A Regularized Formulation for Spectral Clustering with Pairwise Constraints

Carlos Alzate, Johan A. K. Suykens

***Abstract*—A regularized method to incorporate prior knowledge into spectral clustering in the form of pairwise constraints is proposed. This method is based on a weighted kernel principal component analysis (PCA) interpretation of spectral clustering with primal-dual least squares support vector machines (LS-SVM) formulations. The weighted kernel PCA framework allows incorporating pairwise constraints into the primal problem leading to a dual eigenvalue problem involving a modified kernel matrix. This modification on the metric is a regularized rank-1 downdate of the original kernel matrix. The clustering model can also be extended to out-of-sample points which becomes important for generalization, predictive purposes and large-scale data. An extension of an existing model selection criterion is also proposed. This extension introduces an additional term to the criterion measuring the constraint fit. Simulation results with toy examples and an image segmentation problem show the applicability of the proposed method.**

I. Introduction

Spectral clustering corresponds to a popular family of clustering algorithms that make use of the eigenvectors of a similarity matrix derived from the data [1], [2], [3], [4]. These methods arise from graph theory and correspond to relaxations of combinatorial problems leading to eigenvalue problems. Spectral clustering methods are known to perform well in cases where classical methods such as k-means fail. The clustering obtained is typically only for training data with no clear extensions to out-of-sample data although approximation methods have been discussed [5], [6]. A different view on spectral clustering based on weighted kernel PCA was discussed in [7], [8]. These formulations link spectral clustering with kernel PCA via primal/dual insights in a constrained optimization framework typical of least squares support vector machines (LS-SVMs) [9], [10]. One of the main advantages of this intepretation is the possibility to extend the clustering to new points (also called out-of-sample points) via projections onto the eigenvectors.

Incorporating prior knowledge into clustering algorithms has gained significant interest in the machine learning and data mining communities. The type of prior knowledge used is usually in the form of constraints on the cluster assignment. These constraints may come from experts, known or perceived similarity and common sense. Pairwise constraints are cluster assignments restrictions and can be *must-link* or *cannot-link*. *Must-link* constraints specify that two data points should belong to the same cluster while *cannot-link* constraints restrict that two data points should be in different clusters. Incorporating pairwise constraints into the clustering algorithm have been discussed for centroid-based algorithms such as k-means [11], [12]. However, few methods exist for constrained spectral clustering [13], [14]. The method described in [13] uses a Radial Basis Function (RBF) affinity function but replaces the entries corresponding to the constraints either with 0 for *cannot-link* or 1 for *must-link*. This ad-hoc modification does not work well in practice and may cause that the kernel matrix becomes negative-definite. The method in [14] proposes a modification of the kernel matrix arising from a Gaussian process interpretation but with no extensions to new points.

In this paper, we propose a regularized method to incorporate naturally pairwise constraints into spectral clustering. The formulation is based on [7] which is a multiway spectral clustering algorithm with out-of-sample extensions. The constraints lead to a modification of the metric causing a regularized rank-1 downdate of the original kernel matrix. An extension to a constrained model selection criterion is also proposed. The original Balanced Line Fit (BLF) discussed in [7] is modified in order to take into account also the constraint fit. 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 II contains a review of classical multiway spectral clustering. Section III summarizes the multiway spectral clustering approach in which the proposed method is based. In Section IV, we describe the proposed method together with an algorithm. Section V contains the proposed extension of a model selection criterion. Section VI contains the experimental results and in Section VII we give conclusions.

II. Multiway Spectral Clustering

A classical problem in graph theory is the multiway cut problem. In this setting, an undirected graph is formed where the nodes are the data points and the edges represent an affinity between each pair of nodes. The multiway cut problem (also called k-way cut) consists of partitioning the graph into k disjoint sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_k = \emptyset$ such that points in the same cluster are maximally similar and points in different clusters are maximally dissimilar. In general, the k-way cut is NP-Hard due to the combinatorial nature of the problem. However, useful relaxations have been proposed in the literature [15], [2], [1] leading to eigenvalue problems where particular eigenvectors are the relaxed solutions.

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).

