Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA

Carlos Alzate and Johan A.K. Suykens, Senior Member, IEEE

Abstract—A new formulation for multiway spectral clustering is proposed. This method corresponds to a weighted kernel principal component analysis (PCA) approach based on primal-dual least-squares support vector machine (LS-SVM) formulations. The formulation allows the extension to out-of-sample points. In this way, the proposed clustering model can be trained, validated, and tested. The clustering information is contained on the eigendecomposition of a modified similarity matrix derived from the data. This eigenvalue problem corresponds to the dual solution of a primal optimization problem formulated in a high-dimensional feature space. A model selection criterion called the Balanced Line Fit (BLF) is also proposed. This criterion is based on the out-of-sample extension and exploits the structure of the eigenvectors and the corresponding projections when the clusters are well formed. The BLF criterion can be used to obtain clustering parameters in a learning framework. Experimental results with difficult toy problems and image segmentation show improved performance in terms of generalization to new samples and computation times.

Index Terms—Spectral clustering, kernel principal component analysis, out-of-sample extensions, model selection.

1 INTRODUCTION

Spectral clustering comprises a family of clustering methods that make use of the eigenvectors of some normalized affinity matrix derived from the data to group points that are similar. These techniques are known to perform well in cases where classical clustering methods such as k-means and linkage absolutely fail. Many spectral clustering applications can be found in fields like machine learning, computer vision, and data analysis. Starting from graph theory, this family of algorithms [1], [2], [3], [4], [5], [6] is formulated as relaxations of graph partitioning problems that are generally NP-complete. These relaxations take the form of eigenvalue problems involving a normalized affinity matrix containing pairwise similarities. When only two groups are required, the relaxed solution corresponds to a particular eigenvector. The cluster information can then be extracted by binarizing the relaxed solution. However, the clustering problem complicates when more than two clusters are required because converting the relaxed solutions to cluster information is not straightforward. Typical approaches to tackle this issue are recursive cuts and reclustering, which were discussed in [1]. Recursive binary cuts are not optimal because only one eigenvector is used when other eigenvectors may also contain cluster information and it is not clear when to stop cutting. The reclustering approach consists of applying k-means over the eigenvectors. This approach works only when the cluster information contained in the eigenvectors has a spherical structure. More advanced techniques are based on minimizing a metric between the relaxed solutions and the set of allowed cluster indicators [7], [8] and on finding peaks and valleys of a criterion that measures cluster overlap [9].

One of the main drawbacks of spectral clustering is the fact that the clustering model is defined only for training data with no clear extension to out-of-sample points. The methods described in [10], [11] allow extensions to new points by approximating the implicit eigenfunction using the Nyström method [12], [13]. The underlying clustering model is usually not known and the parameter selection is done in a heuristic way.

Kernel principal component analysis (PCA) [14] corresponds to an unsupervised learning technique useful for nonlinear feature extraction, denoising, and dimensionality reduction. This method is a nonlinear extension of PCA by using positive definite kernels. The objective is to find projected variables in a kernel-induced feature space with maximal variance. Links between spectral clustering and kernel PCA have been discussed in [15], [16]. The methods show that some forms of spectral clustering can be seen as a special case of kernel PCA. The method described in [16] shows that classical binary spectral clustering such as the normalized cut [1], the NJW algorithm [3], and the random walks model [6], are cases of weighted kernel PCA with different weighting matrices.

In this paper, we propose a multiway spectral clustering model based on a weighted kernel principal component analysis scheme [16], [17]. The formulation is based on [16], which is a binary clustering technique, and extended to the multiway clustering case with encoding and decoding schemes. The formulation fits into the least-squares support
vector machines (LS-SVMs) framework [18], [19] by considering weighted versions. The proposed approach is cast in a constrained optimization framework with primal and dual insights. This interpretation allows the clustering model to be extended to out-of-sample data points without relying on approximation techniques such as the Nyström method. The eigenvectors of a modified similarity matrix derived from the data are shown to be the dual solutions to a primal optimization problem formulated in a high-dimensional feature space. These solutions contain clustering information and show a special structure when the clusters are well formed. The out-of-sample extension is obtained through projections onto these eigenvector solutions. We also propose a model selection criterion called the BLF. This criterion exploits the structure of the eigenvectors and the corresponding projections and can be used to obtain model parameters.

This paper is organized as follows: Section 2 contains a review of existing spectral clustering techniques. Section 3 contains a description of the existing weighted kernel PCA approach to binary spectral clustering. In Section 4, we propose the multiway spectral clustering model. Section 5 contains the out-of-sample extension, together with an algorithm. In Section 6, we discuss about model selection and propose the BLF criterion. Section 7 contains the empirical results, and in Section 8, we give conclusions.

2 SPECTRAL CLUSTERING

2.1 Graph Theory

Given a set of $N$ data points $\{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^d$ and some similarity measure $s_{ij} \geq 0$ between each pair of points $x_i$ and $x_j$, an intuitive form of representing the data is using a graph $G = (V, E)$. The vertices $V$ represent the data points and the edge $e_{ij} \in E$ between vertices $v_i, v_j$ has a weight determined by $s_{ij}$. If the similarity measure is symmetric, then the graph is undirected. The affinity matrix of the graph is the matrix $A$ such that $A_{ij} = s_{ij}$. The degree of vertex $v_i$ is the sum of all the vertex weights adjacent to $v_i$ and is defined as $\deg_i = \sum_{j=1}^{N} s_{ij}$, the degree matrix $D$ is a diagonal matrix containing the vertex degrees $\deg_1, \ldots, \deg_N$ on the diagonal.

