the training stage: the encoding vector is compared with the codewords in the codebook and it is assigned to the cluster with minimal Hamming distance.

3. Out-of-Sample Eigenvectors

In classical spectral clustering, the cluster decisions are typically taken by applying k-means over the eigenvectors. This scheme works well in most cases because data points in the same cluster are mapped to approximately the same point in the eigenvector space inducing a very localized representation of the clusters. In the case of kernel spectral clustering, we use projections onto the eigenvectors to estimate the cluster membership. The structure of the projections when the eigenvectors are approximately piecewise constant leads to collinearity when two points are in the same cluster. This differs greatly from a localized representation. In

this Section, we summarize the method first introduced in (Alzate and Suykens, 2011) to estimate eigenvectors for out-of-sample data, allowing a localized representation of the clusters. This representation can be used for designing model selection criteria and to obtain a hierarchical view of the clusters as will be explain in the sequel.

3.1. Projections and Collinearity

When the eigenvectors are piecewise constant and the out-of-sample data have been drawn for the same probability distribution as the training dataset $\mathcal{D}$, the projections of these unseen data points onto the eigenvectors display a special structure. Consider the projections of an out-of-sample point x as given in (8) and assume that there are k clusters and that the eigenvectors $\alpha^{(l)}$ are piecewise constant on the partitioning. The piecewise constant property can be written as

$$\alpha_i^{(l)} = \alpha_j^{(l)} = c_p^{(l)}, \text{ if } x_i, x_j \in \mathcal{A}_p,$$

where $c_p^{(l)}$ is the constant value of the l-th eigenvector corresponding to the p-th cluster, $l = 1, \ldots k-1, p = 1, \ldots, k$. Then (8) can be written as:

$$\hat{e}^{(l)}(x) = \sum_{p'=1}^{k} c_{p'}^{(l)} \sum_{j \in \mathcal{A}_{p'}} K(x_j, x) + b_l,$$

now assume that x belongs to the p-th cluster leading to:

$$\hat{e}^{(l)}(x) = c_p^{(l)} \sum_{j \in \mathcal{A}_p} K(x_j, x) + \sum_{p' \neq p} c_{p'}^{(l)} \sum_{u \in \mathcal{A}_{p'}} K(x_u, x) + b_l,$$

perfectly piecewise constant eigenvectors imply that inter-cluster similarities are equal to zero thus, the second term in the previous equation vanishes and we obtain:

$$\hat{e}^{(l)}(x) = c_p^{(l)} \sum_{j \in \mathcal{A}_p} K(x_j, x) + b_l,$$

$l = 1, \ldots, k-1$, which describe the parametric equations of a line in a $(k-1)$-dimensional space. If the inter-cluster similarities are small but non-zero, the eigenvectors are approximately piecewise constant and data points in the same cluster are approximately collinear in the projections space. This structural property of the projections has been used for model selection in (Alzate and Suykens, 2010) for the design of a criterion measuring collinearity of out-of-sample data.

3.2. Eigenvectors for Out-of-Sample Data

The second KKT condition in (4):

$$\alpha_i^{(l)} = (1/\lambda_l) D_{ii}^{-1} e_i^{(l)}, i = 1, \ldots, N, l = 1, \ldots, k-1 \quad (9)$$

establishes the relation between the eigenvectors and the projections for training data points. Cluster decisions are taken on projections because eigenvectors exist only for training data but projections can be obtained for both training and out-of-sample data. However, by extending (9) to out-of-sample data, we can estimate eigenvectors for data that were not used during the training stage. Given a set of N_v validation data points $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_v}$ and the corresponding projections:

$$\hat{e}_{\text{val},t}^{(l)} = w^{(l)^T} \varphi(x_t^{\text{val}}) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_j, x_t^{\text{val}}) + b_l, \quad (10)$$

we can extend (9) to:

$$\hat{\alpha}_{\text{val},t}^{(l)} = \frac{1}{\lambda_l d(x_t^{\text{val}})} \Big(\sum_{j=1}^{N} \alpha_j^{(l)} K(x_j, x_t^{\text{val}}) + b_l \Big), \quad (11)$$

where $d(x) = \sum_{j=1}^{N} K(x_j, x)$ extends the concept of node degree to out-of-sample data. The mean is removed from the out-of-sample eigenvectors in order to fulfill the third KKT condition in (4) by:

$$\hat{\alpha}_{\text{val}}^{(l)} \leftarrow M_c \hat{\alpha}_{\text{val}}^{(l)} \quad (12)$$

where $\hat{\alpha}_{\text{val}}^{(l)} = [\hat{\alpha}_{\text{val},1}^{(l)}, \ldots, \hat{\alpha}_{\text{val},N_v}^{(l)}]$, $M_c = I_{N_v} - (1/N_v) 1_{N_v} 1_{N_v}^T$ is the centering matrix and normalized to have norm 1:

$$\hat{\alpha}_{\text{val}}^{(l)} \leftarrow \frac{\hat{\alpha}_{\text{val}}^{(l)}}{\|\hat{\alpha}_{\text{val}}^{(l)}\|_2}. \quad (13)$$