Consider a dataset $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ with similarity matrix S where S_{ij} measures the similarity between x_i and x_j, D is the graph degree matrix $D = \text{diag}(\text{deg}_1, \ldots, \text{deg}_N)$ and $\text{deg}_i = \sum_{j=1}^N S_{ij}$ is the degree of the i-th node. The k-way normalized cut relaxation [16] can be defined as

$$\max_G \text{Tr}(G^T L G)$$
$$\text{such that } G^T G = I_k,$$

where $G \in \mathbb{R}^{N \times k}$ is a cluster indicator matrix, L is the normalized Laplacian $L = D^{-1/2} S D^{-1/2}$ and I_k is the $k \times k$ identity matrix. This maximization problem is a special form of Fan's theorem [17]. The optimal solution of the relaxed problem is given by $G^\star = BR_2$ where $B \in \mathbb{R}^{N \times k}$ is any orthonormal basis of the k-th principal subspace of L and $R_2 \in \mathbb{R}^{k \times k}$ is an arbitrary orthogonal matrix [17], [18]. This means that the relaxed solution is contained on the eigenvectors corresponding to the k largest eigenvalues of the normalized Laplacian L but the eigenvectors can have arbitrary rotations.

A Markov random walks view of spectral clustering was discussed in [4], [19]. A random walk on a graph consists of random jumps from vertex to vertex. The ij-th entry of the stochastic transition matrix P represents the probability of moving from node i to node j in one step. This transition matrix is defined as $P = D^{-1}S$. The corresponding eigenvalue problem becomes $Pr = \xi r$. The random walks algorithm 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.

III. Spectral Clustering and Weighted Kernel PCA

In this Section, we summarize the multiway spectral clustering method introduced in [7]. 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 desired clusters k, the following primal problem can be formulated [7]:

$$\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2N} \sum_{l=1}^{k-1} \gamma_l e^{(l)^T} V e^{(l)} - \frac{1}{2} \sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} \qquad (1)$$
$$\text{such that } \begin{cases} e^{(1)} = \Phi w^{(1)} + b_1 1_N \\ \vdots \\ e^{(k-1)} = \Phi w^{(k-1)} + b_{k-1} 1_N \end{cases}$$

where $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]$ is the compact form of the projected variables $e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l, i = 1, \ldots, N, l = 1, \ldots, k-1$, $V = \text{diag}([v_1, \ldots, v_N]), v_i \in \mathbb{R}^+$ is a user-defined weighting matrix, $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the mapping to high-dimensional feature space (which can be infinite dimensional), γ_l are regularization parameters, b_l are bias term and $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$ is the $N \times n_h$ feature matrix.

Each score variables vector $e^{(l)}$ induces a binary clustering with cluster indicator $q^{(l)} = \text{sign}(e^l), l = 1, \ldots, k-1$. The final clustering can be obtained from the $k-1$ binary cluster indicators by using encoding and decoding schemes similar to multiclass formulations for LS-SVM classifiers [9].

The dual form of the constrained optimization problem (1) was shown to be the eigenvalue problem [7]:

$$V M_V \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, \qquad (2)$$

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:

$$M_V = I_N - \frac{1}{1_N^T V 1_N} 1_N 1_N^T V$$

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 score variables can now be written in terms of the dual variables:

$$e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l = \sum_{j=1}^N \alpha_j^{(l)} K(x_i, x_j) + b_l, \qquad (3)$$

$l = 1, \ldots, k-1$ where $e_i \in \mathbb{R}^{k-1}$ is the representation of x_i in the score variables space. The bias terms become:

$$b_l = -\frac{1}{1_N^T V 1_N} 1_N^T V \Omega \alpha^{(l)}, \qquad (4)$$

$l = 1, \ldots, k-1$.

Note that setting the weighting matrix to $V = I_N$ leads to the kernel PCA eigenvalue problem which is known to lack discriminatory information for spectral clustering. On the other hand, by setting V to the inverse of the graph degree matrix $V = D^{-1}$, the eigenvalue problem in (2) becomes a centered version of the random walk model for spectral clustering [4]. The matrix M_V removes the weighted mean from the training points in the feature space and the weights are given by $D^{-1}1_N$.

