Out-of-Sample Eigenvectors in Kernel Spectral Clustering

Carlos Alzate and Johan A. K. Suykens

***Abstract*—A method to estimate eigenvectors for out-of-sample data in the context of kernel spectral clustering is presented. The proposed method is within a constrained optimization framework with primal and dual model representations. This formulation allows the clustering model to be extended naturally to out-of-sample points together with the possibility to perform model selection in a learning setting. A model selection methodology based on the Fisher criterion is also presented. The proposed criterion can be used to select clustering parameters such that the out-of-sample eigenvector space show a desirable structure. This special structure appears when the clusters are well-formed and the clustering parameters have been chosen properly. Simulation results with toy examples and images show the applicability of the proposed method and model selection criterion.**

I. Introduction

SPECTRAL clustering methods are a family of unsupervised learning algorithms to find groups of data points that are similar [1], [2], [3]. These techniques are relaxations of graph partitioning problems. The relaxed solutions can be obtained from an eigenvalue decomposition of a similarity matrix derived from the data. Spectral clustering has been shown to outperform classical methods such as k-means and hierarchical clustering typically when the data are highly non-linear. In general, the clustering is obtained only for the data used to compute the similarity matrix and its eigendecomposition. Extending the cluster membership to unseen points (also called out-of-sample points) is unclear from the classical spectral clustering formulation although approximations such as the Nyström method have been discussed [4]. Another issue with classical spectral clustering concerns the choice of the similarity function and its parameters. Parameter tuning is practically non-existent and the values of the parameters are determined via simple heuristics.

A different view of spectral clustering has been discussed in [5]. The formulation can be seen as a weighted version of kernel PCA in a constrained optimization setting with primal and dual model representations typical of least squares support vector machines (LS-SVM) [6]. By defining a clustering model, it is possible to obtain out-of-sample extensions and to perform model selection in a learning framework. The out-of-sample extension is based on projections over the eigenvectors obtained during the training stage. A model selection criterion is also described in [5] which makes use of the structure of the projections when the clusters are well-formed and the clustering parameters have been selected properly. Since eigenvectors only exist for the training points, cluster decisions and model selection using out-of-sample data should be performed on the projections.

In this paper, we propose a method to estimate eigenvectors for out-of-sample data. The methodology is based on the relationship between the eigenvectors and the projections. This relation arises as one of the Karush-Kuhn-Tucker (KKT) conditions which give the optimal representation of the dual problem given the primal problem. The eigenvectors obtained during training and the estimated out-of-sample eigenvectors also display a special structure in ideal cases with no cluster overlap and well-chosen similarity parameters. In this new representation, data points in the same cluster appear very localized in the eigenvector space. A model selection methodology based on the Fisher criterion [7] is also proposed. The objective is to obtain clustering parameters such as the number of clusters and the similarity parameters by maximizing the Fisher criterion on the eigenvector space for validation data. This criterion is maximal when the clusters are represented as single points and the centers are far away from each other.

This paper is organized as follows: Section II summarizes classical spectral clustering. Section III contains a review of the kernel spectral clustering method introduced in [5]. In Section IV, we describe the proposed method for computing out-of-sample eigenvectors together with an algorithm. Section V contains the proposed model selection scheme which is based on the Fisher criterion. In Section VI, we present the experimental results and in Section VII we give conclusions.

II. Classical Spectral Clustering

A. Preliminaries

Given a set of training points $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and a symmetric similarity measure $s_{ij} \geq 0, s_{ij} = s_{ji}$ between each pair of data points $x_i, x_j, i, j = 1, \ldots, N$, a graph can be formed by considering $\mathcal{D}$ as the set of vertices and s_{ij} as the weight of the edge between x_i and x_j[1]. A classical scenario in graph theory is the graph partitioning problem [8], [9]. The objective is to split the graph into k sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\bigcap_{p=1}^k \mathcal{A}_p = \emptyset$ and $\bigcup_{p=1}^k \mathcal{A}_p = \mathcal{D}$ by cutting edges with low weight. The graph partitioning problem is NP-hard due to its combinatorial nature. However, several relaxations leading to eigenvalue problems have been proposed, being the binary normalized cut (NCUT) relaxation [1] and the k-way normalized cut relaxation [10] the most popular methods. Minimizing the $\text{NCUT}(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is defined as

$$\begin{aligned} \min_G \quad & k - \text{trace}(G^T L G) \\ \text{s.t.} \quad & G^T G = I_k \end{aligned} \tag{1}$$

The authors are with the Department of Electrical Engineering ESAT-SCD-SISTA, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium (email: carlos.alzate@esat.kuleuven.be, johan.suykens@esat.kuleuven.be).