An advantage of the estimated *out-of-sample eigenvectors* $\hat{\alpha}_{\text{val}}^{(l)}$ is that they represent the clusters in a localized manner instead of the collinear structure seen in the projections.

4. Model Selection using the Fisher Criterion

Determining good values of the problem parameters is a critical issue in unsupervised learning methods such as clustering. Typically, the clustering parameters are obtained in a heuristic way, due mainly to the lack of a predictive model and the lack of understanding of the underlying process. However, since we have a clustering model with the possibility to predict the cluster membership of out-of-sample data, we can perform

model selection in a learning setting with training, validation and test stages. The representation offered by the out-of-sample eigenvectors when the data contain strong cluster structures allows us to tune the parameters of the model such that the clusters are represented as localized as possible. Measures of cluster distortion are popular for assessing the internal performance of clustering algorithms. The classical Fisher criterion (Bishop, 2006) used in classification can be adapted to select the number of clusters k and the parameters of the kernel (e.g., the bandwidth σ^2 of the RBF kernel) such that the out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}$ have a localized representation of the clusters. Given a validation dataset $\mathcal{D}^{\text{val}} = \{x^{\text{val}}_t\}_{t=1}^{N_{\text{v}}}$, the obtained out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}, l = 1, \ldots, k-1$ and the resulting cluster memberships $\{\mathcal{A}_p\}_{p=1}^{k}$, the *between-cluster* S_B and *within-cluster* S_W scatter matrices are defined by:

$$S_B = \sum_{p=1}^{k} \zeta_p (\hat{\mu}_p - \hat{\mu})(\hat{\mu}_p - \hat{\mu})^T \quad (14)$$

$$S_W = \sum_{p=1}^{k} \sum_{j \in \mathcal{A}_p} (\hat{\alpha}_{\text{val},j} - \hat{\mu}_p)(\hat{\alpha}_{\text{val},j} - \hat{\mu}_p)^T \quad (15)$$

where $\hat{\alpha}_{\text{val},t} = [\hat{\alpha}^{(1)}_{\text{val},t}, \ldots, \hat{\alpha}^{(k-1)}_{\text{val},t}] \in \mathbb{R}^{k-1}$ is the representation of $x^{\text{val}}_t, t = 1, \ldots, N_{\text{v}}$ in the out-of-sample eigenvector space,

$$\hat{\mu}_p = \frac{1}{|\mathcal{A}_p|} \sum_{j \in \mathcal{A}_p} \hat{\alpha}_{\text{val},j}$$

is the mean vector of the p-th cluster,

$$\hat{\mu} = \frac{1}{N_{\text{v}}} \sum_{t=1}^{N_{\text{v}}} \hat{\alpha}_{\text{val},t}$$

is the mean vector of all clusters. Note that, trace(S_B) is characterizing the weighted sum of the squared distances from each cluster centroid to the global centroid. The user-defined weights are given by ζ_p which can be chosen to give preference to certain scenarios (e.g., $\zeta_p = |\mathcal{A}_p|/N_{\text{v}}$ to give preference to balanced clusters (Alzate and Suykens, 2011)). However, hierarchical representations can lead to imbalanced clustering results thus, in this case, we consider an unweighted S_B, i.e., $\zeta_p = 1$. Trace(S_W) is an average measure of cluster compactness with value 0 when the out-of-sample eigenvectors are perfectly localized (piecewise constant). The Fisher criterion (Bishop, 2006) can then be defined as:

$$F(\hat{\alpha}^{(l)}_{\text{val}}, \{\mathcal{A}_p\}_{p=1}^{k}) = \frac{\text{trace}(S_B)}{\text{trace}(S_W + S_B)}. \quad (16)$$

This criterion is bounded between 0 and 1 taking its maximal value for well-separated and compact clusters in the out-of-sample eigenvector space.

5. Hierarchical Representation

Typically, the clusters found by classical clustering methods are disjoint which leads to a flat description of the data. However, in many cases, the clusters have subclusters which in turn might have substructures. A hierarchical visualization of the clusters supersedes the classical way the results of spectral clustering are presented. Rather than just reporting the cluster membership of all points, a hierarchical view provides a more informative description incorporating several scales in the analysis.