A basic problem in graph theory is the graph bipartitioning problem that is to separate the graph $G$ into two disjoint sets $A, B$ based on a cut criterion. The resulting sets should be disjoint: $A \cap B = \emptyset$ and $A \cup B = V$. This problem has been extensively studied [2], [20] and several cut criteria have been proposed [1], [4], [20], [21]. The normalized cut NCut [1] is a common bipartitioning criterion and is defined as $\text{NCut}(A, B) = \text{cut}(A, B) / \text{Vol}(A) + \text{cut}(A, B) / \text{Vol}(B)$, where $\text{Vol}(A) = \sum_{i \in A} d_i$ denotes the volume of the set $A$. The NCut can be extended to the general case in which the problem is to separate the graph into $k$ disjoint sets: $\text{NCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \text{cut}(A_i, \overline{A_i}) / \text{Vol}(A_i)$, where $\overline{A_i}$ denotes the complement of $A_i$.

2.2 The Normalized Cut Relaxation

The NCut algorithm has become a very popular technique for spectral clustering and image segmentation. Its approximate solution follows from the second smallest eigenvector (also known as the Fiedler vector) of a generalized eigenvalue problem. The problem of minimizing the NCut can be written (see [1]) as

$$\min_{q} \text{NCut}(A, B) = \frac{q^T L q}{q^T D q},$$

such that $\{q \in \{-1, 1\}^N, q^T D_1 q = 0\}$,

where $L = D - S$ is the unnormalized graph Laplacian and $c = \text{Vol}(A) / \text{Vol}(B)$. Unfortunately, minimizing the NCut is NP-complete [1]. However, an approximate solution can be found by relaxing the discrete constraint on $q$. If $q$ can take real values, then the objective in (1) corresponds to the Rayleigh quotient of the generalized eigenvalue problem $L q = \nu D q$, where $q \in \mathbb{R}^N$. The remaining constraint $q^T D_1 q = 0$ is automatically satisfied in the generalized eigenproblem. Therefore, the relaxed solution to the NCut is the Fiedler vector.

2.3 Markov Random Walks

A Markov random walks view of spectral clustering was discussed in [6], [22]. A random walk on a graph consists of random jumps from vertex to vertex. This interpretation showed that many properties of spectral clustering methods can be expressed in terms of a stochastic transition matrix $P$ obtained by normalizing the affinity matrix such that its rows sum to 1. The $ij$th entry of $P$ represents the probability of moving from node $i$ to node $j$ in one step. This transition matrix can be defined as $P = D^{-1} S$. The corresponding eigenvalue problem becomes $Pr = \xi r$.

The eigenvalues of $P$ are $1 = \xi_1 \geq \xi_2 \geq \ldots \geq \xi_N$, with corresponding eigenvectors $1_1 = r(1), r(2), \ldots, r(N)$. Consider the general eigenvalue problem of the NCut: $L q = \nu D q$, then premultiplying by $D^{-1}$ leads to $(I_N - P) q = q$. This shows that the eigenvectors of $P$ are identical to the eigenvectors of the NCut algorithm and the eigenvalues of $P$ are $\xi_1 = 1 - \nu$, $i = 1, \ldots, N$. Minimizing the NCut can be interpreted as finding a partition of the graph in such a way that the random walk remains most of the time in the same cluster with few jumps to other clusters.

2.4 The $k$-way NCut Relaxation

Consider $f_p \in \{0, 1\}^N$ as the cluster indicator vector for the $p$th cluster such that $f_p$ has a 1 in the entries corresponding to the data points in the $p$th cluster. The cluster indicator matrix becomes $F = [f_1, \ldots, f_k] \in \{0, 1\}^{N \times k}$. Defining $g_p = D^{1/2} f_p / \|D^{1/2} f_p\|_2$, $p = 1, \ldots, k$, and $G = [g_1, \ldots, g_k]$ leads to the $k$-way NCut criterion:

$$\text{NCut}(A_1, \ldots, A_k) = k - \text{tr}(G^T L G),$$

where $\hat{L}$ is the normalized Laplacian defined as $\hat{L} = D^{-1/2} S D^{-1/2}$ and $G^T G = I_k$. The matrix $G$ that minimizes the $k$-way NCut can be found by maximizing the trace of $G^T L G$:

$$\arg\min_{G} \text{NCut}(A_1, \ldots, A_k) = \arg\max_{G} \text{Tr}(G^T \hat{L} G),$$

such that $G^T G = I_k$. 

This problem is a semidefinite program (SDP) that can be solved using interior-point algorithms.

Authorized licensed use limited to: Katholieke Universiteit Leuven. Downloaded on June 01,2010 at 09:21:40 UTC from IEEE Xplore. Restrictions apply.
Relaxing the discrete constraint and allowing $G$ to be a real-valued matrix leads to the $k$-way $N\text{Cut}$ relaxtion [23]:

$$\arg\min_G N\text{Cut}(A_1, \ldots, A_k) = \arg\max_G \text{Tr}(\tilde{G}^T \tilde{L} \tilde{G}),$$

$$\text{such that } \tilde{G}^T \tilde{G} = I_k.$$  

This maximization problem is a special case of Fan’s theorem [24]. The solution to this relaxed problem is $\tilde{G} = AR_1$, where $A = [a(1), \ldots, a(k)] \in \mathbb{R}^{N \times k}$ is an orthonormal basis of the $k$th principal subspace of $\tilde{L}$ and $R_1 \in \mathbb{R}^{k \times k}$ is an arbitrary orthogonal matrix [7], [25]. Considering the eigenvalue problem of the normalized Laplacian $D^{-1/2}SD^{-1/2}a = \zeta a$, premultiplying by $D^{-1/2}$ leads to $P \tilde{u} = \zeta \tilde{u}$, with $u = D^{-1/2}a$. Thus, the relation between the $k$-way $N\text{Cut}$ and the random walks model is $\zeta = \xi, a = D^{1/2}r$.

### 2.5 Perturbation Analysis and Piecewise Constant Eigenvectors

In the ideal case of a graph with $k$ disconnected components, the corresponding affinity matrix shows a block diagonal structure composed of $k$ blocks if the clusters are ordered. The transition matrix $P$ in this case will have $k$ eigenvalues equal to 1 and the remaining eigenvalues will be less than 1. The eigenvectors corresponding to these unitary eigenvalues are the indicator functions to the respective disconnected components and the clustering is trivial because each cluster is represented as a single point in $\mathbb{R}^k$. This also corresponds to the piecewise constant property of eigenvectors discussed in [6], [25], where the spectral clustering problem is cast in a probabilistic framework using block stochastic matrices.