B. From Eigenvectors to Clusters: Encoding

Once the eigenvectors $\alpha^{(l)}$ and the cluster indicators for training data $q^{(l)}$ have been computed, the final clustering can be obtained by encoding the $k-1$ binary indicators[1]. The codebook $\mathcal{C} = \{c_p\}_{p=1}^k$ can then be obtained from the rows of the binarized eigenvector matrix

$$A = [\text{sign}(\alpha^{(1)}), \ldots, \text{sign}(\alpha^{(1)})], l = 1, \ldots, k-1,$$

with $A \in \{-1, 1\}^{N \times (k-1)}$. Thus, the i-th training point x_i has an associated binary encoding represented by the i-th row of A. If the data are well-represented and the eigenvectors are piecewise constant then there will be exactly k different encodings which form the codebook. However, if the eigenvectors are approximately piecewise constant, then

[1]The cluster indicators for training data $q^{(l)}$ can be obtained by binarizing either the score variables $e^{(l)}$ or the eigenvectors $\alpha^{(l)}$ since $\text{sign}(e_i^{(l)}) = \text{sign}(\alpha_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$. [7]

there will be spurious encodings. In this case, the codebook is formed by the k encodings with most ocurrences.

C. From Projections to Clusters: Decoding

One of the main advantages of using the weighted kernel PCA framework for spectral clustering is that the clustering model can be extended naturally to out-of-sample data. Given a set of N_{test} test points $\{x_t^{\text{test}}\}_{t=1}^{N_{\text{test}}}$, the score variables for these points become:

$$z_t^{(l)} = {w^{(l)}}^T \varphi(x_t^{\text{test}}) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_t^{\text{test}}, x_j) + b_l, \quad (5)$$

and the cluster indicators are given by $\breve{Q}$ with tl-entry:

$$\breve{Q}_{tl} = \breve{q}_t^{(l)} = \text{sign}(z_t^{(l)}), \quad (6)$$

where $\breve{Q} \in \{-1, 1\}^{N_{\text{test}} \times (k-1)}$ is the cluster indicator matrix for test data. The decoding step consists of comparing the cluster indicators (i-th row of $\hat{A}$ for training data, t-th row of $\breve{Q}$ for test data) with respect to the codebook and selecting the nearest codeword in terms of Hamming distance. This procedure corresponds to the ECOC decoding [20].

IV. Spectral Clustering with Pairwise Constraints

In this Section, we extend the multiway spectral clustering algorithm introduced in [7] to deal with pairwise constraints.

A. Hard Pairwise Constraint

Intuitively, if a pair of data points should belong to the same cluster (*must-link* constraint) then these data points should be mapped close in the score variables space such that they have the same codeword. On the other hand, a *cannot-link* constraint should map the points far from each other in the score variables space such that they have different codewords.

Consider a pair of data points $x_a, x_d \in \mathcal{D}, a \neq d$ and a pairwise constraint between x_a and x_d represented by

$${w^{(l)}}^T \varphi(x_a) = c {w^{(l)}}^T \varphi(x_d), l = 1, \ldots, k-1, \quad (7)$$

where $c \in \{-1, 1\}$ defines the relationship between x_a and x_d in the score variables space and takes the value of either -1 or 1 for *cannot-link* and *must-link* constraints respectively. Note that, (7) represents a *hard* pairwise constraint requiring that x_a and x_d should be mapped exactly to the same point in the $\mathbb{R}^{k-1}$ score variables space.

Fixing V to be the inverse of the degree matrix D^{-1} and introducing (7) into (1) leads to the following primal problem:

$$\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2N} \sum_{l=1}^{k-1} \gamma_l {e^{(l)}}^T D^{-1} e^{(l)} - \frac{1}{2} \sum_{l=1}^{k-1} {w^{(l)}}^T w^{(l)} \quad (8)$$

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

The Lagrangian of (8) is:

$$\mathcal{L}(w^{(l)}, e^{(l)}, b_l; \alpha^{(l)}, \beta_l) = \frac{1}{2N} \sum_{l=1}^{k-1} \gamma_l {e^{(l)}}^T D^{-1} e^{(l)} - \frac{1}{2} \sum_{l=1}^{k-1} {w^{(l)}}^T w^{(l)} - \sum_{l=1}^{k-1} {\alpha^{(l)}}^T (e^{(l)} - \Phi w^{(l)} - b_l 1_N) - \sum_{l=1}^{k-1} \beta_l {w^{(l)}}^T (\varphi(x_a) - c\varphi(x_d)),$$

where $\alpha^{(l)} \in \mathbb{R}^N, \beta_l \in \mathbb{R}$ are Lagrange multipliers and $l = 1, \ldots, k-1$. The Karush-Kuhn-Tucker (KKT) conditions for optimality are:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \rightarrow w^{(l)} = \Phi^T \alpha^{(l)} - \beta_l (\varphi(x_a) - c\varphi(x_d)) \\ \frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \rightarrow \alpha^{(l)} = (\gamma_l / N) D^{-1} e^{(l)} \\ \frac{\partial \mathcal{L}}{\partial b_l} = 0 \rightarrow 1_N^T \alpha^{(l)} = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \rightarrow e^{(l)} = \Phi w^{(l)} + b_l 1_N \\ \frac{\partial \mathcal{L}}{\partial \beta_l} = 0 \rightarrow {w^{(l)}}^T (\varphi(x_a) - c\varphi(x_d)) = 0, \end{cases}$$