5.1. Classical Hierarchical Clustering

Traditionally, a hierarchical structure is displayed as a dendrogram where distances are computed for each pair of data points. The bottom-up approach consists of each data point starting as a cluster at the bottom of the tree. Pairs of clusters are then merged as the hierarchy goes up in the tree. Each merge is represented with a horizontal line and the y-axis indicate the similarity (or dissimilarity) of the two merging clusters. Determining which clusters should merge depends on a linkage measure which in turn specifies how dissimilar two clusters are. The linkage criterion takes into account the pairwise distances of the data points in the two sets. Commonly-used linkage criteria include:

- Minimum or single:

$$\text{linkage}(\mathcal{A}_1, \mathcal{A}_2) = \min_{x_i \in \mathcal{A}_1, x_j \in \mathcal{A}_2} \text{dist}(x_i, x_j)$$

- Maximum or complete:

$$\text{linkage}(\mathcal{A}_1, \mathcal{A}_2) = \max_{x_i \in \mathcal{A}_1, x_j \in \mathcal{A}_2} \text{dist}(x_i, x_j)$$

- Average:

$$\text{linkage}(\mathcal{A}_1, \mathcal{A}_2) = \frac{1}{|\mathcal{A}_1||\mathcal{A}_2|} \sum_{x_i \in \mathcal{A}_1} \sum_{x_j \in \mathcal{A}_2} \text{dist}(x_i, x_j)$$

- Ward:

$$\text{linkage}(\mathcal{A}_1, \mathcal{A}_2) = \left(\frac{2|\mathcal{A}_1||\mathcal{A}_2|}{|\mathcal{A}_1| + |\mathcal{A}_2|} \right)^{1/2} \|\hat{\mu}_1 - \hat{\mu}_2\|_2,$$

where dist is a distance measure, $|\mathcal{A}_p|$ is the size of the p-th cluster and $\hat{\mu}_p = (1/|\mathcal{A}_p|)\sum_{x_i \in \mathcal{A}_p} x_i$ is the centroid of the p-th cluster. Single linkage is a local criterion taking into account only the zone where the two clusters are closest to each other. This criterion suffers from an undesirable effect called chaining. Chaining causes unwanted elongated clusters since the overall shape of the formed clusters is not taken into account. Complete linkage is a non-local criterion giving preference to compact clusters but suffers from high sensitivity to outlying data points. Average and Ward's linkage criteria are specialized methods trying to find a compromise between single and complete linkage.

5.2. Hierarchical Kernel Spectral Clustering

In addition to obtaining good values for the number of clusters k and the kernel parameters (e.g., RBF kernel parameter σ^2), the Fisher criterion can also indicate a multiscale structure present on the data. A grid search over different values of k and a σ^2 evaluating the Fisher criterion on validation data in order to find clustering parameter pairs (k, σ^2) such that the criterion is greater than a specified threshold value. The clustering model is then trained for each (k, σ^2) and evaluated at the full dataset using the out-of-sample extension. This procedure results in cluster memberships of all points for each (k, σ^2). A specialized linkage criterion determines which clusters are merging based on the evolution of the cluster memberships as the hierarchy goes up. The y-axis of the dendrogram represents the scale and corresponds to the value of σ^2 at which the merge occurs. Note that during a merge, there might be some *outcast* data points of the merging clusters that go to a non-merging cluster. The ratio between the total size of the newly created cluster and the sum of the sizes of the merging clusters gives an indication of the merge quality. The outcast data points are then forced to join the merging cluster of the majority. The proposed methodology is outlined in the following algorithms.

Algorithm 1 Out-of-Sample Eigenvectors in Kernel Spectral Clustering

Input: Training set $\mathcal{D} = \{x_i\}_{i=1}^N$, positive definite and local kernel function $K(x_i, x_j)$, number of clusters k, validation set $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_v}$.
Output: Partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, cluster codeset C, out-of-sample eigenvectors $\hat{\alpha}_{\text{val}}^{(l)}$.

