Neural Networks 35 (2012) 21–30

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Hierarchical kernel spectral clustering

Carlos Alzate *, Johan A.K. Suykens

Department of Electrical Engineering ESAT-SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 24 May 2012
Accepted 22 June 2012

Keywords:
Hierarchical clustering
Spectral clustering
Kernel methods
Out-of-sample extensions

ABSTRACT

Kernel spectral clustering 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. An eigenvalue problem is solved at the training stage and projections onto the eigenvectors constitute the clustering model. The formulation allows out-of-sample extensions which are useful for model selection in a learning setting. In this work, we propose a methodology to reveal the hierarchical structure present on the data. During the model selection stage, several clustering model parameters leading to good clusterings can be found. These results are then combined to display the underlying cluster hierarchies where the optimal depth of the tree is automatically determined. Simulations with toy data and real-life problems show the benefits of the proposed approach.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Spectral clustering methods have been successfully applied on a variety of applications. These methods correspond to a family of unsupervised learning algorithms to find groups of data points that are similar (Ng, Jordan, & Weiss, 2002; Shi & Malik, 2000; von Luxburg, 2007). The clustering information can be obtained from analyzing the eigenvalues and eigenvectors of a matrix derived from pairwise similarities of the data. 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 (Bach & Jordan, 2006; Ding & He, 2004; Ng et al., 2002). 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, Belongie, Chung, & Malik, 2004; Zhang, Tsang, & Kwok, 2008). Another issue with classical spectral clustering is model selection. The determination of the similarity function, its parameters and the number of clusters is unclear and typically rely on heuristics. This problem is due to the lack of out-of-sample and predictive mechanisms.

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 the context of least squares support vector machines (LS-SVM) (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002). A clustering model is defined for obtaining the cluster membership of the data. The model is expressed in the primal as a set of projections in high-dimensional feature spaces similar to multi-class kernel machines. By using this model, it is possible to predict the cluster membership of unseen (also called out-of-sample) points allowing model selection in a learning framework, i.e., with training, validation and test stages. Having a predictive model is important for ensuring good generalization capabilities because the cluster membership of out-of-sample points can be predicted by evaluating the trained model. During the training stage, an eigenvalue decomposition is solved. The clustering model is then expressed at the dual level as projections which are linear combinations of kernel evaluations with eigenvectors as the coefficients. The sign of the projections is related to the cluster membership: projections of points in the same cluster have the same sign pattern, while projections of points in different clusters have different sign patterns. At the validation and test stages, the cluster membership of out-of-sample points is inferred by evaluating the model. When strong cluster structures are present in the data, the projections of data points in the same cluster are collinear in the projection space. Model selection can then be performed by finding clustering parameters such that the projections of validation data points show strong collinearity. Another advantage of having out-of-sample mechanisms is the possibility to select a small representative subsample of the data, solve an eigenvalue problem of reduced size and predict the cluster membership of the remaining data points. In this way, the computational burden can be greatly reduced.

The results obtained using spectral clustering methods do not provide hints about the relation between the clusters. In some

* Corresponding author.
E-mail addresses: carlos.alzate@esat.kuleuven.be (C. Alzate), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.06.007

applications, a more informative hierarchical representation of the data is desirable (Clauset, Moore, & Newman, 2008). This is typically the case when the data contain different natural groupings depending on the scale. In this paper, we propose a methodology to represent the results of kernel spectral clustering in a hierarchical way. This hierarchical representation is based on a model selection criterion evaluated on validation data. This criterion is called Balanced Linefit (BLF) and was introduced in Alzate and Suykens (2010). The BLF quantifies cluster collinearity and cluster balance. The main idea is to find clustering parameters (e.g., number of clusters and similarity parameters) such that the BLF has a high value, 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 which was first introduced in Alzate and Suykens (2011). In Section 3, we describe the model selection criterion based on the BLF criterion. Section 4 contains the proposed hierarchical representation together with 2 algorithms. In Section 5, we present the experimental results and in Section 6 we give conclusions.

2. Kernel spectral clustering

The kernel spectral clustering framework puts spectral clustering in a constrained optimization setting with primal and dual model representations. The primal problem is formulated as a weighted kernel PCA problem together with 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. 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:

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

where $w^{(l)} \in \mathbb{R}^{d_h}, \varphi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ 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 such that $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_k = \mathcal{D}$ and $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_k = \emptyset$. The primal problem is formulated as:

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

where $\gamma_l \in \mathbb{R}^+$ are regularization constants, $\Phi = [\varphi(x_1), \ldots, \varphi(x_N)]^T$ is the $N \times d_h$ training feature matrix, $V = \text{diag}([v_1, \ldots, v_N])$ and $v_i \in \mathbb{R}^+$ 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 & 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:*

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

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 $\lambda_1 \geq \cdots \geq \lambda_{n_e}$ and $\alpha^{(l)}$ are the corresponding eigenvectors.

