ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1670-1676 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Theoretical developments for interpreting kernel spectral clustering from alternative viewpoints

Diego Peluffo-Ordóñez [*1], Paul Rosero-Montalvo[1,2], Ana Umaquinga-Criollo[1],Luis Suárez-Zambrano[1], Hernan Domínguez-Limaico[1], Omar Oña-Rocha[1], Stefany Flores-Armas[1], Edgar Maya-Olalla[1]

[1] *Universidad Técnica del Norte, Facultad de Ingeniería en Ciencias Aplicadas, 100150, Ecuador*
[2] *Instituto Tecnológico Superior 17 de Julio, Yachay, Ecuador*

ARTICLEINFO

Article history:
Received: 03 April, 2017
Accepted: 23 July, 2017
Online: 27 August, 2017

Keywords:
Kernel
Spectral Clustering
Support Vector Machines

ABSTRACT

To perform an exploration process over complex structured data within unsupervised settings, the so-called kernel spectral clustering (KSC) is one of the most recommended and appealing approaches, given its versatility and elegant formulation. In this work, we explore the relationship between (KSC) and other well-known approaches, namely normalized cut clustering and kernel k-means. To do so, we first deduce a generic KSC model from a primal-dual formulation based on least-squares support-vector machines (LS-SVM). For experiments, KSC as well as other consider methods are assessed on image segmentation tasks to prove their usability.

1 Introduction

In general, for classifying or grouping a set of objects (represented as data points) into subsets holding similar objects, the field of machine learning - specifically, the pattern recognition- provides two great alternatives being essentially different from each other: Supervised- and Unsupervised-learning-based approaches. The former ones normally establish a model from beforehand known information on data normally provided by an expert, while the latter ones form the groups by following a natural clustering criterion based on a (traditionally heuristic) procedure of data exploration [1]. Therefore, unsupervised clustering techniques are preferred when object labelling is either unavailable or unfeasible. In literature, we can find tens of clustering techniques, which are based on different principles and criteria (such as: distances, densities, data topology, and divergences, among others) [2]. Some remarkable, emerging applications are inbalanced data analysis [3] and time-varying data analysis [4]. Particularly, spectral clustering (SC) is a suitable technique to deal with grouping problems involving hardly separable clusters. Many SC approaches have been proposed, among them: Normalized-cut-based clustering (NCC), which, applied as explained in [5], heuristically and iteratively estimates binary cluster indicators [5] or approximates the solution in a one-iteration fashion by solving a quadratic programming problem [6]. Kernel k-means (KKM) that can be formulated using eigenvectors [7]. Kernel spectral clustering (KSC), which uses a latent variable model and a least-squares-support-vector-machine (LS-SVM) formulation [8]. This work has a particular focus on KSC, being one of the most modern approaches. It has been widely used in numerous applications such as time-varying data [9,10], electricity load forecasting [11], prediction of industrial machine maintenance [12], and among others. Also, some improvements and extensions have been proposed [12–14].

The aim of this work is to demonstrate the relationship between KSC and other approaches, namely NCC and KKM. To do so, elegant mathematical developments are performed. Starting from either the primal or dual formulation of KSC, we show clearly the links with the other considered methods. Experimentally, in order to assess the clustering performance, we explore the benefit of each considered method on image segmentation. In this connection, images extracted from the free access Berkeley Segmentation Data Set [15] are used. As a meaningful result of this work, Also, we provide mathematical and experimental evidence of the usability of combining together a LS-SVM formulation and a generic latent variable model for clustering purposes.

The rest of this paper is organized as follows: Section 2 outlines the primal-dual formulation for KSC starting with a LS-SVM formulation regarding a variable model, which naturally yields an eigenvector-

[*] Diego Peluffo, , Contact No (+593)991250815 & dhpeluffo@utn.edu.ec

https://dx.doi.org/10.25046/aj0203208

based solution. Sections 3 and 4 explore and discusses on the links of KSC with NCC and KKM, respectively. Some experimental results are shown in Section 5. Section 6 presents some additional remarks on improved versions of KSC and its relationship with spectral dimensionality reduction. Finally, section 7 draws some final and concluding remarks.

2 Kernel Spectral clustering

Accounting for notation and future statements, let us consider the following definitions: Define a set of N objects or samples represented by d-dimensional feature vectors. Likewise, consider a data matrix holding all the feature vectors, so that $\mathbf{X} \in \mathbb{R}^{N\times d}$: $\mathbf{X} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th d dimensional feature vector or data point. KSC is aiming to split $\mathbf{X}$ into K disjoint subsets, being K the number of desired groups.