[1]A graph can also have been given without being associated to an underlying dataset. This particular case is not considered in this manuscript.

where $L = D^{-1/2}SD^{-1/2}$ is the normalized Laplacian, $S \in \mathbb{R}^{N\times N}$ is the similarity (also called affinity) matrix with ij-th entry $S_{ij} = s_{ij}$, $D = \text{diag}([\deg_1, \ldots, \deg_N])$ is the graph degree matrix, $\deg_i = \sum_{j=1}^{N} S_{ij}$ is the degree of the i-th vertex, $G \in \mathbb{R}^{N\times k}$ is the relaxed cluster indicator matrix and I_k is the $k \times k$ identity matrix. The minimizer of (1) is the matrix $G^\star$ containing any orthonormal basis of the k-th principal subspace of L. In other words, $G^\star$ can be found by computing the eigendecomposition of L and stacking the k dominant eigenvectors on the columns.

B. Reclustering

In order to obtain cluster indicators from the relaxed solution $G^\star$, an extra step called reclustering needs to be performed. Typically k-means is applied over the eigenvectors to obtain the final partitioning but other methods have also been discussed [11], [12]. The main motivation for using k-means relies on the fact that the clusters are very localized in the eigenvector representation. In ideal conditions, all data points in the same cluster are mapped to a single point in the space spanned by the eigenvectors with the k largest eigenvalues. Eigenvectors following this structure are called *piecewise constant* because points in the same cluster have the same constant value in the eigenvectors [13]. The NCUT minimization problem is motivated with insights in graph theory, however there exist other views of spectral clustering with different motivations (e.g. Markov random walks [13] and eigenvalue perturbation theory [14]). The random walks interpretation also leads to an eigenvalue decomposition but in this case of the random walks Laplacian $L_{\text{RW}} = D^{-1}S$. Both Laplacians can be used interchangeably because their eigenvectors are related by a simple matrix multiplication. Figure 1 shows an schematic representation of classical spectral clustering.

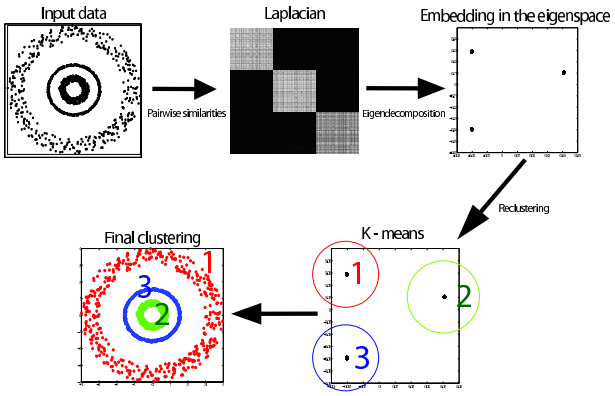

Fig. 1. Schematic of classical spectral clustering. In non-overlapping situations and when the similarity parameters are properly chosen, the clusters are represented in the eigenspace as single points. This different representation allows the use of simple methods for reclustering such as k-means.

C. Issues

One of the main issues in classical spectral clustering is the lack of a systematic parameter tuning scheme. In practice, a spectral clustering problem has two parameters: the number of clusters k and the similarity parameter. The Gaussian similarity (also called RBF kernel) $s_{ij} = K(x_i, x_j) = \exp(-||x_i - x_j||^2/\sigma^2)$ with parameter $\sigma \in \mathbb{R}^+$ is a popular choice. Even though there exist several schemes to determine k and the bandwidth σ, they all rely on heuristics which might not be appropriate in some cases. The clustering results depend greatly on the value of σ and assuming this parameter to be known is unrealistic. Another issue is related to the lack of clear out-of-sample mechanisms. The eigenvectors provide a partial clustering solution for the data points used during the training stage but the clustering results cannot be extended naturally to out-of-sample points. Extensions should rely on approximations such as the Nyström method [4].

III. KERNEL SPECTRAL CLUSTERING

This Section summarizes the multiway spectral clustering method introduced in [5]. This method corresponds to a weighted kernel PCA formulation providing a natural extension to out-of-sample data.

A. Primal - Dual Formulation

Given training data $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of clusters k, consider the following clustering model

$$e_i^{(l)} = {w^{(l)}}^T \varphi(x_i) + b_l, i = 1, \ldots, N, l = 1, \ldots, n_e, \quad (2)$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ is a mapping to a high-dimensional feature space and b_l are bias terms. The projections $e^{(l)} = [e_1^{(l)}; \ldots; e_N^{(l)}] = \Phi w^{(l)} + b_l 1_N, \Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$ represent the latent variables of a set of n_e binary clustering indicator vectors given by $\text{sign}(e_i^{(l)})$ which can be combined to obtain the final k clusters. The primal problem of kernel spectral clustering can be formulated as [5]:

$$\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2N}\sum_{l=1}^{k-1}\gamma_l {e^{(l)}}^T V e^{(l)} - \frac{1}{2}\sum_{l=1}^{k-1} {w^{(l)}}^T w^{(l)} \quad (3)$$

$$\text{such that } \begin{cases} e^{(1)} = \Phi w^{(1)} + b_1 1_N \\ \vdots \\ e^{(k-1)} = \Phi w^{(k-1)} + b_{k-1} 1_N \end{cases}$$

where γ_l are regularization parameters and $V = \text{diag}([v_1, \ldots, v_N]), v_i \in \mathbb{R}^+$ is a user-defined weighting matrix. The Lagrangian of this constrained optimization problem is:

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

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

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

for $l = 1, \ldots, k-1$, where $\alpha^{(l)} \in \mathbb{R}^N$ are the Lagrange multiplier vectors (dual variables). The dual form of the constrained optimization problem (3) is shown to be the eigenvalue problem [5]:

$$V M_V \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, \tag{5}$$

where $\lambda_l = N/\gamma_l$ are the ordered eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{k-1}$, M_V is the weighted centering matrix $M_V = I_N - 1/(1_N^T V 1_N) 1_N 1_N^T V$ and Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. The projected variables can now be written in terms of the dual variables:

$$e_i^{(l)} = {w^{(l)}}^T \varphi(x_i) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_i, x_j) + b_l, \tag{6}$$

$l = 1, \ldots, k-1$ where $e_i \in \mathbb{R}^{k-1}$ is the representation of x_i in the score variables space. The bias terms become $b_l = -1/(1_N^T V 1_N) 1_N^T V \Omega \alpha^{(l)}, l = 1, \ldots, k-1$.

B. Choice of the Weighting Matrix

It has been shown in [5] that when $V = I_N$, the dual problem (7) becomes the eigendecomposition of the centered kernel matrix hence, kernel PCA. When $V = D^{-1} = \text{diag}([1/d_1, \ldots, 1/d_N]), d_i = \sum_{j=1}^{N} K(x_i, x_j)$, the dual problem is related to spectral clustering. This relationship can be seen as a centered version of the random walks Laplacian [15], [13]. Thus, we are interested in the case when $V = D^{-1}$ leading to the eigenvalue problem

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

and centering matrix $M_D = I_N - 1/(1_N^T D^{-1} 1_N) 1_N 1_N^T D^{-1}$. The matrix $D^{-1} M_D \Omega$ has the following properties:

1) It is positive semidefinite with ordered eigenvalues $1 \geq \lambda_1 \geq \ldots \geq \lambda_N \geq 0$.
2) In the case of Ω representing the affinity matrix of a graph with k disconnected subgraphs, the matrix $D^{-1} M_D \Omega$ has $k-1$ eigenvectors $\alpha^{(l)}$ with eigenvalue $\lambda_l = 1$. These eigenvectors are cluster indicators (up to arbitrary rotations) of the underlying partitioning of the graph.
3) Every eigenvector $\alpha^{(l)}$ with associated non-zero eigenvalue has zero mean.

Property 2 is similar[2] to the case of the random walks Laplacian L_{RW}: the clusters are represented as single points in the subspace spanned by the $k-1$ dominant eigenvectors of $D^{-1} M_D \Omega$.. Property 3 is one of the KKT conditions and is essential for obtaining the cluster memberships since the single points representing the clusters will be localized in different orthants. The cluster decisions are taken as $\text{sign}(e^{(l)})$ and since $\text{sign}(e_i^{(l)}) = \text{sign}(\alpha_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$, points in different clusters will be in different orthants in the projection space. Due to this geometric property, there is no need to apply k-means as a final step. Instead, we use a binary encoding scheme that can be combined to form the final k clusters by forming a codebook. This scheme is inspired on LS-SVM multiclass formulations [6].

C. Projections, Collinearity and Model Selection

One of the main advantages of kernel spectral clustering is the possibility to extend the clustering model to out-of-sample points. Given a data point x, the projections can be written as:

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

$l = 1, \ldots, k-1$. The encoding representing x is obtained through $\text{sign}(\hat{e}^{(l)}(x))$ and the cluster membership is determined by comparing this encoding with the codewords in the codebook obtained during training. The cluster membership of x is assigned to the codeword with minimal Hamming distance (also known as ECOC decoding [16]).

When the eigenvectors are piecewise constant and the similarity parameters have been chosen properly, the clusters are represented in the projection space as lines[3]. This property can be used to look for clustering parameters such that out-of-sample data points in the same cluster are as collinear as possible in the projection space. The Balanced Line Fit (BLF) criterion introduced in [5] is an average measure of collinearity and balance of the clusters:

$$\text{BLF} = \frac{1}{k} \sum_{p=1}^{k} \eta \text{linefit}\big(Z_{\text{val}}^{(p)}, k\big) + (1-\eta) \text{balance}\big(Z_{\text{val}}^{(p)}, k\big) \tag{9}$$

where $Z_{\text{val}}^{(p)} \in \mathbb{R}^{|\mathcal{A}_p| \times (k-1)}$ are the projected variables of validation data assigned to the p-th cluster and η is a user-defined constant with $0 \leq \eta \leq 1$ which controls the importance of the linefit and the balance. The linefit ranges between 0 and 1, attaining its minimal value when the projections are spherically distributed and its maximal value when the projections are collinear. The balance index is the ratio between the number of elements in the smallest cluster and the number of elements in the largest cluster. For further details the reader is referred to [5]. Figure 2 illustrates the concept behind the BLF: the clusters form lines in the projection space with well-tuned RBF kernel parameters.