1: Obtain the eigenvectors $\alpha^{(l)}, l = 1, \ldots, k-1$ corresponding to the largest $k-1$ eigenvalues by solving (7) with $n_e = k-1$.
2: Binarize the eigenvectors: $\text{sign}(\alpha_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$, and let $\text{sign}(\alpha_i) \in \{-1, 1\}^{k-1}$ be the encoding vector for the training data point x_i, $i = 1, \ldots, N$.
3: Count the occurrences of the different encodings and find the k encodings with most occurrences. Let the codeset be formed by these k encodings: $C = \{c_p\}_{p=1}^k, c_p \in \{-1, 1\}^{k-1}$
4: $\forall i$, assign x_i to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\text{sign}(\alpha_i), c_p)$ and $d_H(\cdot, \cdot)$ is the Hamming distance.
5: Compute the projections for validation data $\hat{e}_{\text{val}}^{(l)}$ using (10).
6: Estimate the out-of-sample eigenvectors $\hat{\alpha}_{\text{val}}^{(l)}$ using $\hat{e}_{\text{val}}^{(l)}$ and (11). Remove the mean and normalize using (12) and (13).
7: Binarize the out-of-sample eigenvectors $\text{sign}(\hat{\alpha}_{\text{val},t}^{(l)}), t = 1, \ldots, N_v, l = 1, \ldots, k-1$ and let $\text{sign}(\hat{\alpha}_{\text{val},t}) \in \{-1, 1\}^{k-1}$ be the encoding vector of x_t^{val}.
8: $\forall t$, assign x_t^{val} to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\text{sign}(\hat{\alpha}_{\text{val},t}^{(l)}), c_p)$.

Algorithm 2 Hierarchical Kernel Spectral Clustering

Input: Full dataset $\mathcal{D}^{\text{full}} = \{x_f\}_{f=1}^{N_{\text{full}}}$, RBF kernel function with parameter σ^2, maximum number of clusters $k_{\max}$, set of R σ^2 values $\{\sigma_1^2, \ldots, \sigma_R^2\}$, Fisher threshold value θ.
Output: Linkage matrix Z.

1: Obtain the training $\mathcal{D} = \{x_i\}_{i=1}^N$ and validation $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_v}$ sets by subsampling $\mathcal{D}^{\text{full}}$ (e.g., via random subsampling).
2: For every combination of parameter pairs (k, σ^2) and using $\mathcal{D}$ and $\mathcal{D}^{\text{val}}$, train a kernel spectral clustering model using Algorithm 1 and obtain the Fisher criterion value using (16).
3: $\forall k$, find the maximum value of the Fisher criterion across the given range of σ^2 values. If the maximum value is greater than the Fisher threshold θ, create a set of these optimal $(k_\star, \sigma_\star^2)$ pairs.
4: Using $\mathcal{D}$ and the previously found $(k_\star, \sigma_\star^2)$ pairs, train a clustering model and compute the cluster memberships of the full dataset $\mathcal{D}^{\text{full}}$ using the out-of-sample extension.
5: Create the linkage matrix Z by identifying which clusters merge starting from the bottom of the tree which contains max $k_\star$ clusters.

6. Experimental Results

Simulation results are presented in this section. All experiments reported are performed in MATLAB on a Intel Core 2 Duo, 3.06 GHz, 4 GB, Mac OS X. All data have been split into training, validation and test with 50 randomizations of the sets. The Fisher threshold θ is set to 0.7. Thus, if the maximum Fisher criterion for each k is greater than 0.7, the corresponding parameter pair $(k_\star, \sigma_\star^2)$ is considered for building the hierarchical structure.

6.1. Toy Data

The first toy data corresponds to 5 Gaussian clusters in $2D$. The total number of points is $N_{\text{full}} = 2,000$ and the dataset is split into training ($N = 500$), validation ($N_{\text{v}} = 1,000$) and the remaining data points for test. The full dataset and the the model selection plots are presented in Figure 2. Note that according to the Fisher criterion, $k = 2, 3, 4$ and 5 are optimal candidates for building the dendrogram with maximum value equal to 1 indicating the presence of very strong cluster structures. The behavior of the Fisher criterion with respect to σ^2 also displays an underlying hierarchy. The optimal value of the RBF kernel parameter σ^2 decreases as the number of clusters k increases since more complex models are needed to discover the structure of the data. Thus, from this plot the following set of $(k_\star, \sigma^2_\star)$ is obtained: $\{(2, 20), (3, 6.6), (4, 2.6), (5, 1.95)\}$ which is used to create the linkage matrix Z and the corresponding dendrogram. The latter can be seen in Figure 3. At the bottom of the tree we start with 5 clusters which is the maximal $k_\star$ found during the model selection stage. The corresponding clustering results are shown un Figure 4. The underlying hierarchy is revealed by the proposed approach and the membership zones of each clusters display an excellent generalization performance.

The second toy data involves 4 Gaussian clouds with overlap. Overlapping causes non-zero similarities for points in different clusters leading to approximately piecewise constant eigenvectors. Figure 5 shows the model selection plots and the corresponding dendrogram. Note that in this case, $k = 2, 4$ are selected for creating the tree but $k = 3$ does not show a strong cluster structure. Two pairs of clusters merge at the same value of σ^2 and the clustering results are displayed in Figure 6. Despite the overlap, the proposed methodology succeeds in finding the natural hierarchical structure of the data.