Proof. The Lagrangian of (2) is

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

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

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

$l = 1, \ldots, n_e$. Eliminating $w^{(l)}, e^{(l)}$ and expressing the KKT conditions in terms of the dual variables $\alpha^{(l)}$ leads to (3). □

The dual eigenvalue problem (3) has N possible eigenvalues and eigenvectors from which we will select a set of n_e eigenpairs. Note that the primal problem (2) is non-convex. Therefore, the dual solutions obtained through the KKT conditions are stationary points of the Lagrangian. However, the primal objective function equals zero when evaluated at the dual solutions. This means that the eigenvectors and eigenvalues can be interpreted as a pool of candidate solutions.[1] Note also 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:

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

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

[1] The objective function in the primal problem in Alzate and Suykens (2010) has opposite sign compared to (2). Nevertheless, the two problems have identical solutions at the dual level.

to obtain eigenvectors with discriminative clustering information. If $V = D^{-1}$ where $D = \text{diag}([d_1, \ldots, d_N])$, $d_i = \sum_{j=1} \Omega_{ij}$, $i = 1, \ldots, N$ is the so-called degree of x_i, then the dual problem (3) becomes:

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

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 transition matrix of a random walk $P = D^{-1}S$ 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 & 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 $k > 1$ connected components with high intra-cluster similarity and low inter-cluster similarity, then there are k eigenvectors of P with eigenvalue close to 1. When the connected components inter-similarity is equal to zero, these k eigenvectors are linear combinations of the indicator vectors of the partitioning. An indicator vector $\mathbb{1}_{\mathcal{A}_p} \in \{0, 1\}^N$ relative to a set $\mathcal{A}_p$ has i-th entry equal to 1 if $x_i \in \mathcal{A}_p$ and 0 otherwise. This result means that the top k eigenvectors are *piecewise constant* on the partitioning and the corresponding eigenvalues are all equal to 1. A vector $\beta = [\beta_1, \ldots, \beta_N]^T$ is piecewise constant relative to the set $\mathcal{A}_p$ when

$$\beta_i = \beta_j = c_p, \quad \text{if } x_i, x_j \in \mathcal{A}_p,$$

where c_p is an arbitrary constant value corresponding to the p-th cluster, $p = 1, \ldots, k$. In other words, data points belonging to the same cluster are represented with exactly the same value in the eigenvectors. These constant values correspond to the coefficients in the linear combination of the indicator vectors. Thus, spectral clustering 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 (Deuflhard, Huisinga, Fischer, & Schütte, 2000; Meila & Shi, 2001). 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[2] and that $\alpha^{(l)}$ are piecewise constant eigenvectors of P. Expanding (7) leads to:

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

and using the fact that the eigenvectors are piecewise constant on the underlying partitioning of P with eigenvalue 1 (i.e., $P\alpha = \alpha$), the equation above can be further simplified to:

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

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 the KKT condition $1_N^T\alpha^{(l)} = 0$, 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[3] with zero mean corresponding to the eigenvalue 1. With this in mind, we set $n_e = k - 1$ since all clustering information is present on the top $k - 1$ eigenvectors.

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 be combined into the final k clusters. The eigenvectors and the projections have the same sign pattern due to the KKT condition $\alpha_i^{(l)} = \frac{\gamma_l}{N} D_{ii}^{-1} e_i^{(l)}$, $i = 1, \ldots, N, l = 1, \ldots, k - 1$ since γ_l and the diagonal of D^{-1} are positive. Combining this property with the fact that the eigenvectors have zero mean leads to a geometrical structure of the clusters: 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. 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 by appropriate subsampling. Given an arbitrary data point x, the latent variable of the clustering model becomes:

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

$l = 1, \ldots, k - 1$, 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. When the out-of-sample points are very different from the training points, the kernel evaluations with respect to all training data are very close to zero. This indicates that the out-of-sample points can be categorized as outliers. In other words, if $\|[K(x_1, x), \ldots, K(x_N, x)]^T\|_2 < \theta_0$ where θ_0 is a pre-specified threshold then x is assigned to the *zero* cluster $\mathcal{A}_0$ which contains outlying data points.

2.5. 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

[2] In general, the RBF kernel function does not become exactly zero. A discussion about this issue is presented later.

[3] Note that 1_N is an eigenvector of P but not of $D^{-1}M_D\Omega$, since it does not fulfill the KKT condition $1_N^T\alpha^{(l)} = 0$.