2.1 Latent variable model and problem formulation

In the following, the clustering model is described. Let $\mathbf{e}^{(l)} \in \mathbb{R}^N$ be the l-th projection vector, which is assumed in the following latent variable form:

$$\mathbf{e}^{(l)} = \boldsymbol{\Phi}\mathbf{w}^{(l)} + b_l \mathbf{1}_N, \tag{1}$$

where $\mathbf{w}^{(l)} \in \mathbb{R}^{d_h}$ is the l-th weighting vector, b_l is a bias term, n_e is the number of considered latent variables, notation $\mathbf{1}_N$ stands for a N dimensional all-ones vector, and the matrix $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1)^\top, \ldots, \phi(\mathbf{x}_N)^\top]^\top$, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times d_h}$, is a high dimensional representation of data. The function $\phi(\cdot)$ maps data from the original dimension to a higher one d_h, i.e., $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_h}$. Therefore, $\mathbf{e}^{(l)}$ represents the latent variables from a set of n_e binary cluster indicators obtained with $\text{sign}(\mathbf{e}^{(l)})$, which are to be further encoded to obtain the K resultant groups.

From the least-squares SVM formulation of equation (1), the following optimization problem can be stated:

$$\max_{\mathbf{e}^{(l)},\mathbf{w}^{(l)},\mathbf{b}^{(l)}} \frac{1}{2N}\sum_{l=1}^{n_e} \gamma_l \mathbf{e}^{(l)\top}\mathbf{V}\mathbf{e}^{(l)} - \frac{1}{2}\sum_{l=1}^{n_e} \mathbf{w}^{(l)\top}\mathbf{w}^{(l)} \tag{2a}$$

$$\text{s.t. } \mathbf{e}^{(l)} = \boldsymbol{\Phi}^\top \mathbf{w}^{(l)} + b_l \mathbf{1}_N, \tag{2b}$$

where $\gamma_l \in \mathbb{R}^+$ is the l-th regularization parameter and $\mathbf{V} \in \mathbb{R}^{N\times N}$ is a diagonal matrix representing the weight of projections.

2.2 Matrix problem formulation

For the sake of simplicity, we can express the primal formulation (2) in matrix terms, as follows:

$$\max_{\mathbf{E},\mathbf{W},\mathbf{b}} \quad \frac{1}{2N}\text{tr}(\mathbf{E}^\top\mathbf{V}\mathbf{E}\Gamma) - \frac{1}{2}\text{tr}(\mathbf{W}^\top\mathbf{W}) \tag{3a}$$

$$\text{s.t. } \mathbf{E} = \boldsymbol{\Phi}\mathbf{W} + \mathbf{1}_N \otimes \mathbf{b}^\top, \tag{3b}$$

where $\mathbf{b} = [b_1, \ldots, b_{n_e}]$, $\mathbf{b} \in \mathbb{R}^{n_e}$, $\Gamma = \text{Diag}([\gamma_1, \ldots, \gamma_{n_e}])$, $\mathbf{W} = [\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(n_e)}]$, $\mathbf{W} \in \mathbb{R}^{d_h\times n_e}$, and $\mathbf{E} = [\mathbf{e}^{(1)}, \cdots, \mathbf{e}^{(n_e)}]$, $\mathbf{E} \in \mathbb{R}^{N\times n_e}$. Notations $\text{tr}(\cdot)$ and $\otimes$ denote the trace and the Kronecker product, respectively. By minimizing the previous cost function, the goals of minimizing the weighting variance of $\mathbf{E}$ and maximizing the variance of $\mathbf{W}$ are reached simultaneously. Let $\Sigma_\mathbf{E}$ be the weighting covariance matrix of $\mathbf{E}$ and $\Sigma_\mathbf{W}$ be the covariance matrix of $\mathbf{W}$. Since matrix $\mathbf{V}$ is diagonal, we have that $\text{tr}((\mathbf{V}^{1/2}\mathbf{E})^\top\mathbf{V}^{1/2}\mathbf{E}) = \text{tr}(\Sigma_\mathbf{E})$. In other words, $\Sigma_\mathbf{E}$ is the covariance matrix of weighted projections, i.e., the projections scaled by square root of matrix $\mathbf{V}$. As well, $\text{tr}(\mathbf{W}^\top\mathbf{W}) = \text{tr}(\Sigma_\mathbf{W})$. Then, KSC can be seen as a kernel, weighted principal component analysis (KWPCA) approach [8].

2.3 Solving KSC by using a dual formulation

To solve the KSC problem, we form the corresponding Lagrangian of the problem from equation (2) as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{E},\mathbf{W},\Gamma,\mathbf{A}) = &\frac{1}{2N}\text{tr}(\Gamma\mathbf{E}^\top\mathbf{V}\mathbf{E}) - \frac{1}{2}\text{tr}(\mathbf{W}^\top\mathbf{W}) \\ &- \text{tr}(\mathbf{A}^\top(\mathbf{E} - \boldsymbol{\Phi}\mathbf{W} - \mathbf{1}_N \otimes \mathbf{b}^\top)),\end{aligned} \tag{4}$$

where matrix $\mathbf{A} \in \mathbb{R}^{N\times n_e}$ holds the Lagrange multiplier vectors $\mathbf{A} = [\alpha^{(1)}, \cdots, \alpha^{(n_e)}]$, and $\alpha^{(l)} \in \mathbb{R}^N$ is the l-th vector of Lagrange multipliers.