**Definition 1** [6]. A piecewise constant vector $\alpha = [\alpha_1, \ldots, \alpha_N] \in \mathbb{R}^N$ relative to a partition of $\Delta = \{A_p\}_{p=1}^k$ and a data set $D = \{x_i\}_{i=1}^N$ is defined as

$$\alpha_j = c_p, \text{ for all } x_j \in A_p, p = 1, \ldots, k,$$

where $c_p$ is a constant value corresponding to the $p$th cluster.

Perturbation analysis is the study of how the eigenvalues and eigenvectors of a symmetric matrix $B$ change if an additive perturbation $E$ is considered. The perturbed matrix is denoted by $\tilde{B} \equiv B + E$. The Davis-Kahan theorem [26] states that the distance between the top $k$ eigenvalues and eigenvectors of the perturbed matrix $\tilde{B}$ and the top $k$ eigenvalues and eigenvectors of $B$ is bounded by $\|E\|_F/\delta$, where $\|\cdot\|_F$ denotes the Frobenius norm, $\delta = \tilde{\eta}_k - \tilde{\eta}_{k+1}$ corresponds to the additive eigengap, and $\tilde{\eta}_k$ is the $k$th ordered eigenvalue of $\tilde{B}$. Therefore, if the perturbation is small and the eigengap is large enough, the perturbed eigenvectors will be approximately piecewise constant, and consequently, easy to cluster [27]. This represents one of the fundamental ideas of spectral clustering.

### 2.6 From Eigenvectors to Clusters

The relaxed solutions provided by $k$-way spectral clustering methods are real-valued and do not provide cluster indicator matrices. Hence, it is necessary to convert these solutions in order to obtain indicators of the partitioning. In [1], two methods were proposed. The first one is based on applying binary cuts in a recursive way. The current partition is subsequently divided if the $N\text{Cut}$ is below some prespecified value. The process is repeated until $k$ clusters have been found. This method is slow and not optimal due to the fact that only the second smallest eigenvector is used when other eigenvectors may contain useful information. The second method called reclustering is one of the most popular approaches for $k$-way spectral clustering and consists of computing the top $k$ eigenvectors of the $k$-way $N\text{Cut}$ problem, and then, applying $k$-means onto them. This works for piecewise constant eigenvectors because every cluster $A_i \in \Delta$ is mapped to a unique point in $\mathbb{R}^k$. A different approach, called rounding, was introduced in [28]. It is based on the minimization of a metric between the relaxed solution and the set of discrete allowed solutions. This optimization problem can be solved either with respect to the partition resulting in a weighted $k$-means algorithm [29] or with respect to the similarity matrix yielding an algorithm for learning $W$ given a particular partition. In [9], a method called linearized cluster assignment was discussed. This method transforms the problem of obtaining clusters from the eigenvectors into finding peaks and valleys of a 1D quantity called cluster crossing.

### 3 Kernel PCA and LS-SVMs

#### 3.1 Classical Kernel PCA Formulation

A kernel-based algorithm for nonlinear PCA was described in [14]. This algorithm performs linear PCA in a kernel-induced feature space nonlinearly related to the original input space. Given a set of $N$ data points $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$, kernel PCA finds directions in which the projected variables $w^T \varphi(x_i), i = 1, \ldots, N$ have maximal variance. Here, $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^k$ corresponds to the mapping to a high-dimensional feature space $F$ of dimension $d_0$ (which can be infinite dimensional). Assuming zero-mean-mapped data $\sum_{i=1}^N \varphi(x_i) = 0$, the covariance matrix in the feature space becomes $C \approx (1/N) \sum_{i=1}^N \varphi(x_i) \varphi(x_i)^T$ with eigendecomposition $Cw = \lambda w$. The computation of $\varphi(\cdot)$ and $C$ is complicated due to the high dimensionality of $F$. However, by making use of the representer theorem [14], it is possible to avoid dealing with the mapped data explicitly. A positive definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a positive definite function that corresponds to a dot product in a high-dimensional feature space $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. The kernel PCA problem can be solved by means of the eigendecomposition of the kernel matrix $\Omega \in \mathbb{R}^{N \times N}$ which contains pairwise evaluations of the kernel function $\Omega_{ij} = K(x_i, x_j), i, j = 1, \ldots, N$. Typically, the data are preprocessed in the feature space by removing the mean. This step is accomplished by centering the kernel matrix: $\Omega = M\Omega M_*$, where $M_0 = I_N - (1/N)1_N1_N^T$ is the centering matrix.

The kernel PCA eigenvalue problem becomes $\Omega \beta = \lambda \beta$. The resulting eigenvectors have to be normalized in $F$ by requiring $\|\beta\|_2 = 1$. The LS-SVM approach to kernel PCA was introduced in [19]. This approach showed that kernel PCA is the dual
solution to a primal optimization problem formulated in a kernel-induced feature space. The underlying loss function associated with kernel PCA was shown to be $L_2$. Given the training set $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$ sampled i.i.d. from an underlying distribution $P$, the LS-SVM approach to kernel PCA is formulated in the primal as:

$$
\min_{w, e, b} J_p(w, e) = \frac{\gamma}{2N} \sum_{i=1}^N e_i^2 - \frac{1}{2} w^T w,
$$

such that

$$
e_i = w^T \varphi(x_i) + b, \quad i = 1, \ldots, N,
$$

where $b$ is a bias term.

**Proposition 1** [17], [19]. Given a positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ with $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, the Karush-Kuhn-Tucker (KKT) conditions of the Lagrangian of (5) are satisfied by each eigenvector $\alpha$ of the dual problem $\Omega \alpha = \lambda \alpha$, where $\Omega$ denotes the centered kernel matrix and $\lambda = N/\gamma$.