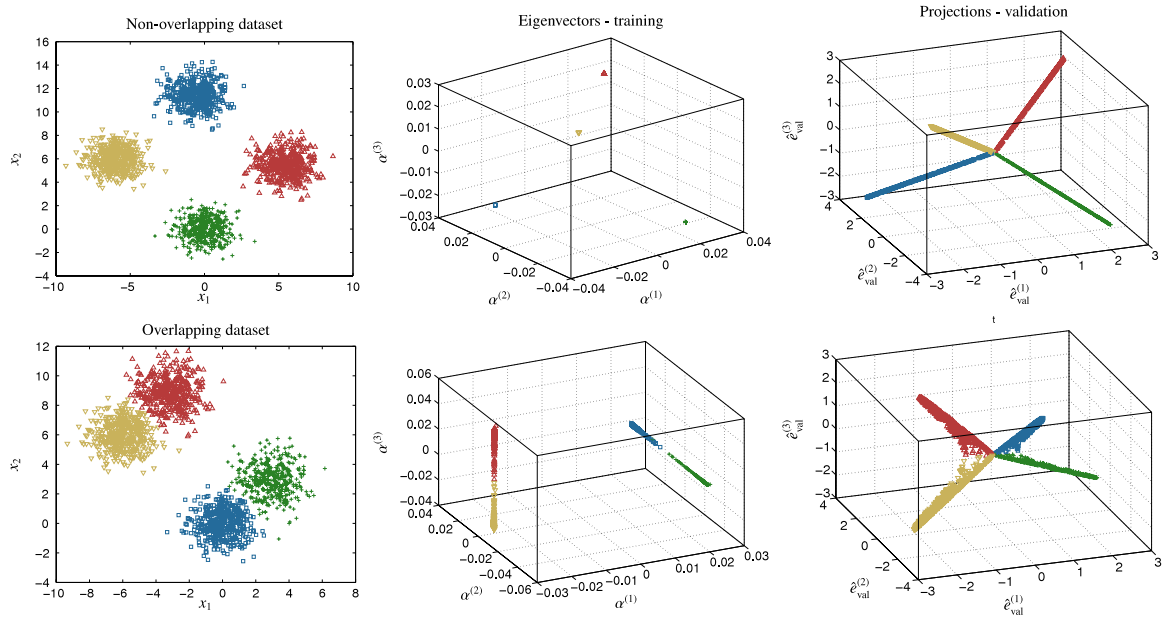


Fig. 1. Illustration of overlapping and non-overlapping clusters and the effect on the eigenvectors and projections. The first row depicts an ideal situation with 4 non-overlapping Gaussian clouds. For an appropriate choice of the RBF kernel parameter, the eigenvectors are piecewise constant and the projections are collinear. The second row corresponds to overlapping clouds causing non-zero inter-cluster similarities. The cluster overlap distorts the piecewise constant structure only slightly. Since decisions are taken considering the sign of the projections, the clustering is robust with respect to small distortions caused by overlap.

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. Fig. 1 shows an illustrative example of a clustering problem with four Gaussian clouds in non-overlapping and overlapping situations. The non-overlapping case leads to very small inter-cluster similarities (tending to zero) which leads to piecewise constant eigenvectors and collinear projections for out-of-sample data. Each data point in a given cluster is mapped exactly to the same point in the eigenvectors and these points are in different orthants. Likewise, the clusters form lines in the projection space also in different orthants. This property will be exploited for performing model selection in the next section. As can be seen in the figure, the overlapping case leads to approximately piecewise constant eigenvectors and approximately collinear projections.

3. Model selection

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.

3.1. Cluster 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. 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. The balanced linefit (BLF) criterion

Consider the following measure of cluster collinearity of a validation dataset $\mathcal{D}^v = \{x_m^{\text{val}}\}_{m=1}^{N_v}$, with respect to a partitioning $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ and $k > 2$:

$$\text{collinearity}(\mathcal{D}^v, k > 2, p) = \frac{k-1}{k-2}\left(\frac{\zeta_1^{(p)}}{\sum_{l=1}^{k-1}\zeta_l^{(p)}} - \frac{1}{k-1}\right), \qquad (9)$$

where $\zeta_1^{(p)} \geq \cdots \geq \zeta_{k-1}^{(p)}$ are the ordered eigenvalues of the covariance matrix $C_Z^{(p)} = (1/|\mathcal{A}_p|){Z^{(p)}}^T Z^{(p)}$, $p = 1, \ldots, k$, $Z^{(p)} \in \mathbb{R}^{|\mathcal{A}_p|\times(k-1)}$ is the matrix containing the projections (with the mean removed) of validation data assigned to the p-th cluster on the rows. The term $\zeta_1^{(p)} / \sum_l \zeta_l^{(p)}$ equals 1 when all energy is present on the first principal direction, i.e., when the rows of $Z^{(p)}$ are collinear. The additional terms in (9) scale the collinearity between 0 and 1. Note that the collinearity is defined for $k > 2$. In case of $k = 2$

clusters, only one eigenvector and projection is need to induce a bipartitioning. To be able to compute collinearity for $k = 2$, the following ad hoc modification has been proposed in Alzate and Suykens (2010):

$$\text{collinearity}(\mathcal{D}^v, k = 2, p) = \sum_{p=1}^{2} \left(\frac{\tilde{\zeta}_1^{(p)}}{\tilde{\zeta}_1^{(p)} + \tilde{\zeta}_2^{(p)}} - \frac{1}{2} \right), \tag{10}$$

where $\tilde{\zeta}_1 \geq \tilde{\zeta}_2$ are the ordered eigenvalues of $C_{\tilde{Z}}^{(p)} \in \mathbb{R}^{2\times 2}$ defined as $C_{\tilde{Z}}^{(p)} = (1/|\mathcal{A}_p|)\tilde{Z}^{(p)^T}\tilde{Z}^{(p)}$ and $\tilde{Z}^{(p)} \in \mathbb{R}^{|\mathcal{A}_p|\times 2}$ is an augmented projection matrix containing the zero-mean projections on the first column. The second column contains $(1/N_v)\sum_{i=1}^{N} K(x_i, x_m^{\text{val}}) + b$ on the m-th entry. The minimal linefit is defined as the minimum collinearity value between the clusters:[4]

$$\text{minlinefit}\,(\mathcal{D}^v, k) = \min_p \text{collinearity}(\mathcal{D}^v, k, p). \tag{11}$$

In some applications, the balance of the obtained clustering is also important. Cluster balance can be measured as $\text{balance}(\mathcal{D}^v, k) = \min\{|\mathcal{A}_p|\}/\max\{|\mathcal{A}_p|\}$, $p = 1, \ldots, k$. The BLF model selection criterion is then defined as:

$$\begin{aligned}\text{BLF}\,(\mathcal{D}^v, k) &= \eta\ \text{minlinefit}\,(\mathcal{D}^v, k)\\ &\quad + (1-\eta)\ \text{balance}\,(\mathcal{D}^v, k),\end{aligned} \tag{12}$$

where $0 \leq \eta \leq 1$ is a user-defined parameter determining the weight given to the linefit with respect to the balance index.