[2]The difference relies on the fact that 1_N is an eigenvector of L_{RW} but not of $D^{-1} M_D \Omega$.

[3]This holds when the out-of-sample data have been drawn from the same probability distribution as $\mathcal{D}$ and when the similarity tends to zero for points in different clusters. The latter is typically fulfilled with localized kernels such as the RBF kernel [5].

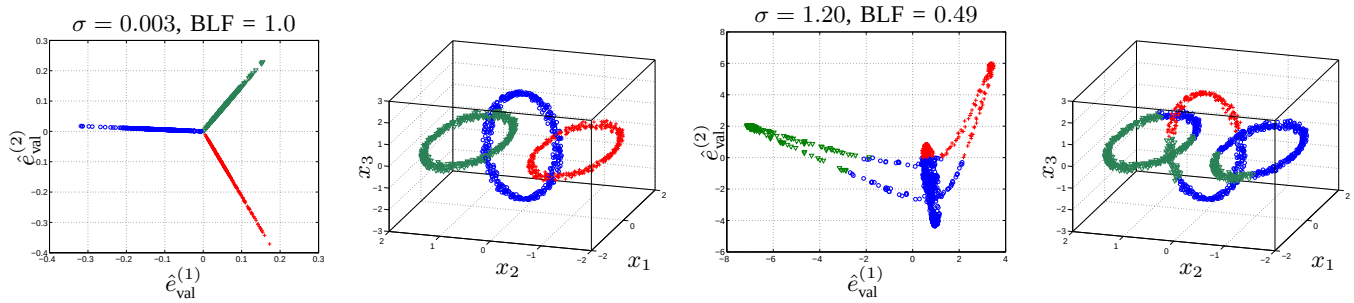

Fig. 2. Idea behind the BLF. For well-chosen kernel parameters, the clusters for out-of-sample data are collinear in the projection space. The BLF measures collinearity and balance. **From left to right:** Projection space for a good σ value, the BLF is maximal; Resulting clustering for $\sigma = 0.003$; Projection space for a wrongly-chosen kernel parameter; Corresponding clustering in case of $\sigma = 1.20$.

IV. New Notion: Out-of-Sample Eigenvectors

As discussed before, cluster decisions for out-of-sample data are taken on the projections onto the eigenvectors. In this space, the clusters form lines and each line is in a different orthant. Thus, applying k-means over the projections is not appropriate since this method is optimal only for spherical clusters. In this paper, we propose a new method for approximating eigenvectors for out-of-sample data.

A. From projections back to eigenvectors

Consider a set of N_{v} validation data points $\mathcal{D}^{\text{val}} = \{x^{\text{val}}_t\}_{t=1}^{N_{\text{v}}}$. The latent variables of the clustering model (projections) are:

$$\hat{e}^{(l)}_{\text{val}} = \Phi_{\text{val}} w^{(l)} + b_l 1_{N_{\text{v}}} = \Omega_{\text{val}} \alpha^{(l)} + b_l 1_{N_{\text{v}}}, \qquad (10)$$

$l = 1, \ldots, k-1$ where $\Omega_{\text{val}} = \Phi_{\text{val}} \Phi^T$ is the $N_{\text{v}} \times N$ validation kernel matrix with tj-th entry $\Omega_{\text{val},tj} = K(x^{\text{val}}_t, x_j), t = 1, \ldots, N_{\text{v}}, j = 1, \ldots, N$, $\Phi_{\text{val}} = [\varphi(x^{\text{val}}_1)^T; \ldots; \varphi(x^{\text{val}}_{N_{\text{v}}})^T]$ and $1_{N_{\text{v}}}$ is a vector of N_{v} ones. The second KKT condition in (4): $\alpha^{(l)} = (\gamma_l / N) D^{-1} e^{(l)}$ links the eigenvectors $\alpha^{(l)}$ and the projections for training data $e^{(l)}$. In the same spirit, it can also be used to compute *out-of-sample* eigenvectors $\hat{\alpha}^{(l)}_{\text{val}} \in \mathbb{R}^{N_{\text{v}}}$ by:

$$\hat{\alpha}^{(l)}_{\text{val}} = \frac{1}{\lambda_l} D^{-1}_{\text{val}} \hat{e}^{(l)}_{\text{val}}, l = 1, \ldots, k-1, \qquad (11)$$

where $D^{-1}_{\text{val}} = \text{diag}([1/\text{deg}(x^{\text{val}}_1), \ldots, 1/\text{deg}(x^{\text{val}}_{N_{\text{v}}})]) \in \mathbb{R}^{N_{\text{v}} \times N_{\text{v}}}$ is the inverse degree matrix for validation data and $\text{deg}(x) = \sum_{j=1}^{N} K(x, x_j)$ extends the concept of degree to out-of-sample data. The mean is removed to fulfill the third KKT condition: $1_N^T \alpha^{(l)} = 0$:

$$\hat{\alpha}^{(l)}_{\text{val}} \leftarrow \hat{\alpha}^{(l)}_{\text{val}} - \frac{1}{N_{\text{v}}} 1_{N_{\text{v}}} 1^T_{N_{\text{v}}} \hat{\alpha}^{(l)}_{\text{val}} \qquad (12)$$

and normalized to have norm 1:

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

The out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}$ can now be used as a localized representation of the grouping with the same properties as the original eigenvectors $\alpha^{(l)}$ for training data. Cluster decisions can still be taken based on orthants since this property is induced by the bias terms and the out-of-sample eigenvectors have the same sign as the projections. The algorithm for obtaining out-of-sample eigenvectors in kernel spectral clustering is shown in Algorithm 1.