**Remark 1** [19]. Note that including a bias term in (5) leads to centering the kernel matrix. The mapped points in the feature space have zero mean. The bias term becomes

$$
b = -\frac{1}{N} 1_N^T \Omega \alpha.
$$

### 3.3 Weighted Kernel PCA

A weighted kernel PCA formulation based on the LS-SVM framework was discussed in [16], [17]. This formulation has been used for extending kernel PCA to general loss functions in order to achieve sparseness and robustness in an efficient way [17]. A different kind of weighting was used in [16] to link several binary spectral clustering methods. The weighted kernel PCA framework has also been used to provide links with spectral clustering techniques such as the NCut [1], the approximate kernel alignment [30], [31], the NJW algorithm [3], and the random walks [6], emphasizing primal-dual insights in addition to out-of-sample extensions for binary clustering [16]. Table 1 shows the relation between weighted kernel PCA and some spectral clustering methods. The last column indicates which eigenvector of (7) is the relaxed solution to the original spectral method given $V$.

### 3.4 Links with Binary Spectral Clustering Methods

The weighted kernel PCA framework has also been used to provide links with spectral clustering techniques such as the NCut [1], the approximate kernel alignment [30], [31], the NJW algorithm [3], and the random walks [6], emphasizing primal-dual insights in addition to out-of-sample extensions for binary clustering [16]. Table 1 shows the relation between weighted kernel PCA and some spectral clustering methods. The last column indicates which eigenvector of (7) is the relaxed solution to the original spectral method given $V$.

### 4 Multiway Formulation to Spectral Clustering

The multiclass problem in classification consists of obtaining $n_c$ classes by using a set of $n_y$ binary classification problems. The number of additional classifiers is an important issue and typically corresponds to different encoding/decoding schemes. A unique codeword $y_l \in \{-1, 1\}^{n_y}$, $l = 1, \ldots, n_c$, is assigned to each class. The one-versus-all coding scheme consists of setting $n_c$ equal to the number of classes $n_c$ and making binary decisions between each class and all the other classes. In one-versus-one, the number of binary classifiers is set to $n_c(n_c - 1)/2$ and the discrimination is made for each possible pair of classes. The number of one-versus-one classifiers is usually higher than the number of classifiers in the one-versus-all case, but the resulting decision boundaries are typically simpler. Another encoding scheme is to use $n_y$ problems to encode up to $2^{n_y}$ classes. The error-correcting output code (ECOC) [32] approach usually introduces redundancy ($n_y > n_c$) and a careful design of codewords (e.g., good Hamming distance between codewords, codewords motivated by known features). In order to classify a new point, the

<table>
<thead>
<tr>
<th>Method</th>
<th>Original Problem</th>
<th>$V$</th>
<th>Relaxed Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>$\Omega g = \lambda q$</td>
<td>$I_N$</td>
<td>$N \alpha$</td>
</tr>
<tr>
<td>NCut</td>
<td>$L_q = \lambda D_q$</td>
<td>$D^{-1}$</td>
<td>$\alpha(2)$</td>
</tr>
<tr>
<td>Random walks</td>
<td>$D^{-1}Wq = \lambda q$</td>
<td>$D^{-1}$</td>
<td>$\alpha(2)$</td>
</tr>
<tr>
<td>NJW</td>
<td>$D^{-1}W D^{-1/2} q = \lambda q$</td>
<td>$D^{-1}$</td>
<td>$D^{1/2}\alpha(2)$</td>
</tr>
</tbody>
</table>
output variables are then compared with the set of code-
words and the new point is assigned to the class of the
closest codeword in terms of Hamming distance. This
scheme has shown good results in terms of robustness
against some errors induced by noise.

In this paper, we extend the weighted kernel PCA model
for binary spectral clustering to the k-way case by using
some concepts of multiclass theory and introducing addi-
tional score variables as constraints. This extension is
inspired in the multiclass formulation for LS-SVM models
discussed in [18]. Introducing additional regularization
constants, score variables, bias terms, and using the inverse
of the degree matrix as the weighting matrix leads to the
following primal problem:

$$\min_{u^{(l)}, e^{(l)}, b_l} \frac{1}{2N} \sum_{l=1}^{n} \gamma_l e^{(l)^T} D^{-1} e^{(l)} - \frac{1}{2} \sum_{l=1}^{n} u^{(l)^T} w^{(l)}$$

subject to

$$e^{(l)} = \Phi u^{(l)} + b_l 1_N,$$

for some positive definite kernel function $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, regularization
constants $\gamma_l \in \mathbb{R}^+$, $l = 1, \ldots, n_e$, and the inverse of the degree
matrix $D^{-1}$ which is diagonal with positive elements, then the
KKT optimality conditions of the Lagrangian of (8) are
satisfied by the following eigenvalue problem:

$$D^{-1} M_D \alpha_l = \lambda_l \alpha_l,$$

where $\lambda_l = N/\gamma_l$, $l = 1, \ldots, n_e$, and

$$M_D = I_N - \frac{1}{k} D^{-1} \frac{1}{N} D^{-1}.$$

Proof. Consider the Lagrangian of the problem (8):

$$\mathcal{L}(u^{(l)}, e^{(l)}, b_l; \alpha_l) = \frac{1}{2} \sum_{l=1}^{n} \gamma_l e^{(l)^T} D^{-1} e^{(l)} - \frac{1}{2} \sum_{l=1}^{n} u^{(l)^T} w^{(l)} - \sum_{l=1}^{n} \alpha_l^T (e^{(l)} - \Phi u^{(l)} - b_l 1_N),$$

with KKT optimality conditions:

$$\begin{align*}
\frac{\partial \mathcal{L}}{\partial u^{(l)}} &= 0 \Rightarrow u^{(l)} = \Phi^T \alpha_l, \\
\frac{\partial \mathcal{L}}{\partial e^{(l)}} &= 0 \Rightarrow e^{(l)} = \frac{\gamma_l}{N} D^{-1} e^{(l)}, \\
\frac{\partial \mathcal{L}}{\partial b_l} &= 0 \Rightarrow 1_N^T \alpha_l = 0, \\
\frac{\partial \mathcal{L}}{\partial \alpha_l} &= 0 \Rightarrow e^{(l)} = \Phi u^{(l)} + b_l 1_N,
\end{align*}$$

for $l = 1, \ldots, n_e$. The bias terms become:

$$b_l = \frac{1}{k} I_N^T D^{-1} \alpha_l, l = 1, \ldots, n_e.$$
corresponds to the ECOC decoding procedure.