4. Hierarchical representation

Typically, the clusters found by classical clustering methods do not give hints about the relationship between the clusters. 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.

4.1. 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 indicates 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: single, complete, average and Ward's criterion. 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.

[4] The original linefit in Alzate and Suykens (2010) is expressed as the average cluster collinearity rather than the minimum.

4.2. Hierarchical kernel spectral clustering

In this section, we propose a methodology based on KSC to discover cluster hierarchies. During the model selection process, the BLF criterion can indicate that there are several cluster parameter pairs (k, σ^2) leading to good clustering results. A grid search over different values of k and σ^2 evaluating the BLF on validation data is performed. The parameter pairs (k, σ^2) for which the BLF is larger than a specified threshold value θ are selected. 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 typically represents the scale and corresponds to the value of σ^2 at which the merge occurs. In cases when all merges occur at the same scale, the y-axis represents the merge order.

Note that during a merge, there might be some 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 and dendrogram 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 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 $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_v}$ and zero cluster threshold θ_0.

Output: Partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, cluster codeset $\mathcal{C}$, zero cluster $\mathcal{A}_0$.

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: Compute the projections for training data $e^{(l)} = \Omega\alpha^{(l)} + b_l 1_N$.
3: Binarize the projections: $\text{sign}(e_i^{(l)})$, $i = 1, \ldots, N$, $l = 1, \ldots, k-1$, and let $\text{sign}(e_i) \in \{-1, 1\}^{k-1}$ be the encoding vector for the training data point x_i, $i = 1, \ldots, N$.
4: Count the occurrences of the different encodings and find the k encodings with most occurrences. Let the codeset be formed by these k encodings: $\mathcal{C} = \{c_p\}_{p=1}^k$, $c_p \in \{-1, 1\}^{k-1}$
5: $\forall i$, 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.
6: Compute the projections for validation data $\hat{e}_{\text{val}}^{(l)} = \Omega_{\text{val}}\alpha^{(l)} + b_l 1_{N_v}$.
7: Find all data points x for which $\|[K(x_1, x), \ldots, K(x_N, x)]^T\|_2 < \theta_0$ and assign them to $\mathcal{A}_0$.
8: Binarize the projections for validation data: $\text{sign}(\hat{e}_{\text{val},t}^{(l)})$, $t = 1, \ldots, N_v$, $l = 1, \ldots, k-1$, and let $\text{sign}(\hat{e}_{\text{val},t}) \in \{-1, 1\}^{k-1}$ be the encoding vector for the training data point x_i, $i = 1, \ldots, N$.
9: $\forall t$, assign x_t^{val} to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\text{sign}(\hat{e}_{\text{val},t}^{(l)}), c_p)$.

Note also that the proposed hierarchical methodology does not use classical linkage criteria to build up the tree. Instead, the dendrogram is built upon the predicted cluster memberships of the full dataset using the out-of-sample extension. The maximum number of clusters (depth of the tree) is determined automatically by model selection. This also differs with classical hierarchical clustering where the depth of the tree is always the total number of observations and the number of clusters still need to be determined. A MATLAB implementation of the proposed approach is available at http://www.esat.kuleuven.be/sista/lssvmlab/KSC/HKSCDemo.m which is part of the KSC toolbox available at http://www.esat.kuleuven.be/sista/lssvmlab/KSC.

5. 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,

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 values of σ^2 $\{\sigma_1^2, \ldots, \sigma_R^2\}$, BLF threshold value θ.
Output: Linkage matrix Z.

1: Obtain the training $\mathcal{D} = \{x_i\}_{i=1}^{N}$ and validation $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_{\text{v}}}$ 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 BLF criterion value using (12).
3: $\forall k$, find the maximum value of the BLF criterion across the given range of σ^2 values. If the maximum value is larger than the BLF 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.

4 GB, Mac OS X. All data have been split into training, validation and test with 10 randomizations of the sets. The BLF threshold θ is problem dependent and is set to 75% of the global maximal BLF. Thus, if the maximum BLF value for each k is greater than this threshold, the corresponding parameter pair $(k_\star, \sigma_\star^2)$ is considered for building the hierarchical structure. The tradeoff η characterizing the weight given to the collinearity and the balance in the BLF criterion is fixed to 0.75. This choice gives more weight to finding collinear clusters than to balanced ones.

5.1. Toy data

The first toy data corresponds to 5 Gaussian clusters in 2D. The total number of points is $N_{\text{full}} = 2000$ and the dataset is split into training ($N = 500$), validation ($N_{\text{v}} = 1000$) and the remaining data points for test. The full dataset, model selection plots and inferred dendrogram are presented in Fig. 2. Note that according to the BLF 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 BLF 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 $(k_\star, \sigma_\star^2)$ is obtained: $\{(2, 20.0), (3, 4.0), (4, 3.0), (5, 2.0)\}$ which is used to create the linkage matrix Z and the corresponding dendrogram. At the bottom of the tree we start with 5 clusters which is the maximal $k_\star$ found during the model selection stage. Fig. 3 shows a dendrogram calculated using classical hierarchical clustering with average linkage. The tree was arbitrarily cut at $k = 8$ since there is no clear mechanism to find the optimal number of clusters in the classical hierarchical methods. Moreover, the full dataset is used in building the dendrogram since there are no out-of-sample and predictive mechanisms. The clustering results of the proposed approach are shown in Fig. 4. The underlying hierarchy is revealed by the proposed approach and the membership zones of each cluster display excellent generalization performance.