Solving the partial derivatives on $\mathcal{L}(\mathbf{E},\mathbf{W},\Gamma,\mathbf{A})$ to determine the Karush-Kuhn-Tucker conditions, we obtain:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{E}} = 0 \quad &\Rightarrow \mathbf{E} = N\mathbf{V}^{-1}\mathbf{A}\Gamma^{-1}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{W}} = 0 \quad &\Rightarrow \mathbf{W} = \boldsymbol{\Phi}^\top\mathbf{A}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{A}} = 0 \quad &\Rightarrow \mathbf{E} = \boldsymbol{\Phi}\mathbf{W}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{b}} = 0 \quad &\Rightarrow \mathbf{b}^\top\mathbf{1}_N = 0.\end{aligned}$$

Therefore, by eliminating the primal variables from initial problem (2) and assuming a kernel trick such that $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \boldsymbol{\Omega}$, being $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ a given kernel matrix, the following eigenvector-based dual solution is obtained:

$$\mathbf{A}\Lambda = \mathbf{V}(\mathbf{I}_N + (\mathbf{1}_N \otimes \mathbf{b}^\top)(\boldsymbol{\Omega}\Lambda)^{-1})\boldsymbol{\Omega}\mathbf{A}, \tag{5}$$

where $\Lambda = \text{Diag}(\lambda)$, $\Lambda \in \mathbb{R}^{N\times N}$, $\lambda \in \mathbb{R}^N$ is the vector of eigenvalues with $\lambda_l = N/\gamma_l$, $\lambda_l \in \mathbb{R}^+$.

Also, taking into account that the kernel matrix represents the similarity matrix of a graph with K connected components as well as $\mathbf{V} = \mathbf{D}^{-1}$ where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the degree matrix defined as $\mathbf{D} = \text{Diag}(\boldsymbol{\Omega}\mathbf{1}_N)$; then the $K-1$ eigenvectors contained in $\mathbf{A}$, associated to the largest eigenvalues, are piecewise constant and become indicators of the corresponding connected parts of the graph. Therefore, value n_e is fixed

to be $K-1$ [8]. With the aim of achieving a dual formulation, but satisfying the condition $\mathbf{b}^\top \mathbf{1}_N = 0$ by centering vector $\mathbf{b}$ (i.e. with zero mean), the bias term should be chosen in the form

$$b_l = -1/(\mathbf{1}_N^\top \mathbf{V}\mathbf{1}_N)\mathbf{1}_N^\top \mathbf{V}\boldsymbol{\Omega}\alpha^{(l)}. \quad (6)$$

Thus, the solution of problem of equation (3) is reduced to the following eigenvector-related problem:

$$\mathbf{A}\boldsymbol{\Lambda} = \mathbf{V}\mathbf{H}\boldsymbol{\Omega}\mathbf{A}, \quad (7)$$

where matrix $\mathbf{H} \in \mathbb{R}^{N\times N}$ is the centering matrix that is defined as

$$\mathbf{H} = \mathbf{I}_N - \frac{1}{\mathbf{1}_N^\top \mathbf{V}\mathbf{1}_N}\mathbf{1}_N\mathbf{1}_N^\top, \mathbf{V},$$

where $\mathbf{I}_N$ denotes a N-dimensional identity matrix and, $\boldsymbol{\Omega} = [\Omega_{ij}]$, $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$, being $\Omega_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), i, j \in [N]$. Notation $\mathcal{K}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ stands for the kernel function. As a result, the set of projections can be calculated as follows:

$$\mathbf{E} = \boldsymbol{\Omega}\mathbf{A} + \mathbf{1}_N \otimes \mathbf{b}^\top. \quad (8)$$

Once projections are calculated, we proceed to carry out the cluster assignment by following an encoding procedure applied on projections. Because each cluster is represented by a single point in the $K-1$-dimensional eigenspace, such that those single points are always in different orthants due also to the KKT conditions, we can encode the eigenvectors considering that two points are in the same cluster if they are in the same orthant in the corresponding eigenspace [8]. Then, a code book can be obtained from the rows of the matrix containing the $K-1$ binarized leading eigenvectors in the columns, by using $\text{sign}(\mathbf{e}^{(l)})$. Then, matrix $\widetilde{\mathbf{E}} = \text{sgn}(\mathbf{E})$ is the code book being each row a codeword.

2.4 Out-of-sample extension

KSC can be extended to out-of-samples analysis without re-clustering the whole data to determine the assignment cluster membership for new testing data [8]. In particular, defining $\mathbf{z} \in \mathbb{R}^{n_e}$ as the projection vector of a testing data point $\mathbf{x}_{\text{test}}$, and by taking into consideration the training clustering model, the testing projections can be computed as:

$$\mathbf{z} = \mathbf{A}^\top \boldsymbol{\Omega}_{\text{test}} + \mathbf{b}, \quad (9)$$

where $\boldsymbol{\Omega}_{\text{test}} \in \mathbb{R}^{n_e}$ is the kernel vector such that

$$\boldsymbol{\Omega}_{\text{test}} = [\Omega_{\text{test}_1}, \ldots, \Omega_{\text{test}_N}]^\top,$$

and $\Omega_{\text{test}_i} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_{\text{test}})$. Once, the test projection vector $\mathbf{z}$ is computed, a decoding stage is carried out that consists of comparing the binarized projections with respect to the codewords in the code book $\widetilde{\mathbf{E}}$ and assigning cluster membership based on the minimal Hamming distance [8].