indicators with the codebook and selecting the nearest
projection of the mapped out-of-sample points onto
extension is based on the score variables which correspond
approximating the underlying eigenfunction. Our proposed
problem of obtaining clusters from the out-of-sample
are more suited to data visualization than to clustering since
PCA and manifold learning was presented in [15], where

The eigenvectors \( \alpha^{(l)} \) are piecewise constant; therefore,
\( \alpha^{(l)} = c_p^{(l)} \) for all \( x_j \in A_p \), \( l = 1, \ldots, k - 1, \ p = 1, \ldots, k \),
where \( c_p^{(l)} \) is the constant value corresponding to the \( p \)th
cluster in the \( l \)th eigenvector. The projections become:

\[
\alpha^{(l)}_m = c_p^{(l)} \sum_{j \in A_p} K(x_j, x_m) + \sum_{u \notin A_p} \alpha^{(l)}_u K(x_u, x_m) + b_l,
\]

where \( l = 1, \ldots, k - 1, \ p = 1, \ldots, k, \ m = 1, \ldots, N_v \). Note
that \( K(x_j, x_m) \) tends to zero when \( x_j \) and \( x_m \) are not in
the same cluster; therefore, the term \( \sum_{u \notin A_p} \alpha^{(l)}_u K(x_u, x_m) \)
is very small compared to the term \( c_p^{(l)} \sum_{j \in A_p} K(x_j, x_m) \).
In this case, validation points belonging to the same
cluster are represented as collinear points in the score
variables space.

The proposed encoding scheme is based on an arbitrary
set of orthogonal hyperplanes induced by the eigenpro-
blem. Due to the fact that the eigenvectors can be arbitrarily
rotated, the hyperplanes can bisect data points in the same
cluster if the data are not well represented. However, this is
improbable for well-represented data because, in that case,
data points belonging to the same cluster will be located in
the same orthant. The algorithm of the proposed multiway
spectral clustering approach is shown in Table 2.

6 Model Selection

The results obtained using spectral clustering depend
greatly on the choice of the affinity function and its
parameters. The most used affinity function is the radial
basis function (RBF) kernel \( K(x_j, x_j) = \exp(-0.5|x_j - x_j|/\sigma^2) \) with tuning parameter \( \sigma^2 \). The selection of this
parameter is not considered as part of the learning process
and is commonly done in a manual way using grid search
and minimizing some cluster distortion function [3]. How-
ever, counterexamples can be easily found in which clusters
with small distortion do not correspond to the natural
grouping of the data. The tuning method proposed in [8]
introduces local scaling by selecting a \( \sigma_i \) parameter per data
point \( x_i \). The selection of this scaling parameter is done in a
heuristic way using the distance of the point \( x_i \) to some
nearest neighbor. Again, the results of this approach depend
on the choice of the nearest neighbor. A different method
that can be used to learn the parameters given a particular
partition was proposed in [7]. In this case, the method

\[
\text{sign}^{(l)} = \text{sign}(\Omega_{\text{test}}^{(l)} + b_l N_{\text{tot}}), l = 1, \ldots, k - 1.
\]

This natural extension to out-of-sample points without
relying on approximations like the Nystrom method corre-
sponds to the main advantage of the weighted kernel PCA
framework. In this way, the clustering model can be trained,
validated, and tested in an unsupervised learning scheme.

In this section, we provide an implementation of the
multiway formulation to spectral clustering together with
an extension to out-of-sample data points. According to
Fan’s theorem [24], the relaxed solutions to the \( k \)-way
normalized cut are arbitrary rotated eigenvectors. An
intuitive approach to obtain the clusters from the eigenvec-
tors is to find the rotation matrix that makes the eigenvector
matrix a cluster indicator matrix (i.e., only one nonzero
entry per row). The spectral method discussed in [8] uses
this approach and proposes an incremental gradient
descent algorithm to obtain the optimal number of clusters.
However, finding this rotation matrix increases the com-
putational cost. The computation of the rotation matrix can be
avoided using \( k \)-means because this method is invariant to
rotated eigenvectors but is prone to local minima, requires
good initialization, and assumes spherical clusters.

\textbf{Proposition 2.} Consider a training data set \( D = \{x_i\}^N_{i=1}, x_i \in \mathbb{R}^d \)
sampled i.i.d. from an underlying distribution \( \mathcal{P} \), a
validation data set \( D' = \left\{ x_i' \right\}^N_{i=1}, x_i' \in \mathbb{R}^d \) also sampled
i.i.d. from \( \mathcal{P} \), and an isotropic kernel function \( K(x_j, x_j) \),
\( i, j = 1, \ldots, N \). Let \( \Omega \) be the kernel matrix with \( ij \)-entry \( \Omega_{ij} =
K(x_i, x_j) \) and let \( \alpha^{(l)} \), \( l = 1, \ldots, k - 1 \), be the eigenvectors of
(9). Consider the following assumptions:

- \( \mathcal{P} \) contains \( k \) clusters denoted by \( \Delta = \{A_1, \ldots, A_k\} \).
- The eigenvectors \( \alpha^{(l)} \), \( l = 1, \ldots, k - 1 \), are piecewise
constant.
- \( K(x_i, x_j) \to 0 \), when \( x_i \) and \( x_j \) are in different clusters.

Then, validation points in the same cluster are collinear in
the \( k - 1 \)-dimensional subspace spanned by the columns of \( Z' \),
where \( Z' \in \mathbb{R}^{N_v \times (k-1)} \) is the validation score variables matrix
with \( ml \)-entry \( Z'^{ml} = \alpha^{(l)}_m = \sum_{i=1}^{N_v} \alpha^{(l)}_i K(x_i, x'_m) + b_l \) and
\( m = 1, \ldots, N_v \), \( l = 1, \ldots, k - 1 \).