5.2. 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, Fowlkes, Tal, & Malik, 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

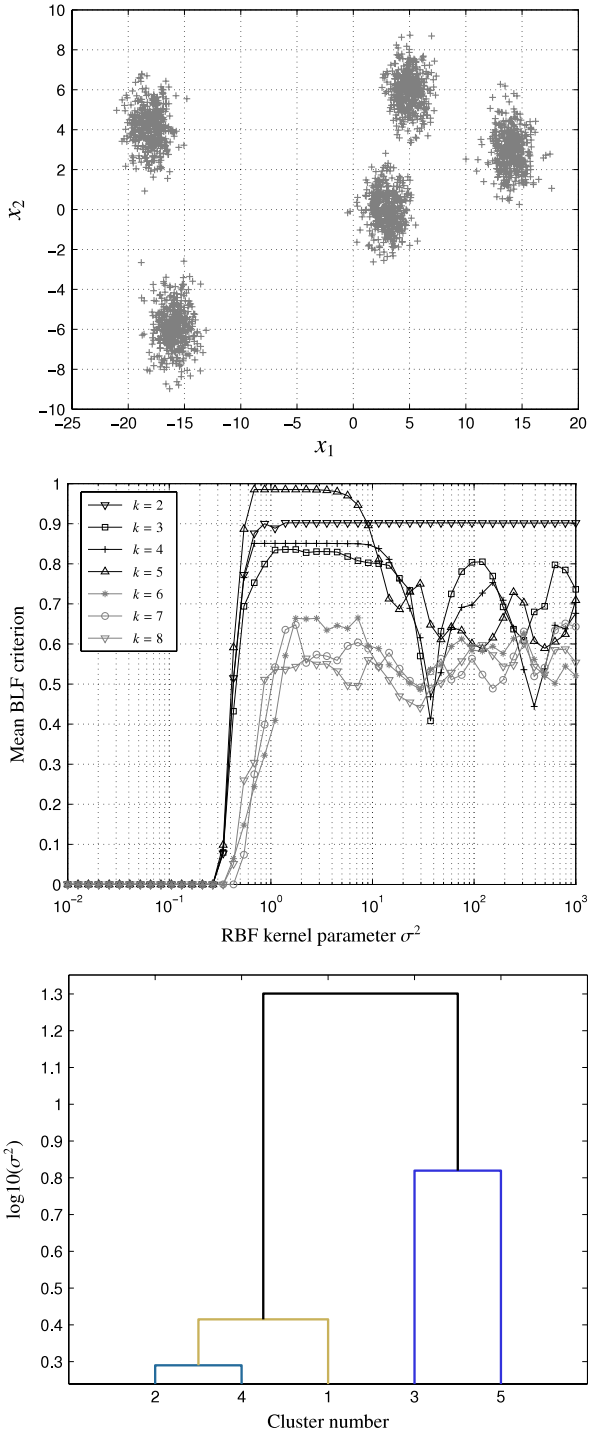


Fig. 2. *Left:* full dataset. *Center:* model selection plot on validation data to determine the parameter pairs (k, σ^2). The multiscale nature of this toy dataset is revealed by the fact that only $k = 2, 3, 4$ and 5 exceed the BLF threshold value $\theta = 0.7$. *Right:* resulting tree displaying the hierarchical structure of the data. The bottom of the tree contains the largest k that gives high average maximum BLF criterion on validation data. The y-axis shows the value of σ^2 representing the scale.

the distance between two local color histograms $h^{(i)}$ and $h^{(j)}$ (Puzicha, Hofmann, & Buhmann, 1997): $\chi_{ij}^2 = 0.5 \sum_{b=1}^{B} (h_b^{(i)} - h_b^{(j)})^2 / (h_b^{(i)} + h_b^{(j)})$, 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 $K(h^{(i)}, h^{(j)}) = \exp(-\chi_{ij}^2/\sigma_\chi)$ with parameter $\sigma_\chi \in \mathbb{R}^+$ is positive definite and has shown to be very performant for color discrimination and image segmentation. The image is 321×481

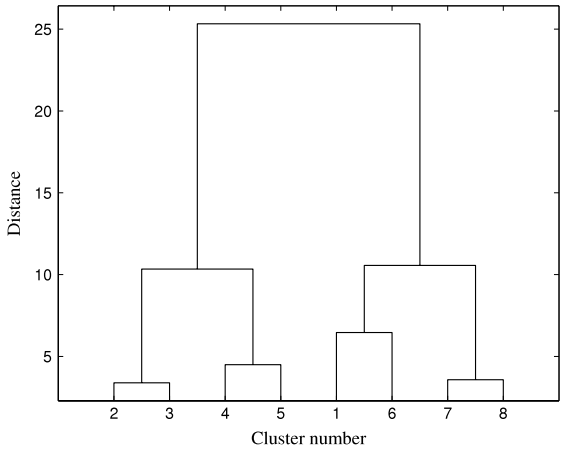


Fig. 3. Dendrogram using classical hierarchical clustering with average linkage. The tree was arbitrarily cut at $k = 8$ clusters. Note that all points are used for calculating the tree since there are no out-of-sample mechanisms. If the tree is cut at $k = 5$, the clusters are correctly detected. However, determining the right k is not clear from the classical hierarchical clustering methodologies.