6.2. Image segmentation

Image segmentation is a difficult application for spectral clustering due to the large number of data points leading to eigendecomposition of big matrices. In this experiment, we used an image from the Berkeley image database (Martin et al., 2001). We computed a local color histogram with a 5×5 pixels window around each pixel using minimum variance color quantization of 8 levels. We used the χ^2 test to compute the distance between two local color histograms $h^{(i)}$ and $h^{(j)}$ (Puzicha et al., 1997) $\chi^2_{ij} = 0.5 \sum_{b=1}^{B} (h_b^{(i)} - h_b^{(j)})^2 / (h_b^{(i)} + h_b^{(j)})$, where B is total number of quantization levels. The histograms are normalized $\sum_{b=1}^{B} h_b^{(i)} = 1, i = 1, \ldots, N$. The χ^2 kernel $K(h^{(i)}, h^{(j)}) = \exp(-\chi^2_{ij}/\sigma_\chi)$ with parameter $\sigma_\chi \in \mathbb{R}^+$

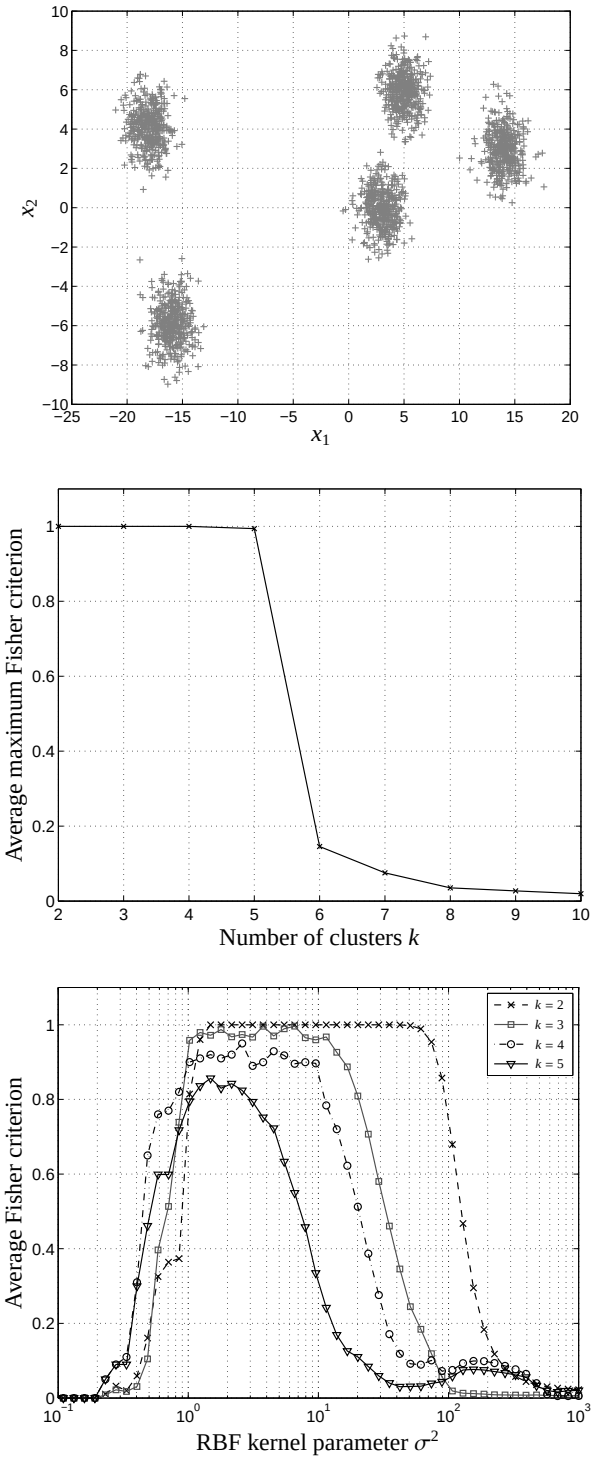

Figure 2: **Top:** Full dataset. **Center:** Average maximum Fisher criterion for determining the number of clusters k. The multiscale nature of this toy dataset is revealed by the fact that $k = 2, 3, 4$ and 5 give a very high value of the Fisher criterion. **Bottom:** Optimal RBF kernel parameter σ^2 with respect to k. Note that as k increases the range of the optimal σ^2 decreases giving preference to smaller scales.