\textbf{Proof.} The eigenvectors \( \alpha^{(l)} \) are piecewise constant; therefore,
\( \alpha^{(l)} = c_p^{(l)} \) for all \( x_j \in A_p \), \( l = 1, \ldots, k - 1, p = 1, \ldots, k \).
where \( c_p^{(l)} \) is the constant value corresponding to the \( p \)th
cluster in the \( l \)th eigenvector. The projections become:

\[
\alpha^{(l)}_m = c_p^{(l)} \sum_{j \in A_p} K(x_j, x_m) + \sum_{u \notin A_p} \alpha^{(l)}_u K(x_u, x_m) + b_l,
\]
corresponds to a supervised setting and the clustering cost function characterizes the ability of the induced affinity matrix to produce clusters that fit the given partition. The selection of the number of clusters \( k \) is also a general problem for all clustering methods and several approaches have been proposed. Typical algorithms are the eigengap heuristics, the elbow criterion, the gap statistic [37], and the silhouette index [38]. The simplest approach is the eigengap heuristic which consists of choosing a value for \( k \) such that the eigengap \( |\lambda_{k+1} - \lambda_k| \) is greater than some threshold value \( \theta \). The justification of this procedure relies on the fact that in the ideal case when the graph has \( k \) disconnected components, the multiplicity of the eigenvalue 1 is \( k \). Thus, the \((k+1)\)th eigenvalue will be less than 1 resulting in some eigengap. In this ideal case, the threshold can be set to zero, but in practical situations, this simple heuristic absolutely fails. The elbow criterion is a simple measure of the compactness of the clusters and the gap statistic together with the silhouette index are advanced cluster distortion indicators.

In this paper, we propose a criterion called the balanced line fit for estimating suitable spectral clustering parameters, namely, the RBF kernel parameter \( \sigma^2 \) (or the \( \chi^2 \) kernel parameter \( \sigma^2_z \)) and the number of clusters \( k \). This criterion exploits the eigenstructure of piecewise constant eigenvectors and the corresponding projections. Consider the following average measure of collinearity of a validation data set \( \mathcal{D}' \) with respect to a clustering \( \Delta = \{ A_1, \ldots, A_k \} \), \( k > 2 \),

\[
\text{linefit}(\mathcal{D}', k) = \frac{1}{k} \sum_{p=1}^{k-2} \left( \frac{\zeta(p)}{\sum p} - \frac{1}{k-1} \right),
\]

(16)

where \( \zeta(p) \geq \cdots \geq \zeta(1) \) are the ordered eigenvalues of the sample covariance matrix:

\[
\zeta(p) = \frac{1}{|A_p|} \mathbf{Z}(p)^\top \mathbf{Z}(p), p = 1, \ldots, k,
\]

(17)

where \( \mathbf{Z}(p) \in \mathbb{R}^{|A_p| \times (k-1)} \) is the matrix representing the zero mean score variables for validation data assigned to the \( p \)th cluster. The term \( \zeta(p)/\sum p \zeta(p) \) indicates how much of the total variance is contained on the eigenvector corresponding to the largest eigenvalue of \( \mathbf{Z}(p)^\top \mathbf{Z}(p) \). In this way, if the validation score variables for the \( p \)th cluster are collinear, then all variance is contained on the first eigenvector and \( \zeta(p)/\sum p \zeta(p) \) equals 1. On the other hand, if the total variance is evenly distributed on the eigenvectors, then \( \zeta(p)/\sum p \zeta(p) \) equals \( 1/(k-1) \). The additional terms in (16) fix the criterion such that the linefit equals 0 when the score variables are distributed spherically (i.e., the eigenvalues are identical) and equals 1 when the score variables are collinear.

Note that the linefit is defined only for \( k > 2 \). When \( k = 2 \), only one eigenvector is needed to obtain a binary clustering. Consider the matrix \( \mathbf{Z}^i \in \mathbb{R}^{N \times 2} \), where the \( m \)th entry in the first column corresponds to the score variables \( z_m = \sum_{p=1}^N \alpha_p K(x_i, x_m^p) + b \) and the \( m \)th entry in the second column to \( \sum_{p=1}^N K(x_i, x_m^p) + b \). This modification allows the application of the linefit criterion when \( k = 2 \):

\[
\text{linefit}(\mathcal{D}', k) = \begin{cases} 
\frac{2}{\sum_{p=1}^{k-2} \left( \frac{\zeta(p)}{\sum p} - \frac{1}{k-1} \right)} & \text{if } k = 2, \\
\frac{1}{k} \sum_{p=1}^{k-2} \frac{\zeta(p)}{\sum p} - \frac{1}{k-1} & \text{if } k > 2,
\end{cases}
\]

(18)

where \( \zeta(p) \) are the ordered eigenvalues of the matrix \( \mathbf{Z}^i \) representing the zero mean score variables for validation data assigned to the \( p \)th cluster when \( k = 2 \).

The balance index equals 1 when the clusters have the same number of elements and tends to 0 in extremely unbalanced cases. Combining the linefit with the balance index leads to the BLF:

\[
\text{BLF}(\mathcal{D}', k) = \eta \text{ linefit}(\mathcal{D}', k) + (1 - \eta) \text{ balance}(\mathcal{D}', k),
\]

(19)

where \( \eta \) is a parameter controlling the importance given to the linefit with respect to the balance index and \( 0 \leq \eta \leq 1 \).

7 Empirical Results

In this section, some experimental results are presented to illustrate the proposed approach. All experiments reported are carried out in MATLAB on an Intel Dual Core, 3.0 GHz, 2 GB RAM. All data have been separated into training, validation, and test sets. The BLF criterion is used on validation data to find the number of clusters \( k \) and the kernel parameters and the \( \eta \) parameter is fixed to \( \eta = 0.75 \). Hence, more emphasis is given to the linefit. Simulations on toy data and image segmentation are presented. The adjusted Rand index (ARI) [39] is used as an external validation in order to compare the clustering results with some external cluster indicators. This clustering performance index ranges from 0 to 1 and takes the unitary value when the clustering
results matches perfectly the external cluster indicators. We empirically observed that, for the used kernels, the learned bias terms are very small compared to the magnitude of the projections. This means that the score variables value is dominated by the kernel expansions.