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 $\mathcal{D}^{\text{val}} = \{x^{\text{val}}_t\}_{t=1}^{N_{\text{v}}}$.
Output: Partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, cluster codeset $\mathcal{C}$, out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}$.

1: Obtain the eigenvectors $\alpha^{(l)}, l = 1, \ldots, k-1$ corresponding to the largest $k-1$ eigenvalues by solving (7).
2: Binarize the eigenvectors: $\text{sign}(\alpha^{(l)}_i), i = 1, \ldots, N, l = 1, \ldots, k-1$, 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: $\mathcal{C} = \{c_p\}_{p=1}^{k}, c_p \in \{-1, 1\}^{k-1}$
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 $\hat{e}^{(l)}_{\text{val}}$ using (10).
6: Estimate the out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}$ using $\hat{e}^{(l)}_{\text{val}}$ and (11). Remove the mean and normalize using (12) and (13).
7: Binarize the out-of-sample eigenvectors $\text{sign}(\hat{\alpha}^{(l)}_{\text{val},t}), t = 1, \ldots, N_{\text{v}}, l = 1, \ldots, k-1$ and let $\text{sign}(\hat{\alpha}_{\text{val},t}) \in \{-1, 1\}^{k-1}$ be the encoding vector of x^{val}_t.
8: $\forall t$, assign x^{val}_t to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\text{sign}(\hat{\alpha}^{(l)}_{\text{val},t}), c_p)$.

B. Differences with the Nyström method

The Nyström method is a classical approach to compute numerical approximations of integral eigenvalue problems using a quadrature rule. This approach approximates the corresponding eigenfunction using a fixed small subset of points and then extrapolates the approximation to the full set of points. It has been used for speeding up kernel machines [17] and in the context of spectral clustering mainly for large-scale problems of image segmentation (large number of pixels) [4]. The idea is to approximate the full eigenvectors of the NCUT problem using only a reduced set of points. Typically random subsampling is used to determine the subset of points. The Nyström method requires a matrix inversion and a singular value decomposition to approximate the eigenvectors in spectral clustering. This differs from the proposed out-of-sample eigenvector methodology since our

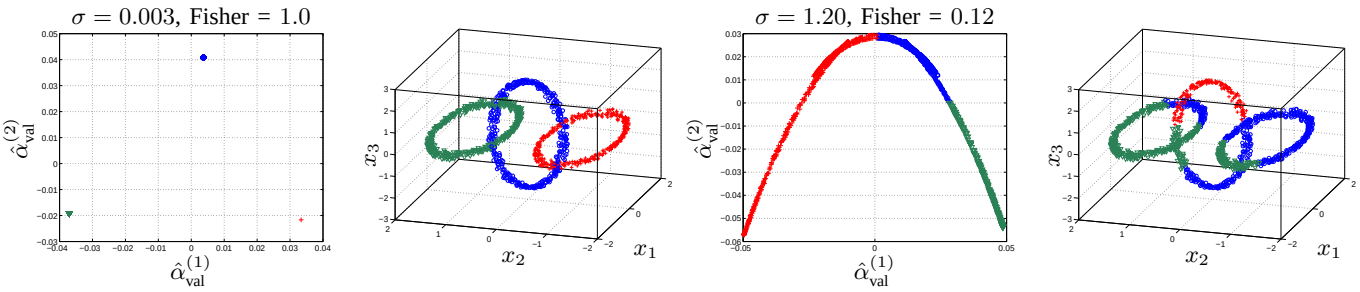

Fig. 3. Idea behind the out-of-sample eigenvectors and the Fisher criterion. For well-chosen kernel parameters, the clusters are localized in the out-of-sample eigenvector space. **From left to right:** Estimated eigenvector space for a good sigma value, the Fisher criterion is maximal; Corresponding clustering for $\sigma = 0.003$; Estimated eigenvector space for a wrongly-chosen kernel parameter; Corresponding clustering in case of $\sigma = 1.2$.

method is based on insights from the primal - dual representations and the KKT conditions. Morever, the proposed method has lower computational cost since it only requires matrix multiplications.