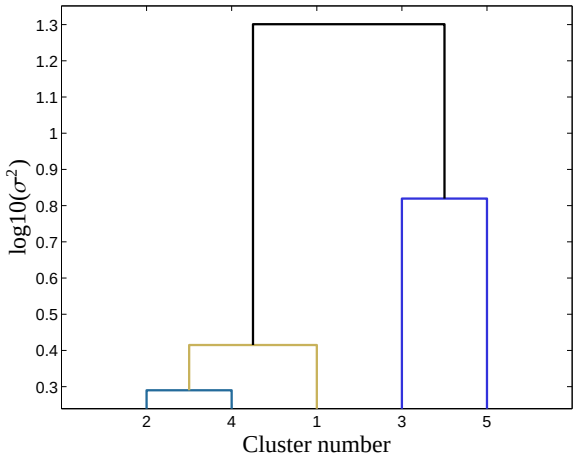

Figure 3: Resulting tree displaying the hierarchical structure of the data. The bottom of the tree contains the largest k that gives high average maximum Fisher criterion on validation data. The y-axis shows the value of σ^2 representing the scale.

is positive definite and has shown to be very performant for color discrimination and image segmentation. The image is 321×481 pixels leading to a total number of data points of $N_{\text{full}} = 154,401$. The training set consisted of 600 randomly selected pixel histograms and $20,000$ pixel histograms were used as validation. Model selection was performed to find strong cluster structures at $k = 2, 3$. Figure 7 shows the dendrogram and the segmentation results. Note that less than 0.4% of the total pixel histograms in the image are used for training. The cluster membership of the remaining 99.6% of the data is inferred using the out-of-sample extension. The obtained clustering display visually appealing clusters, a coherent merge and an enhance generalization capability. The average computation time for training the model and computing the out-of-sample eigenvectors was 5.14 ± 1.9 seconds.

7. Conclusions

A method to show a hierarchical tree in the context of kernel spectral clustering is presented. The proposed methodology is based on the Fisher criterion which can be used to find clustering parameters such that the resulting clusters are well-formed. The concept of eigenvector is also extended to out-of-sample data leading to a localized representation of the clusters. The clustering model can be trained in a learning setting ensuring good generalization capabilities. The hierarchical representation is obtained by evaluating the trained model

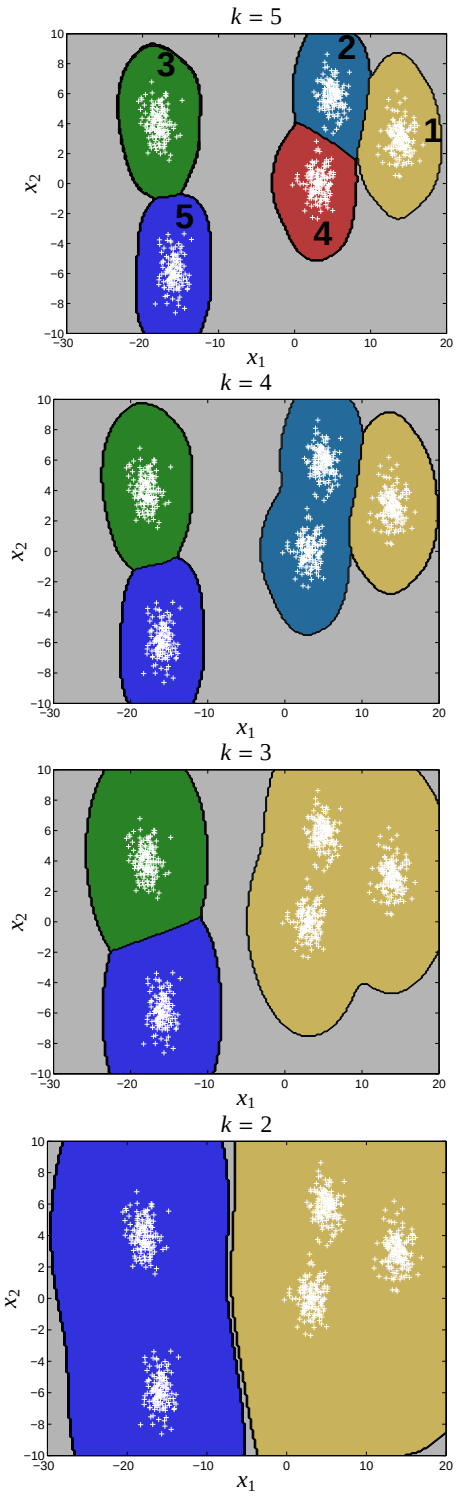

Figure 4: The plots show the obtained hierarchical representation and the cluster decision boundaries according to the out-of-sample extension. The gray areas represent the zero cluster corresponding to data points very close to zero in the out-of-sample eigenvector space leading to low confidence in the cluster membership assignment.