We compare our proposed method in terms of performance and computation times with the spectral clustering scheme discussed in [11]. This method uses random subsampling to solve an eigenvalue problem of a small subset of training points, and then, extrapolates the eigenvectors to the complete set of data points using the Nyström approximation [12], [13]. The resulting approximated eigenvectors are not mutually orthogonal and additional orthogonalization steps are needed. The cluster indicators are obtained by applying k-means on the approximated orthogonalized eigenvectors. We used the one-shot implementation described in [11].

7.1 Toy Problems
The first toy problem consists of three Gaussian clouds in a 2D space. The total number of data points is 800. The RBF kernel parameter \( \sigma^2 \) was set to 0.08. The training set consisted of 200 randomly selected data points. Fig. 1 shows the
binary clustering boundaries induced by the encoding scheme. The first eigenvector $\alpha^{(1)}$ of (9) groups the left and bottom Gaussian clouds into one cluster. The second eigenvector $\alpha^{(2)}$ groups the bottom and right clouds into one cluster. The cluster indicators for out-of-sample points were inferred using the out-of-sample extension and the ECOC decoding scheme.

The second toy problem shows the influence of increasing number of clusters $k$ with small training set sizes. Fig. 2 shows the results of the proposed approach and the Nyström algorithm [11]. This data set consists of 10 Gaussian clouds for a total of 2,000 data points. We performed 10 randomizations of the training set and the total number of training points was set to one-fifth of the total number of points. We used the average ARI over the randomizations to assess the performance with respect to the optimal clustering. The proposed method performs better than the Nyström algorithm in terms of ARI.

The third toy problem consists of three concentric rings in a 2D space. The training, validation, and test sets contain 600, 1,200, and 800 data points, respectively. This problem is known to be difficult for spectral clustering because of its nonlinearity, the fact that every data point has few neighbors, resulting in a very sparse affinity matrix and the multiscale nature of the rings. Fig. 3 shows the training set and the model selection curves. In this particular case, the BLF and the silhouette index find the correct number of clusters, but for the selection of $\sigma^2$, the silhouette index selects a nonreliable value. The BLF is very close to the adjusted Rand index yielding maxima in the same range of perfect performance. The eigengap is very small in this experiment, making the eigendecomposition problematic. Despite this fact, the BLF succeeds in selecting optimal parameters. Fig. 4 shows the score variables and the clustering results for an optimal value of $\sigma^2$ and a value far away from the optimal range found by the BLF. The score variables in the optimal case are perfect lines resulting in an optimal clustering. In the nonoptimal case, the score variables do not show the line structure leading to a wrong
clustering. Fig. 5 shows comparisons of the proposed method with the Nystro¨m method in [11]. We compute the ARI between cluster indicators obtained through different random subsamplings. This can be seen as a measure of pairwise clustering agreement between different randomizations of the training data. Comparisons with respect to the optimal clustering are also reported. The proposed method shows less variability with respect to random subsamplings compared to the Nystro¨m method. This issue can be due to convergence problems of the reclustering step using $k$-means. In the proposed method, the agreement between cluster indicators tends to stabilize as the number of training points increases. Computation times are reported in Fig. 6.

### 7.2 Image Segmentation

For the image segmentation experiments, we used color images from the Berkeley image data set$^1$ [40]. We computed a local color histogram with a $5 \times 5$ pixels window around each pixel using minimum variance color quantization of eight levels. We used the $\chi^2$ test to compute the distance between two local color histograms $h^{(i)}$ and $h^{(j)}$ [41] $\chi^2_{ij} = 0.5 \sum_{b=1}^{B} (h^{(i)}_b - h^{(j)}_b)^2 / (h^{(i)}_b + h^{(j)}_b)$, where $B$ is total number of quantization levels. The histograms are assumed to be normalized $\sum_{b=1}^{B} h^{(i)}_b = 1$, $i = 1, \ldots, N$. The $\chi^2$ kernel $K(h^{(i)}, h^{(j)}) = \exp(-\chi^2_{ij} / \sigma^2)$ with parameter $\sigma^2 \in \mathbb{R}^+$ is positive definite and has shown to be a robust and efficient way of comparing the similarity between two histograms for color discrimination and image segmentation [11].

We performed model selection using the BLF for determining the optimal number of clusters $k$ and the $\chi^2$ kernel parameter $\sigma^2$. The training scenario consisted of

![Fig. 7. Model selection using the BLF for image with ID 167062. (a) Tuning the number of clusters $k$. (b) Tuning the kernel parameter $\sigma^2$. The model selection scenario consists of 1,000 pixel histograms for training and 20,000 pixel histograms for validation. The training and validation sets were randomized 20 times. Tuned parameters are $k = 2$, $\sigma^2 = 0.09$.](image)

![Fig. 8. Model selection using the BLF. (a) Original image. (b) Score variables for the validation set. The tuned parameters are $k = 4$, $\sigma = 0.084$. Note that the strong line structures present on the score variables. (c) Segment-label image.](image)

![Fig. 9. Mean computation times for the Segbench test data set over 10 randomizations of the training set. Solid line: Proposed method. Dashed line: Nystro¨m method [11] using the one-shot implementation. The main bottleneck in [11] is the orthogonalization of the approximated eigenvectors.](image)

---

1. [http://www.eecs.berkeley.edu /Research/Projects/CS/vision/ grouping/segbench/](http://www.eecs.berkeley.edu /Research/Projects/CS/vision/ grouping/segbench/)
1,000 pixel histograms for training and 20,000 pixel histograms for validation. The training and validation sets were randomized 20 times. Fig. 7 shows the model selection results for image with ID 167062. The obtained tuned parameters are $k = 2$, $\sigma_v = 0.09$. Similar tuning plots were used to determine the optimal parameters of each of the 10 images used. Fig. 8 shows the obtained clustering and the validation score variables for image with ID 145086 using tuned $k = 4$, $\sigma_v = 0.084$. Each cluster in the image is represented as a line in the score variables space. Mean computation times can be seen in Fig. 9. The bottleneck of the Nystroem method corresponds to the orthogonalization steps. For training set sizes $N > 2,000$, the proposed methods run faster than [11] by at least one order of magnitude.