2.5 KSC algorithm

Following the pseudo-code (Algorithm 1) to perform KSC is shown.

Algorithm 1 Kernel spectral clustering: $[q_{\text{train}}, q_{\text{test}}] = \text{KSC}(\mathbf{X}, \mathcal{K}(\cdot,\cdot), K)$

1: Input: K, $\mathbf{X}$, $\mathcal{K}(\cdot,\cdot)$
2: Form the kernel matrix $\boldsymbol{\Omega}$ such that $\Omega_{ij} = \mathcal{K}(\mathbf{y}_i, \mathbf{x}_j)$
3: Determine $\mathbf{E}$ through (8)
4: Form the training codebook by binarizing $\widetilde{\mathbf{E}} = \text{sgn}(\mathbf{E})$
5: Assign the output training labels q_{train} according to similar codewords
6: Compute the training codewords for testing
7: Assign the output testing labels q_{test} according to the minimal Hamming distance when comparing with training codewords
8: Output: $q_{\text{train}}, q_{\text{test}}$

3 Links between KSC and NCC

This section deals with the relationship between KSC and NCC, starting from the formulation of the NC problem until reaching a weighting principal component analysis (WPCA) formulation in a finite domain.

3.1 Multi-cluster spectral clustering (MCSC) from two point of view

In [5], the so-called Multi-cluster spectral clustering (MCSC) is introduced, which is based on the well-known k-way normalized cut-based formulation given by:

$$\max \frac{1}{K}\frac{\text{tr}(\mathbf{M}^\top \boldsymbol{\Omega}\mathbf{M})}{\text{tr}(\mathbf{M}^\top \mathbf{D}\mathbf{M})} = \max_{\mathbf{m}^{(k)}} \frac{1}{K}\frac{\sum_{k=1}^K \mathbf{m}^{(k)\top}\widehat{\boldsymbol{\Omega}}\mathbf{m}^{(k)}}{\sum_{k=1}^K \mathbf{m}^\top \mathbf{m}} \quad (10a)$$

$$\text{s.t. } \mathbf{M} \in \{0,1\}^{N\times K}, \quad \mathbf{M}\mathbf{1}_K = \mathbf{1}_N. \quad (10b)$$

Expressions (10a) and (10b) are the formulation of the NC optimization problem, named (*NCPM*). Previous formulation can also be expressed as follows. Let $\widehat{\boldsymbol{\Omega}} = \mathbf{D}^{-1/2}\boldsymbol{\Omega}\mathbf{D}^{-1/2}$ be a normalized kernel matrix and $\mathbf{L} = \mathbf{D}^{1/2}\mathbf{M}$ be a binary matrix normalized by the square root of the kernel degree. Then, a new *NCPM* version can be expressed as:

$$\max_{\mathbf{L}} \frac{1}{K}\frac{\text{tr}(\mathbf{L}^\top \widehat{\boldsymbol{\Omega}}\mathbf{L})}{\text{tr}(\mathbf{L}^\top \mathbf{L})} = \max_{\ell^{(k)}} \frac{1}{K}\frac{\sum_{k=1}^K \ell^{(k)\top}\widehat{\boldsymbol{\Omega}}\ell^{(k)}}{\sum_{k=1}^K \ell^{(k)\top}\ell^{(k)}} \quad (11a)$$

$$\text{s.t. } \mathbf{D}^{-1/2}\mathbf{L} \in \{0,1\}^{N\times K}, \quad \mathbf{D}^{-1/2}\mathbf{L}\mathbf{1}_K = \mathbf{1}_N, \quad (11b)$$

where $\ell^{(k)}$ is the column k of $\mathbf{L}$.

Solution of former problem has been addressed in [5, 16] by introducing a relaxed version, in which numerator is maximized subject to denominator is constant, so

$$\max_{\mathbf{L}} \frac{1}{K}\text{tr}(\mathbf{L}^\top \widehat{\boldsymbol{\Omega}}\mathbf{L}) \qquad \text{s.t. tr}(\mathbf{L}^\top \mathbf{L}) = \text{const.} \quad (12)$$

Indeed, authors assume the particular case $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_K$, i.e. letting $\mathbf{L}$ be an orthonormal matrix. Then, solution correspond to any K-dimensional basis of normalized matrix eigenvectors. Despite that in [16] it is presented an one-iteration solution for *NCPM* with suboptimal results avoiding the calculation of SVD per iteration, the omitting of the effect of denominator $\text{tr}(\mathbf{L}^\top\mathbf{L})$ by assuming orthogonality causes that the solution cannot be guaranteed to be a global optimum. In addition, this kind of formulation provide non-stable solutions due to the heuristic search carried out to determine an optimal rotation matrix [16].