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 $k = 2, 3$. Fig. 5 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 size of the eigenvalue problem is only 600×600. 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 enhanced generalization capability. The average computation time for training the model and computing the cluster memberships of all pixel histograms was 5.14 ± 1.9 s.

5.3. Text clustering

We used a subset of the 20`newsgroups` text dataset.[5] There are a total of 16,242 postings (observations) from the groups `comp.*` (computed-related), `rec.*` (recreational activities), `sci.*` (scientific) and `talk.*` (controversial, e.g., politics, religion). Each posting is represented with binary occurrence information of 100 terms (variables). The objective here is to find groups of similar postings. The training set consists of 1000 postings, the validation set has 4000 postings and the remaining observations constitute the test set. These sets were randomized 10 times. An RBF-type kernel was used for this application: $K_{\cos}(x_i, x_j) = \exp\left(-d_{\cos}(x_i, x_j)^2/\sigma_{\cos}^2\right)$ where $d_{\cos}(x_i, x_j) = 1 - x_i^T x_j/(\|x_i\|_2\|x_j\|_2)$ is the cosine distance which is commonly used in text mining applications. Model selection and clustering results are shown in Fig. 6. The BLF threshold is set to $\theta = 0.6$. Only mean BLF values on validation data exceeding θ are considered to build the tree leading to a number of clusters from $k = 2$ until $k = 7$. The model selection plot does not show a multiscale behaviour as in the previous experiments. All selected possible candidates for k exceeded θ around the same scale $\sigma_{\cos}^2 = 0.1$. The obtained dendrogram shows the hierarchical partitioning at the same scale. Table 1 shows the 5 terms with most occurrences per cluster confirming a coherent tree. A comparison of the proposed method and classical hierarchical clustering in terms of agreement with respect to external classification labels is also shown in Fig. 6. In the case of classical hierarchical

[5] Downloaded from http://cs.nyu.edu/roweis/data.html.

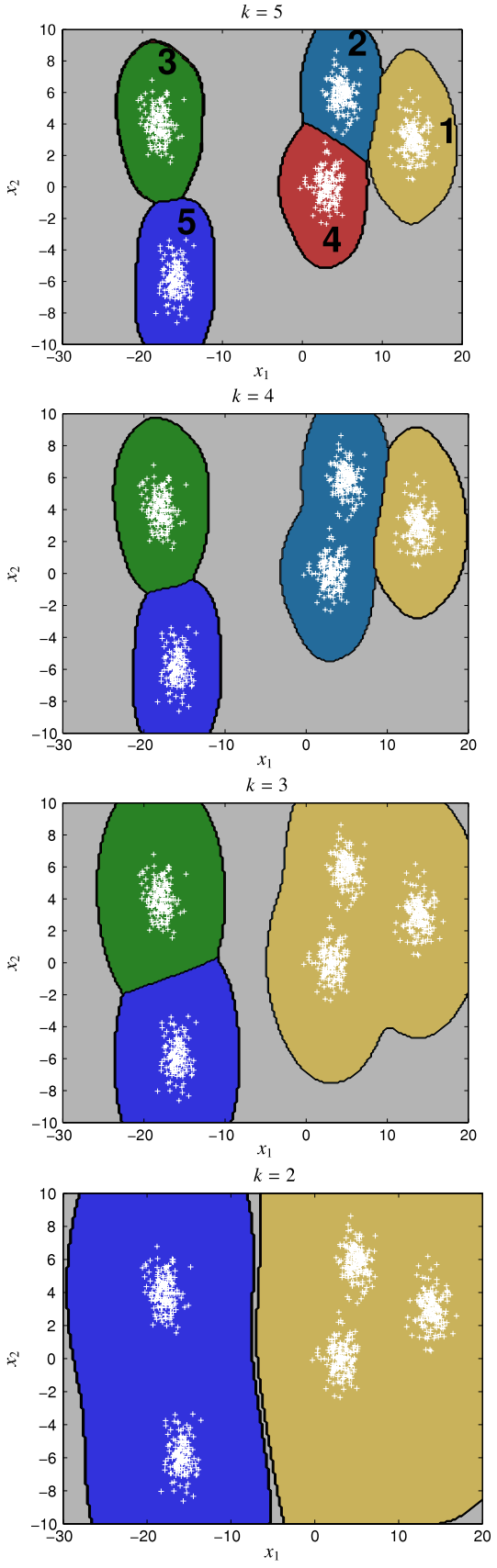


Fig. 4. The plots show the obtained hierarchical representation and the cluster decision boundaries according to the out-of-sample extension. The gray areas represent the zero cluster $\mathcal{A}_0$ corresponding to outlying data points where $\|[K(x_1, x), \ldots, K(x_N, x)]^T\|_2 < \theta_0$ where $\theta_0 = 1 \times 10^{-3}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