V. Model Selection Based on the Fisher Criterion

The estimated out-of-sample eigenvectors $\hat{\alpha}^{(l)}_{\text{val}}$ allows us to use measures of cluster distortion (rather than collinearity) for selecting the clustering parameters. In ideal conditions, when there is no cluster overlap and the kernel parameters are chosen such that $K(x_i, x_j) \simeq 0$ when x_i and x_j are in different clusters, the eigenvectors and the out-of-sample eigenvectors will appear localized. The classical Fisher criterion [7] popular in classification problems can be adapted here to select k and the RBF kernel parameter σ such that $\hat{\alpha}^{(l)}_{\text{val}}$ represent the clusters as localized as possible. Given $\hat{\alpha}^{(l)}_{\text{val}}$ which is considered as a new representation of the validation data $\mathcal{D}^{\text{val}}$ and the clustering $\{\mathcal{A}_p\}_{p=1}^k$ found by decoding, 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 $\hat{\alpha}_{\text{val},t} \in \mathbb{R}^{k-1}$ is the representation of $x^{\text{val}}_t, t = 1, \ldots, N_{\text{v}}$,

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

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 weights are given by ζ_p which can be chosen as $\zeta_p = |\mathcal{A}_p|/N_{\text{v}}$ to give preference to balanced clusters. Trace(S_W) is an average measure of cluster compactness with value 0 when the out-of-sample eigenvectors are perfectly localized (piecewise constant). Thus, the idea is to look for clustering parameters such that trace(S_B) is maximized and trace(S_W) is minimized. This can be accomplished by maximizing the Fisher criterion [7] with respect to k and the RBF kernel parameter σ:

$$\max_{k,\sigma} \frac{\text{trace}(S_B)}{\text{trace}(S_W + S_B)}. \quad (16)$$

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.

VI. 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 adjusted Rand index (ARI) [18] is used as a performance measure when comparing two clustering results. The ARI equals 1 when the cluster memberships agree completely.

A. Toy data

The first toy data corresponds to 10 Gaussian clusters in $2D$. The data have been created such that there are also other good clusterings rather than only 10. The total number of points is $N_{\text{full}} = 2,000$ and the dataset is split into training ($N = 400$), validation ($N_{\text{v}} = 1,200$) and the remaining data points for test. Figure 4 shows the tuning results after applying the proposed model selection criterion. The number of clusters k and the RBF kernel parameter σ can be determined from these plots. The Fisher criterion is able to detect the multiscale nature of the datasets indicating strong cluster structures from $k = 2$ until $k = 10$. Figure 5 shows the clustering results for several parameter pairs (k, σ) that display a high average Fisher criterion value. The *zero* cluster indicates out-of-sample data points mapped close to the origin in the projections space. Typically data points in the zero cluster are very different from the points observed during the training stage. The results show a natural partitioning of the data with good generalization capabilities.

The second toy data consists of 5 clusters in $2D$. The data show a nonlinear structure and classical methods such

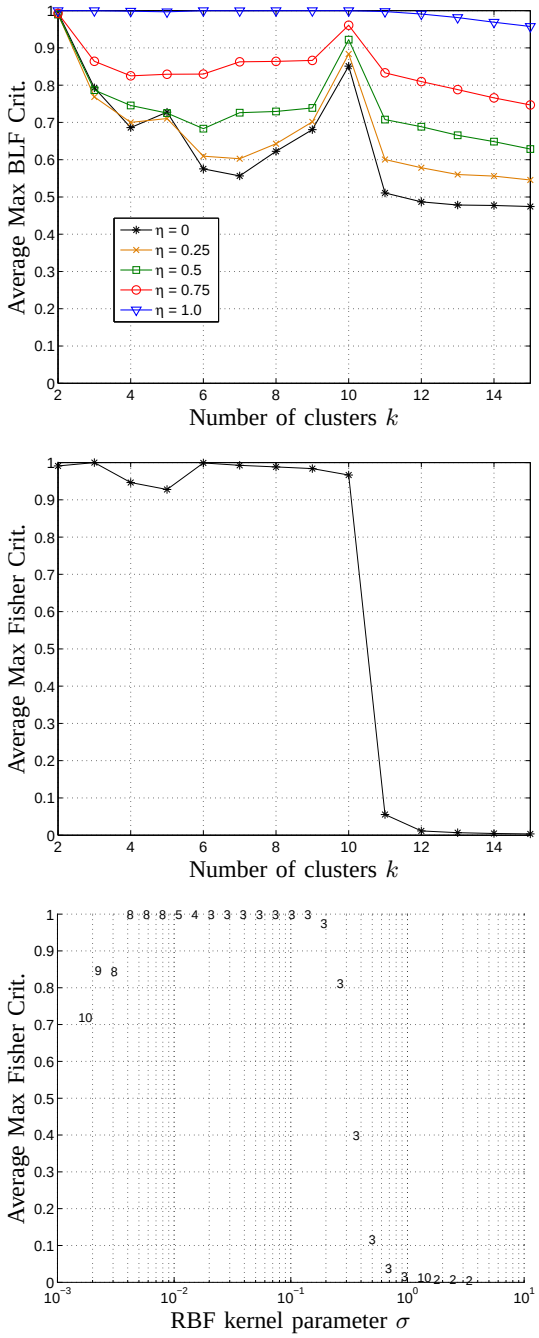

Fig. 4. Toy Example 1 - Model selection. Comparison of the BLF criterion and the proposed Fisher criterion. **Top:** Average maximal BLF criterion for a fixed k with several η values. The BLF selects 2 and 10 clusters. **Center:** Average maximal Fisher criterion for a fixed k. The proposed model selection criterion detects the multiscale nature of the data and suggest several possibilities for k. **Bottom:** Mode (most frequent value) of k for a fixed σ using the Fisher criterion.