3.2 Solving the problem by a difference: Empirical feature map

Recalling original problem 3.1, we introduce another way to solve the *NCPM* formulation via a minimization problem where the aims for maximizing $\text{tr}(\mathbf{L}^\top\widehat{\mathbf{\Omega}}\mathbf{L})$ and minimizing $\text{tr}(\mathbf{L}^\top\mathbf{L})$ can be accomplished simultaneously, so:

$$\max_{\mathbf{L}} \text{tr}(\mathbf{L}^\top\widehat{\mathbf{\Omega}}\mathbf{L}\,\text{Diag}(\gamma)) - \text{tr}(\mathbf{L}^\top\mathbf{L}) \tag{13}$$

where $\gamma = (\gamma_1,\ldots,\gamma_N)^\top$ is a vector containing the regularization parameters.

Let us assume $\mathbf{\Omega} = \mathbf{\Psi}\mathbf{\Psi}^\top$ where $\mathbf{\Psi}$ is a $N \times N$ dimensional auxiliary matrix, and consider the following equality:

$$\begin{aligned}\text{tr}(\widehat{\mathbf{\Omega}}) &= \text{tr}(\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{D}^{-1/2}) = \text{tr}(\mathbf{D}^{-1}\mathbf{\Omega}) \\ &= \text{tr}(\mathbf{D}^{-1}\mathbf{\Psi}\mathbf{\Psi}^\top) = \text{tr}(\mathbf{\Psi}^\top\mathbf{D}^{-1}\mathbf{\Psi}),\end{aligned}$$

then

$$\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{D}^{-1/2} = \mathbf{\Psi}^\top\mathbf{D}^{-1}\mathbf{\Psi}.$$

Previous formulation is possible since kernel matrix Ω is symmetric. Now, let us define $\mathbf{h}^{(k)} \in \mathbb{R}^N = \mathbf{\Psi}^\top\ell^{(k)}$ as the k-th projection and $\mathbf{H} = (\mathbf{h}^{(1)},\cdots,\mathbf{h}^{(K)})$ as the projections matrix. Then, formulation given by (13) can be expressed as follows:

$$\max_{\mathbf{h}^{(k)},\ell^{(k)},\gamma_k} \frac{1}{2K}\sum_{k=1}^{K}\gamma_k\mathbf{h}^{(k)\top}\mathbf{V}\mathbf{h}^{(k)} - \frac{1}{2}\sum_{k=1}^{K}\ell^{(k)\top}\ell^{(k)} \tag{14a}$$

$$\text{such that } \mathbf{h}^{(k)} = \mathbf{\Psi}\ell^{(k)}, \tag{14b}$$

where matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$ can be chosen as:

- $\mathbf{I}_N$: We can normalize matrix $\mathbf{\Omega}$ in such way for all i condition $\sum_j^N \omega_{ij} = 1$ is satisfied and therefore we would obtain a degree matrix equaling the identity matrix. Then, $\sum\mathbf{h}^{(l)\top}\mathbf{h}^{(l)} = \text{tr}(\mathbf{H}^\top\mathbf{H})$, which corresponds to a PCA-based formulation.
- Diag($\mathbf{v}$): With $\mathbf{v} \in \mathbb{R}^N$ such that $\mathbf{v}^\top\mathbf{v} = 1$, we have a WPCA approach.
- $\mathbf{D}^{-1}$: Given the equality $\mathbf{V} = \mathbf{D}^{-1}$, optimization problem can be solved by means of a procedure based on random walks; being the case of interest in this study.

3.2.1 Gaussian processes

In terms of Gaussian processes, variable $\mathbf{\Psi}$ represents a mapping matrix such that $\mathbf{\Psi} = (\psi(\mathbf{x}_1),\ldots,\psi(\mathbf{x}_N))$ and where $\psi(\cdot) : \mathbb{R}^d \to \mathbb{R}^N)$ is mapping function, which provides a new N-dimensional data representation where resultant clusters are assumed to be more separable. Also, matrix $\mathbf{\Omega}$ is to be chosen as a Gaussian kernel [17]. Therefore, according to optimization problem given by (14), term $\mathbf{h}^{(k)}$ is to be the k-th projection of normalized binary indicators as $\mathbf{h}^{(k)} = \mathbf{\Psi}\ell^{(k)}$.

3.2.2 Eigen-solution

We present a solution for 14, which after solving the KKT conditions on its corresponding Lagrangian, an eigenvectors problem is yielded. Then, we first solve the Lagrangian of problem (14) so:

$$\mathcal{L}(\mathbf{h},\ell,\gamma,\alpha) = \frac{1}{2K}\mathbf{h}^\top\mathbf{V}\mathbf{h} - \frac{1}{2}\ell^\top\ell - \alpha^\top(\mathbf{h} - \mathbf{\Psi}\mathbf{w}), \tag{15}$$

where α is a N-dimensional vector containing the Lagrange multipliers.