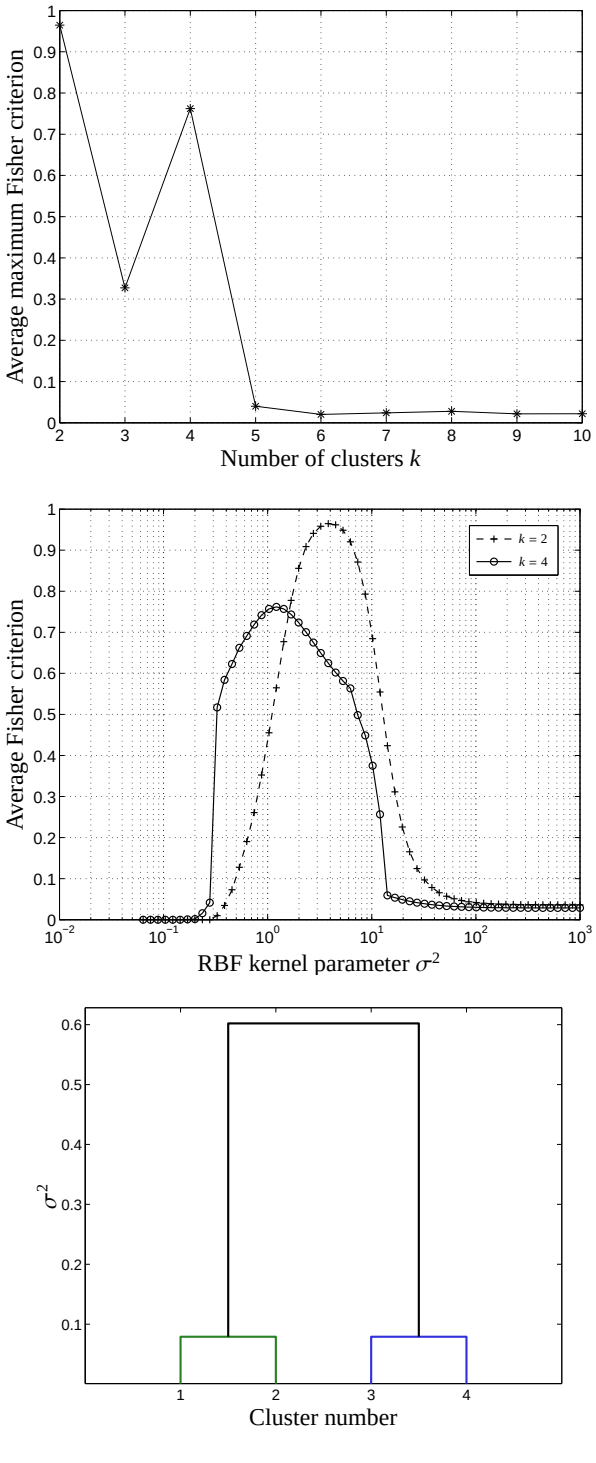

Figure 5: **Top:** Average maximum Fisher criterion for determining k. Note that only $k = 2, 4$ produced maximum Fisher criterion greater than 0.7 **Center:** Optimal RBF kernel parameter σ^2 with respect to k. **Bottom:** Resulting dendrogram displaying the hierarchy.

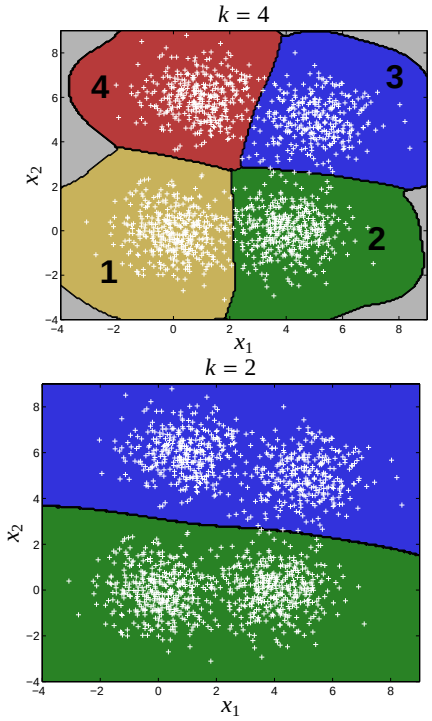

Figure 6: Clustering results showing good generalization performance in presence of cluster overlap.

on the whole data using the tuned clustering parameters. A dendrogram is formed by merging clusters in a bottom-up approach given by the predicted cluster memberships. The experimental results confirm the applicability of the proposed hierarchical method.