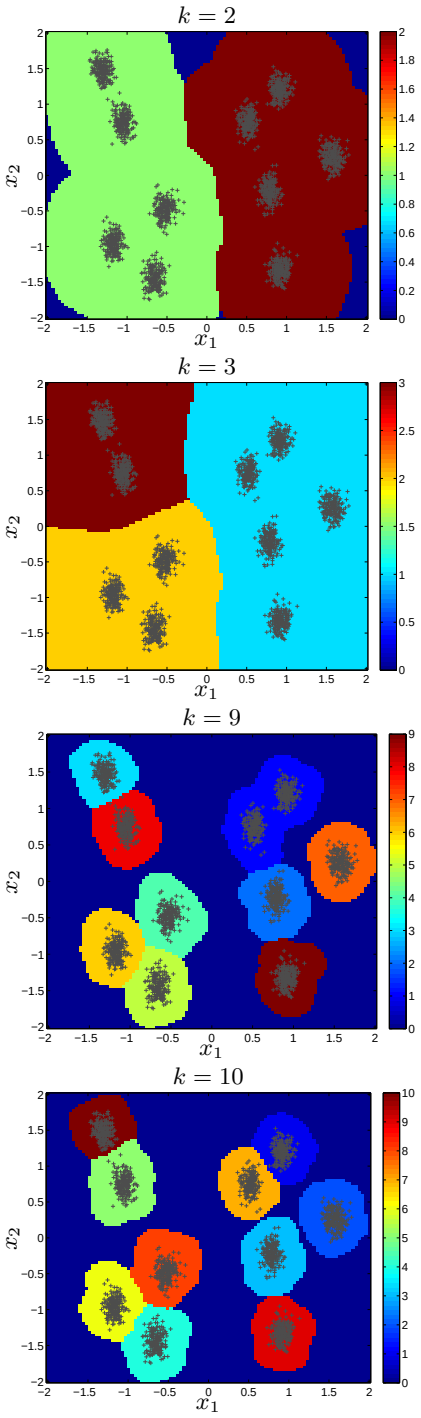

Fig. 5. Toy Example 1 - Clustering results. Four different partitionings of the data that showed a high average maximal Fisher criterion. The *zero* cluster indicates points very close to the origin in the projection space leading to predicted cluster membership with very low confidence. The results show good generalization capabilities.

as k-means fail in this type of datasets. Figure 6 shows the clustering results after model selection using the Fisher criterion. The proposed method is compared with the Nyström approximation for spectral clustering discussed in [4] showing an improved performance with less computational effort.

B. Image Segmentation

Image segmentation is a classical application for spectral clustering due to the large number of data points (pixels) leading to eigendecomposition of big matrices. In this experiment, we used an image from the Berkeley image database

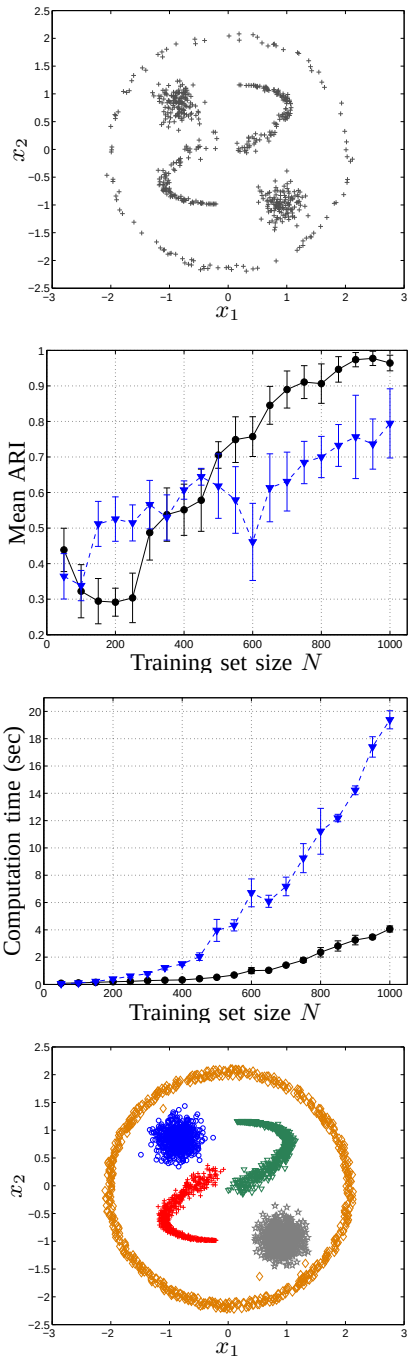

Fig. 6. Toy Example 2 - Comparison with the Nyström method [4] when the clustering parameters are known. Proposed method in solid black (circles) and [4] in dashed blue (triangles). **From top to bottom:** Training dataset; Agreement measured in terms of the ARI with respect to underlying grouping of the data and an increasing number of training points; Computation times in seconds of the proposed method and the Nyström method; Final clustering of the complete dataset.