Solving the partial derivatives to determine the KKT conditions, we have:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{h}} = 0 \quad &\Rightarrow \mathbf{h} = \frac{K}{\gamma}\mathbf{D}\alpha, \\ \frac{\partial\mathcal{L}}{\partial\ell} = 0 \quad &\Rightarrow \ell = \mathbf{\Psi}^\top\alpha, \\ \frac{\partial\mathcal{L}}{\partial\alpha} = 0 \quad &\Rightarrow \mathbf{h} = \mathbf{\Psi}\ell.\end{aligned}$$

Eliminating the primal variables, we obtain the following eigenvector problem:

$$\lambda\alpha = \mathbf{D}^{-1}\mathbf{\Omega}\alpha, \tag{16}$$

where $\lambda = N/\gamma$. Then, matrix $\mathbf{\Delta}_K = (\alpha^{(1)},\cdots,\alpha^{(K)})$ can be computed as the eigenvectors associated with the first K longest eigenvalues of $\mathbf{D}^{-1}\mathbf{\Omega}$.

Finally, projections matrix $\mathbf{H}$ is in the form

$$\mathbf{H} = \mathbf{\Psi}\mathbf{L} = \mathbf{\Psi}\mathbf{D}^{1/2}\mathbf{M} = \mathbf{\Omega}\mathbf{\Delta}_K, \tag{17}$$

and therefore $\mathbf{M} = \mathbf{\Psi}^{-1}\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{\Delta}_K$, where $\mathbf{\Psi}$ can be obtained from a Cholesky decomposition.

Then, within a finite domain, both solution and formulation of NCC can be expressed similarly as done in KSC. So it is demonstrated the relationship between a kernel-based model and Gaussian processes.

4 Links between KSC and KKM

Kernel K-means method (KKM) is a generalization of standard K-means that can be seen as a spectral relaxation when introducing a mapping function in the objective function formulation [18]. As mentioned throughout this paper, spectral clustering approaches usually are performed on a lower-dimensional space, keeping the pairwise relationships among nodes. Then, it often leads to a relaxed NP-problems where continuous solutions are obtained by a eigen-decomposition. Such an eigen-decomposition is regarding the normalized similarity matrix (Laplacian, as well). In a Kernel K-means framework, eigenvectors are considered as geometric coordinates and then K-means methods is applied over the eigen-space to get the resultant clusters [19, 20].

Previous instance is as follows: Suppose that we have a gray scale matrix $m \times n$ pixels in size. Characterizing each image pixel with d features -e. g., color spaces, morphological descriptors- it is yielded as a result a data matrix in the form $\mathbf{X} \in \mathbb{R}^{N \times d}$, where $N = mn$. Afterwards, the eigenvectors $\mathbf{V}\mathbb{R}^{N \times N}$ of a normalized kernel matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ such that $\mathbf{P} = \mathbf{D}^{-1}\mathbf{\Omega}$, being $\mathbf{\Omega}$ the kernel matrix and $\mathbf{D} \in \mathbb{R}^{N \times N}$ the corresponding degree matrix. Then, we proceed to cluster $\mathbf{V}$ into K groups using K-means algorithm: $\mathbf{q} = \text{kmeans}(\mathbf{V}, K)$, being $\mathbf{q} \in \mathbb{R}^N$ the output cluster indicator such that $q_i \in [K]$. The segmented image is then a $m \times n$ sized matrix holding regions in accordance with $\mathbf{q}$.

Briefly put, one simple way to perform a KKM procedure is applying k-means over the eigen-space. In Equation (7), the dual formulation is regarding the matrix $\mathbf{S} = \mathbf{VH\Omega}$ where weighting matrix can be chosen as $\mathbf{V} = \mathbf{D}^{-1}$ and $\mathbf{D}$ is the degree of the data-related graph. Since $\mathbf{H}$ causes a centering effect, matrix $\mathbf{S}$ is the same as $\mathbf{P}$ when kernel matrix $\mathbf{\Omega}$ is centered. In other words, KKM can be seen as a KSC formulation with an incomplete latent variable model being a non-centered one (with no bias term).

5 Results and discussion

In order to show how considered methods work, we conduct some experiments to test their clustering ability on segmenting images. To do so, the segmentation performance is quantified by a supervised index noted as Probabilistic Rand Index (PR), explained in [21], such that $PR \in [0, 1]$, being 1 when regions are properly segmented. Images are drawn from the free access Berkeley Segmentation Data Set [15]. To represent each image as a data matrix, we characterize the images by color spaces (RGB, YCbCr, LAbB, LUV) and the xy position of each pixel. At the end, data matrix $\mathbf{X}$ gathers N pixels represented by d characteristics (variables). To run the experiment, we resize the images at 20% of the original size due to memory usage restrictions. All the methods are performed with a given number of clusters K manually set as shown in shown in Fig. 5 and using the scaled exponential similarity matrix as described in [19], setting the number of neighbors to be 9.