Acknowledgement

This work was supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet , CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants;Flemish Government:FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other:Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. Carlos Alzate is a postdoctoral fellow of the Research Foundation - Flanders (FWO). Johan Suykens is a professor at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

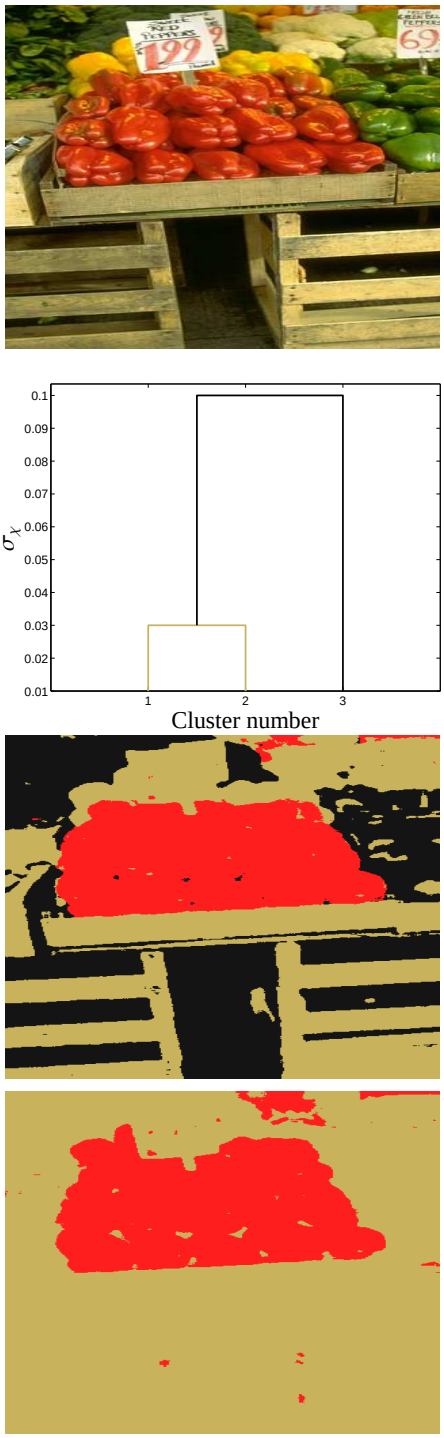

Figure 7: **From top to bottom:** 481 × 321 pixels image; Obtained dendrogram showing 3 clusters and one merge; Segmentation results for $k = 3$; Segmentation results for $k = 2$. The wooden crates and the background clusters merge into one big cluster while the red peppers cluster is practically unchanged.

References

Alzate, C., Suykens, J.A.K., 2010. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 335–347.

Alzate, C., Suykens, J.A.K., 2011. Out-of-sample eigenvectors in kernel spectral clustering, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN'11), pp. 2349–2356.

Bach, F.R., Jordan, M.I., 2006. Learning spectral clustering, with application to speech separation. Journal of Machine Learning Research 7, 1963–2001.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Chung, F.R.K., 1997. Spectral Graph Theory. American Mathematical Society.

Clauset, A., Moore, C., Newman, M., 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101.

Deuflhard, P., Huisinga, W., Fischer, A., Schütte, C., 2000. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Linear algebra and its applications 315, 39–59.

Ding, C., He, X., 2004. Linearized cluster assignment via spectral ordering, in: ICML '04: Proceedings of the twenty-first International Conference on Machine Learning, ACM Press, New York, NY, USA. p. 30.

Fowlkes, C., Belongie, S., Chung, F., Malik, J., 2004. Spectral grouping using the Nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 214–225.

Kato, T., 1995. Perturbation theory for linear operators. Springer.

von Luxburg, U., 2007. A tutorial on spectral clustering. Statistics and Computing 17, 395–416.

Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proc. 8th Int'l Conf. Computer Vision, pp. 416–423.

Meila, M., Shi, J., 2001. A random walks view of spectral segmentation, in: Artificial Intelligence and Statistics AISTATS.

Ng, A.Y., Jordan, M.I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm, in: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA. pp. 849–856.

Puzicha, J., Hofmann, T., Buhmann, J., 1997. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, in: Computer Vision and Pattern Recognition, pp. 267–272.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 22, 888–905.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.

Zhang, K., Tsang, I., Kwok, J., 2008. Improved nystrom low-rank approximation and error analysis, in: McCallum, A., Roweis, S. (Eds.), Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008), pp. 1232–1239.