Fig. 10. Image segmentation results using the proposed method and the Nyström algorithm [11]. The training set consisted of 1,000 randomly chosen pixel histograms. The proposed approach performs better than [11] with respect to human segmentation. The performance in terms of F-measure of the full test set can be seen in Table 3.
Performance of the Test Images from the Berkeley Image Data Set: (a) Proposed Method and (b) Nyström Method [11]

<table>
<thead>
<tr>
<th>Image ID</th>
<th>Proposed Method</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>115006</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>39080</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td>38905</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>115046</td>
<td>0.57</td>
<td>0.52</td>
</tr>
<tr>
<td>100001</td>
<td>0.63</td>
<td>0.60</td>
</tr>
<tr>
<td>115030</td>
<td>0.61</td>
<td>0.59</td>
</tr>
<tr>
<td>115058</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>115033</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>115036</td>
<td>0.58</td>
<td>0.54</td>
</tr>
</tbody>
</table>

(a) The proposed method performs better than [11] on 83 out of 100 images with respect to human segmentation.

(b) Proposed Method | Nyström Method [11]

<table>
<thead>
<tr>
<th>Image ID</th>
<th>Proposed Method</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>570083</td>
<td>0.54</td>
<td>0.51</td>
</tr>
<tr>
<td>570084</td>
<td>0.55</td>
<td>0.52</td>
</tr>
<tr>
<td>570085</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>570086</td>
<td>0.57</td>
<td>0.54</td>
</tr>
<tr>
<td>570087</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>570088</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>570089</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>570090</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>570091</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>570092</td>
<td>0.63</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The table compares the two methods using a performance measure with respect to human segmentation. The performance criterion used is the F-measure [42, 43]. This measure is known to be meaningful for evaluating image segmentation boundaries. The F-measure is closely related to the area under Receiver Operating Characteristic (ROC) curves and is defined as the harmonic mean between precision and recall. The proposed method outperforms the Nyström algorithm in terms of agreement with human segmentation.

8 CONCLUSIONS

A new formulation for multiway spectral clustering with out-of-sample extensions is proposed. This formulation is based on a weighted kernel PCA framework in which the clustering model can be extended to new points. The proposed approach is cast in a constrained optimization framework providing primal-dual insights in a learning scheme. In the case of well-formed clusters, the eigenvectors of a modified similarity matrix derived from the data display a special structure. Training data points in the same cluster are mapped into a single point in the space spanned by the eigenvectors. In the case of validation data, points in the same cluster are collinear in the space spanned by the projected variables. This structure is exploited by the proposed model selection criterion to obtain useful model parameters leading to visually appealing and interpretable clusters. Experiments with known difficult toy examples and image segmentation show the applicability of the proposed multiway clustering model and model selection criterion.

ACKNOWLEDGMENTS

This work was supported by grants and projects for the Research Council K.U. Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD/Postdocs and fellow grants), the Flemish Government FWO: PhD/Postdocs grants, projects G.0240.99, G.0211.05, G.0407.02, G.0197.02, G.0080.01, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0226.06, G.0302.07, ICCoS, ANMM; AWI; IWT: PhD grants, GBOU (McKnow) Soft4s, the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/01/40), the EU (FP5-Quprodis, ERNSI, Eureka 2063-Impact; Eureka 2419-FLiTE) and Contracts Research/Agreements (ISMIC/IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). The scientific responsibility is assumed by its authors.

REFERENCES


Carlos A. Alzate P. received the degree in electronic engineering from the Universidad Nacional de Colombia, Manizales, in 2002, the master’s degree in artificial intelligence and the PhD degree in engineering from the Katholieke Universiteit Leuven, Belgium, in 2004 and 2009, respectively. He is currently a postdoctoral researcher in the Department of Electrical Engineering (ESAT) of the K.U. Leuven. His research interests include kernel methods, unsupervised learning, model selection, and optimization.

Johan A.K. Suykens received the degree in electro-mechanical engineering and the PhD degree in applied sciences from the Katholieke Universiteit Leuven (K.U. Leuven), in 1989 and 1995, respectively. He has been a postdoctoral researcher with the Fund for Scientific Research FWO Flanders and is currently a professor with K.U. Leuven. He is an author of the books Artificial Neural Networks for Modelling and Control of Non-Linear Systems (Kluwer Academic Publishers) and Least Squares Support Vector Machines (World Scientific), a coauthor of the book Cellular Neural Networks, Multi-Scroll Chaos and Synchronization (World Scientific), and an editor of the books Nonlinear Modeling: Advanced Black-Box Techniques (Kluwer Academic Publishers) and Advances in Learning Theory: Methods, Models and Applications (IOS Press). He is a senior member of the IEEE and has served as an associate editor for the IEEE Transactions on Circuits and Systems (1997-1999 and 2004-2007), and since 1998, he has been serving as an associate editor for the IEEE Transactions on Neural Networks. He received an IEEE Signal Processing Society 1999 Best Paper (Senior) Award and several Best Paper Awards at International Conferences. He is a recipient of the International Neural Networks Society INNS 2000 Young Investigator Award for significant contributions in the field of neural networks. He has served as a director and organizer of the NATO Advanced Study Institute on Learning Theory and Practice (Leuven, 2002), as a program cochair for the International Joint Conference on Neural Networks 2004 and the International Symposium on Nonlinear Theory and Its Applications 2005, and as an organizer of the International Symposium on Synchronization in Complex Networks 2007.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.