To test all the methods in a fair scenario, kernel-based methods (KSC and KKM) use $\mathbf{\Omega}$ as kernel matrix, whereas such a matrix is the affinity matrix for NCC. As well, to perform the clustering procedure, the number of clusters is the same for all the considered methods. As can be readily appreciated, KSC overcome the rest of studied clustering methods. This fact can be attributed to the KSC formulation, which involves a whole latent variable model being in turn incorporated within a LS-SVM framework. Indeed, just like principal component analysis (PCA), KSC optimizes an energy term. Differently, such an energy term is regarding a latent variable instead of directly the input data matrix. Concretely, a latent variable model is used, which is linear and formulated in terms of projections of the input data. The versatility of KSC relies on the kernel matrix required during the optimization procedure of its cost function. Such a matrix holds pairwise similarities, then KSC can be seen as data-driven approach that not only consider the nature of data but yields a true clustering model. It is important to quote that -depending on the difficulty of the segmentation task- data matrices representing images yield features spaces, which may present hardly separable classes. Then, we have demonstrated the benefit of the KSC approach that uses a model along with a LS-SVM formulation -everything within a primal-dual scheme. Other studies have also proven the usability and versatility of this kind of approaches [8, 22].

6 Additional remarks

As explained in [23], KSC performance can be enhanced in terms of cluster separability by optimally projecting original input data and performing the clustering procedure over the projected space. Given the unsupervised nature, spectral clustering becomes very often a parametric approach, involving then a stage of selection/tunning of collection of initial parameters to avoid any local-optimum solution. Typically, the initial parameters are the kernel or similarity matrix and the number of groups. Nonetheless, in some problems when data are represented in a high-dimensional space and/or data-sets are non-linearly separable, a proper feature extraction may be an advisable alternative. In particular, a projection generated by a proper feature extraction procedure may provide a new feature space wherein the clustering procedure can reach more accurate cluster indicators. In other words, data projection accomplishes a new representation space, where the clustering can be improved, in terms of a given mapping criterion, rather than performing the clustering procedure directly over the original input data.

The work developed in [23] introduces a matrix projection focusing on a better analysis of the structure of data that is devised for a KSC. Since data projection can be seen as a feature extraction process,

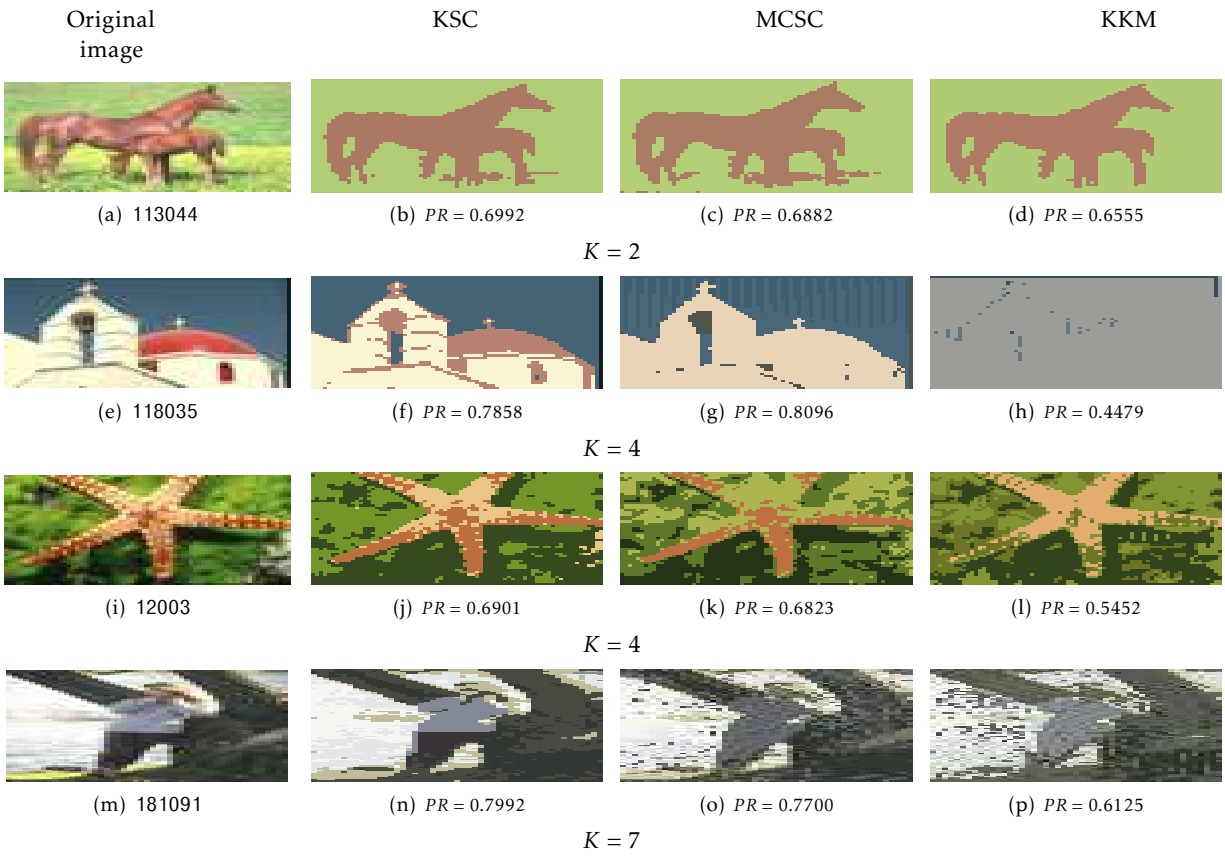

Figure 1: Clustering performance on image segmentation reached by all the considered methods. It is noticeable that KSC overcome the remaining methods. Images are data that traditionally involve highly non-separable clusters. Therefore, the benefit of using a whole latent variable model within a LS-SVM formulation is verified.

we propose the M-inner product-based data projection, in which the similarity matrix is also considered within the projection framework, similarly as discussed in [24]. There are two main reasons for using data projection to improve the performance of kernel spectral clustering: firstly, the data global structure is taken into account during the projection process and, secondly, the kernel method exploits the information of local structures.