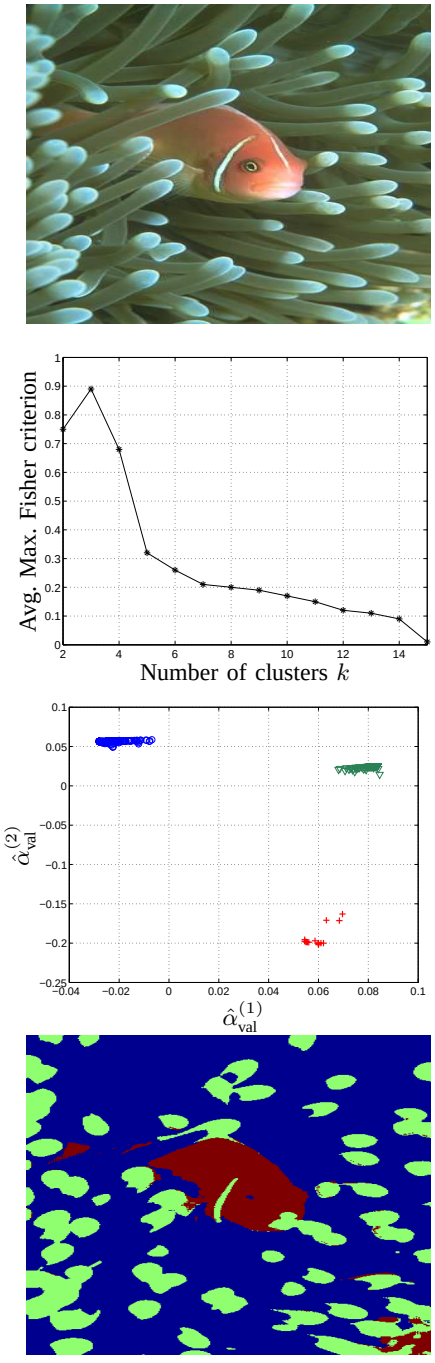

Fig. 7. Image segmentation. **From top to bottom:** Full 321×481 pixels image; Model selection using the Fisher criterion on validation data ($20,000$ pixels). The chosen parameters are $\sigma_\chi = 0.02$ which maximizes the criterion for $k = 3$; Visualization of the out-of-sample eigenvector space. The clusters appear localized; Clustering results shown as a segment-label image. Different colors indicate different clusters.

[19]. 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 $h^{(j)}$ [20] $\chi^2_{ij} = 0.5 \sum_{b=1}^{B} (h_b^{(i)} - h_b^{(j)})^2 / (h_b^{(i)} + h_b^{(j)})$, 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 $\sigma_\chi \in \mathbb{R}^+$ 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

800 randomly selected pixel histograms and $20,000$ pixel histograms were used as validation. Figure 7 shows the model selection and clustering results. The Fisher criterion found strong cluster structures for $k = 2, 3, 4$ peaking at $k = 3$. The χ^2 kernel parameter that maximized the criterion for $k = 3$ was $\sigma_\chi = 0.02$. The out-of-sample eigenvector space shows a localized behavior indicating good clustering results which are confirmed with the visually appealing segment-label image. The average computation time for training the model and computing the out-of-sample eigenvectors was 6.78 ± 1.3 seconds.

VII. Conclusions

A method to estimate out-of-sample eigenvectors is presented. The method fits in the kernel spectral clustering framework and is based on insights from the primal and dual representations. A model selection scheme based on the Fisher criterion is also presented. The proposed criterion can be used to obtain the underlying number of clusters present on the data and the kernel parameters by using training, validation and test stages. The experimental results on difficult toy datasets and an image segmentation problem confirm the applicability of the proposed method and model selection criterion.

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.

References

[1] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

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

[3] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[4] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[5] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, February 2010.

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

[7] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[9] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory," *Czech. Math. J.*, vol. 25, no. 100, pp. 619–633, 1975.

[10] M. Gu, H. Zha, C. Ding, X. He, and H. Simon, "Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering," Penn. State Univ, Computer Science and Engineering., Technical Report, 2001.

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

[12] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

[13] M. Meila and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001.

[14] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.

[15] M. Meila and J. Shi, "A random walks view of spectral segmentation," in *Artificial Intelligence and Statistics AISTATS*, 2001.

[16] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Machine Learning Research*, vol. 2, pp. 263–286, 1995.

[17] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[18] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, pp. 193–218, 1985.

[19] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "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*, vol. 2, 2001, pp. 416–423.

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