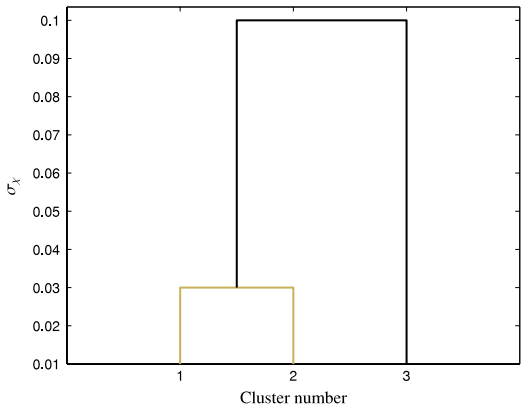
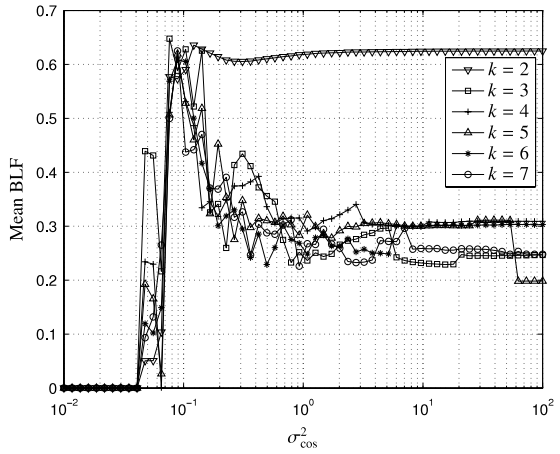
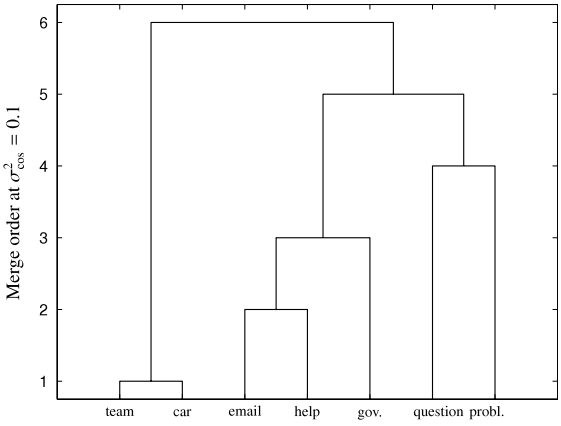






Fig. 5. *From top to bottom:* 481 × 321 pixels image; obtained dendrogram showing 3 clusters and one merge; segmentation results for $k = 3$; segmentation results for $k = 2$. The wooden crates and the background clusters merge into one big cluster while the red peppers cluster is practically unchanged. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)





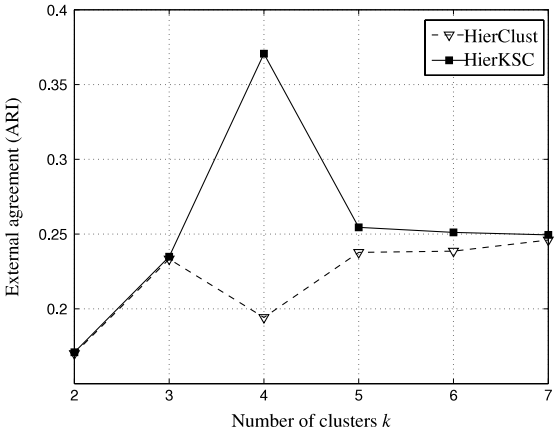


Fig. 6. *Top:* model selection plot on validation data. Only $k = 2, 3, 4, 5, 6, 7$ exceed the BLF threshold $\theta = 0.6$. The BLF gives high value around $\sigma^2_{\cos} = 0.1$. *Center:* induced dendrogram using the cluster parameters found during model selection. Note that the dendrogram displays several possible groupings at the same scale. *Bottom:* agreement with respect to external classification labels for classical and kernel hierarchical clustering. The number of classes is 4. The dendrograms of the compared methods are cut at different k and the agreement with the labels is reported in terms of ARI. The proposed method shows improved agreement with the labels.

clustering, the complete dataset was used to build the tree since it does not possess an out-of-sample mechanism. Ward linkage with cosine distance was used to build the tree. The proposed method compares favorably with respect to classical hierarchical clustering giving a better agreement at all k, particularly for $k = 4$ which corresponds to the number of classes in the classification labels.