Another study [25] explores the links of KSC with spectral dimensionality reduction from a kernel viewpoint. Particularly, the proposed formulation is LS-SVM in terms of a generic latent variable model involving the projected input data matrix. In order to state a kernel-based formulation, such a projection maps data onto a unknown high-dimensional space. Again, the solution of the optimization problem is addressed through a primal-dual scheme. Finally, once latent variables and parameters are determined, the resultant model outputs a versatile projected matrix able to represent data in a low-dimensional space. To do so, since the optimization is posed under a maximization criterion and dual version has a quadratic from, the eigenvectors associated with the largest eigenvalues can be chosen as a solution. Therefore, the generalized kernel model may represente a weighted version of kernel principal component analysis.

7 Conclusions

This works explores a widely-recommended method for unsupervised data classification, namely kernel spectral clustering (KSC). From elegant developments, the relationship between KSC and two other well-known spectral clustering approaches (normalized cut clustering and kernel k-means) is demonstrated. As well, the benefit of KSC-like approaches is mathematically and experimentally proved. The goodness of KSC relies on the nature of its formulation, which is based on a latent variable model incorporated into a least-square-support-vector-machine framework. Additionally, some key aspects and hints to improve KSC performance as well as its ability to represent dimensionality reduction approaches are briefly outlined and discussed.

As a future work, a generalized clustering framework is to be designed so that a wide range of spectral approaches can be represented. Doing so, the task of selecting and/or testing a spectral clustering method would become easier and fairer.

References

[1] F. Schwenker and E. Trentin, "Pattern classification and clustering: a review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[2] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Advances in Science, Technology and Engineering Systems Journal*, no. 3, pp. 116–124.

[4] H. Chebi, D. Acheli, and M. Kesraoui, "Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks," *Advances in Science, Technology and Engineering Systems Journal*, no. 5, pp. 56–63.

[5] Y. S. X. and S. Jianbo, "Multiclass spectral clustering," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 313.

[6] D. H. Peluffo-Ordóñez, C. Castro-Hoyos, C. D. Acosta-Medina, and G. Castellanos-Domínguez, "Quadratic problem formulation with linear constraints for normalized cut clustering," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2014, pp. 408–415.

[7] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 551–556.

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

[9] D. H. Peluffo-Ordónez, S. García-Vega, A. M. Alvarez-Meza, and C. G. Castellanos-Domínguez, "Kernel spectral clustering for dynamic data," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 238–245.

[10] R. Langone, C. Alzate, and J. A. Suykens, "Kernel spectral clustering with memory effect," *Physica A: Statistical Mechanics and its Applications*, 2013.

[11] C. Alzate and M. Sinn, "Improved electricity load forecasting via kernel spectral clustering of smart meters," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 943–948.

[12] R. Langone, R. Mall, C. Alzate, and J. A. Suykens, "Kernel spectral clustering and applications," in *Unsupervised Learning Algorithms*. Springer, 2016, pp. 135–161.

[13] D. H. Peluffo-Ordóñez, C. Alzate, J. A. K. Suykens, and G. Castellanos-Domínguez, "Optimal data projection for kernel spectral clustering," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2014, pp. 553–558.

[14] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. Suykens, "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 4, pp. 720–733, 2015.

[15] 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, July 2001, pp. 416–423.

[16] D. Peluffo, C. D. Acosta, and G. Castellanos, "An improved multi-class spectral clustering based on normalized cuts," *XVII Iberoamerican Congress on Pattern Recognition - CIARP*, 2012.

[17] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 1.

[18] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral relaxation for k-means clustering," *Advances in neural information processing systems*, vol. 14, pp. 1057–1064, 2001.

[19] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.

[20] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts," 2004.

[21] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 929–944, 2007.

[22] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 171–177.

[23] D. H. Peluffo-Ordóñez, C. Alzate, J. A. Suykens, and G. Castellanos-Domínguez, "Optimal data projection for kernel spectral clustering." in *European Symposium on Artificial Neural Networks - ESANN*, 2014.

[24] J. Rodríguez-Sotelo, D. Peluffo-Ordóñez, D. Cuesta-Frau, and G. Castellanos-Domínguez, "Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering," *Computer Methods and Programs in Biomedicine*, 2012.

[25] X. Blanco-Valencia, M. Becerra, A. Castro-Ospina, M. Ortega-Adarme, D. Viveros-Melo, J. Alvarado-Pérez, and D. Peluffo-Ordóñez, "Kernel-based framework for spectral dimensionality reduction and clustering formulation: A theoretical study," *ADCAIJ: Advances in Distributed Computing and Articial Intelligence Journal*, 2017.