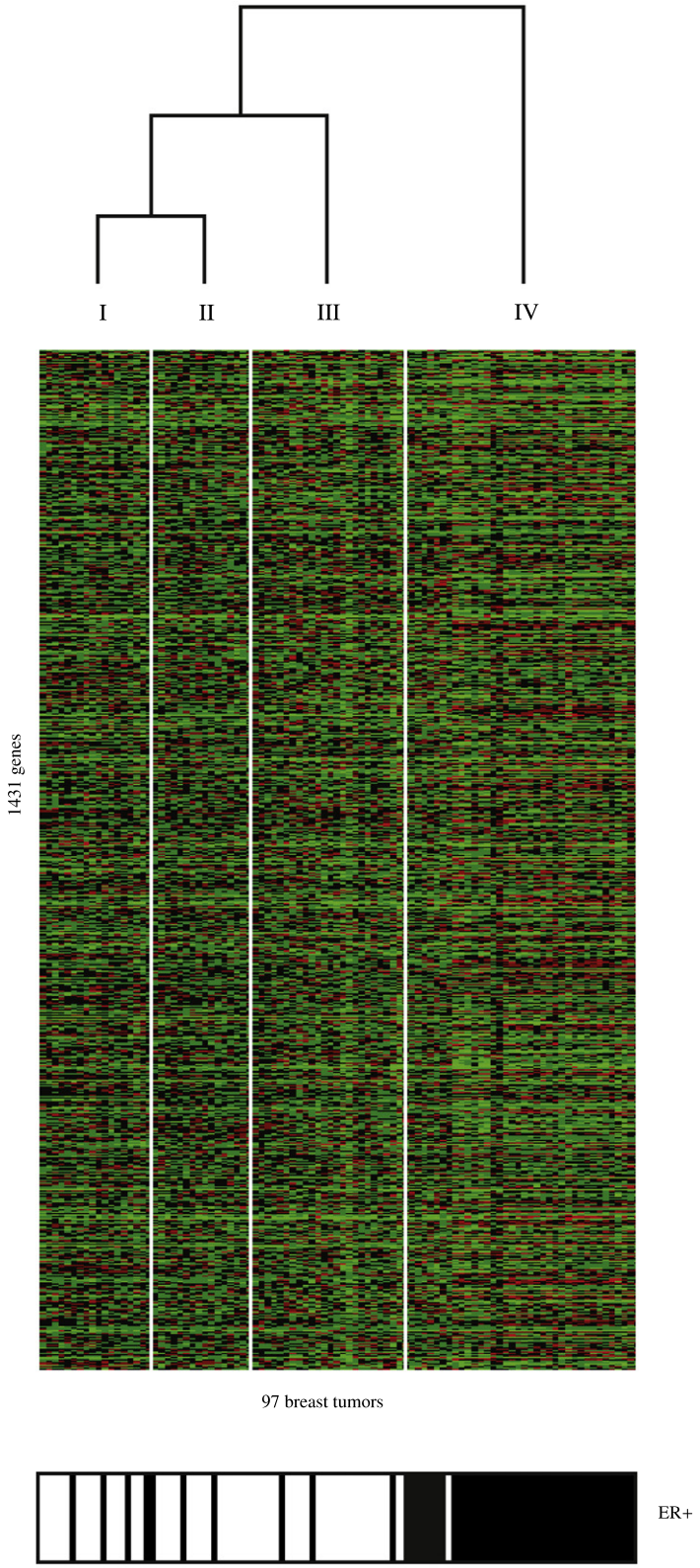


Fig. 7. *Top:* induced dendrogram on a microarray dataset using hierarchical KSC. From model selection, the maximum number of clusters found on the data is $k = 4$. *Center:* microarray with cluster boundaries depicted as white lines. *Bottom:* estrogen receptor status (ER) of each tumor. White color indicates positive status. The clustering shows a coherent agreement with respect to the ER: 84.1% of clusters I ∪ II ∪ III is ER+ while only 2.9% of cluster IV is ER+.

5.4. Microarray data

We used the breast cancer microarray dataset introduced in van 't Veer et al. (2002). The dataset contains 97 breast cancers and around 250,000 gene expression values. The data are already normalized and log-transformed. To select informative genes, we used a similar procedure as in van 't Veer et al. (2002). Gene expressions with missing values were discarded and only those

Table 1
Best terms representing the clusters. The 5 terms with most occurrences per cluster are shown. The partitioning shows a coherent relationship between the terms. The term with most occurrences is shown in bold.

Cluster	Best 5 terms
I	**Team**, games, players, hockey, season
II	**Car**, course, case, fact, number
III	**Email**, university, help, phone, computer
IV	**Help**, windows, system, program, problem
V	**Government**, state, jews, rights, evidence
VI	**Question**, god, christian, fact, jesus
VII	**Problem**, space, nasa, system, drive

with at least a three-fold difference and p-value less than 0.01 in more than 5 tumors were selected. This procedure leads to 1431 genes. The idea here is to obtain groups of similar tumors. We performed model selection on validation data as explained in the previous experiments. The BLF threshold value was set to $\theta = 0.7$. Only $k = 2, 3, 4$ exceed θ at the same scale given by the RBF kernel parameter $\sigma^2 = 110$. The hierarchical results are shown in Fig. 7. The clustering induced by the dendrogram is compared with external clinical data given by the estrogen receptor (ER) status which is known to play a role in breast cancer (Hoshida, Brunet, Tamayo, Golub, & Mesirov, 2007; van 't Veer et al., 2002). The results indicate good agreement with respect to the ER status particularly for $k = 2$. The metacluster composed of clusters I ∪ II ∪ III has 53/63 tumors which are ER+Òn the other hand, cluster IV contains only 1/34 tumors with positive estrogen receptor status which constitute an improvement with respect to previous results (50/61, 3/36) as reported in Hoshida et al. (2007).

6. Conclusions

A method to show a hierarchical tree in the context of kernel spectral clustering is presented. The proposed methodology is based on the BLF criterion which can be used to find clustering parameters such that the resulting clusters are well-formed. The clustering model can be trained in a learning setting ensuring good cluster prediction capabilities on unseen points. The proposed methodology can also be used to reduce the computational cost by subsampling. The hierarchical representation is obtained by evaluating the trained model 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.

Acknowledgments

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 Ph.D./postdoc & fellow grants;Flemish Government:FWO: Ph.D./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), G.0377.12 (Structured systems) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: Ph.D. 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), ERC AdG A-DATADRIVE-B, COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. CA is a postdoctoral fellow of the Research Foundation — Flanders (FWO). JS is a professor at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

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.

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 *Proceedings of the twenty-first international conference on machine learning*, ICML '04. (p. 30). New York, NY, USA: ACM Press.

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.

Hoshida, Y., Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2007). Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS ONE*, *2*, e1195.

Kato, T. (1995). *Perturbation theory for linear operators*. Springer.

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 T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 849–856). Cambridge, MA: MIT Press.

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 Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

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

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).