for $l = 1, \ldots, k-1$. Eliminating the primal variables $w^{(l)}, e^{(l)}, b_l$ and using Mercer's theorem $\varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$ where $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a positive definite function, leads to the following eigenvalue problem:

$$\lambda_l \alpha^{(l)} = D^{-1} M_D \Omega^\star \alpha^{(l)} \quad (9)$$

where $\lambda_l = N/\gamma_l, l = 1, \ldots, k-1$, M_D is the weighted centering matrix:

$$M_D = I_N - \frac{1}{1_N^T D^{-1} 1_N} 1_N 1_N^T D^{-1},$$

$\Omega^\star$ is the matrix:

$$\Omega^\star = \Omega - \frac{1}{k_1} (\Omega(:, a) - c\Omega(:, d))(\Omega(:, a) - c\Omega(:, d))^T,$$

Ω is the kernel matrix $\Omega_{ij} = K(x_i, x_j)$, $\Omega(:, i) \in \mathbb{R}^N$ is the i-th column of Ω and

$$k_1 = K(x_a, x_a) - 2cK(x_a, x_d) + c^2 K(x_d, x_d).$$

Note that the modified kernel matrix $\Omega^\star$ corresponds to a rank-1 downdate of the original kernel matrix Ω. This modification changes the metric according to the pairwise constraints and the equivalent kernel function can be written

as:

$$K^{\star}_{\text{eq}}(x_i,x_j) = K(x_i,x_j) - \frac{1}{k1}\big(K(x_i,x_a) - cK(x_i,x_d)\big)\big(K(x_j,x_a) - cK(x_j,x_d)\big). \quad (10)$$

The bias terms become

$$b_l = -\frac{1}{1_N^T D^{-1} 1_N} 1_N D^{-1}\Omega^{\star}\alpha^{(l)} \quad (11)$$

and the projected variables for an out-of-sample point x are:

$$z^{(l)}(x) = {w^{(l)}}^T \varphi(x) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K^{\star}_{\text{eq}}(x,x_j) + b_l,$$

for $l = 1,\ldots,k-1$.

A constrained spectral clustering algorithm from a completely different perspective was discussed in [14]. This method starts from a Gaussian process interpretation of the kernel matrix and leads to the same equivalent kernel (10). However, [14] only aims at modifying the kernel matrix with no clear extensions to out-of-sample points. Moreover, interaction between *cannot-link* constraints create false similarities and negative entries in the equivalent kernel matrix. This fact may cause data points having negative graph degrees and therefore invalidating the computation of a graph Laplacian.

B. Soft Pairwise Constraint via Regularization

Consider the following extension of (8) by introducing a *soft* pairwise constraint between x_a and x_d:

$$\min_{w^{(l)},e^{(l)},b_l,r_l} \frac{1}{2}\sum_{l=1}^{k-1}\frac{\gamma_l}{N}{e^{(l)}}^T D^{-1} e^{(l)} - {w^{(l)}}^T w^{(l)} - \rho r_l^2 \quad (12)$$

$$\text{such that } \begin{cases} e^{(1)} = \Phi w^{(1)} + b_1 1_N \\ \vdots \\ e^{(k-1)} = \Phi w^{(k-1)} + b_{k-1} 1_N \\ {w^{(1)}}^T \varphi(x_a) = c{w^{(1)}}^T \varphi(x_d) + r_1 \\ \vdots \\ {w^{(k-1)}}^T \varphi(x_a) = c{w^{(k-1)}}^T \varphi(x_d) + r_{k-1} \end{cases}$$

where $r_l, l = 1,\ldots,k-1$ are residual terms that allow some distance between the mapped x_a and x_b in the score variables space. The amount of discrepancy between the points is included in the objective function using a new regularization term $\rho \in \mathbb{R}^+$.

The Lagrangian of (8) becomes:

$$\mathcal{L}(w^{(l)},e^{(l)},b_l;\alpha^{(l)},\beta_l) = \frac{1}{2}\sum_{l=1}^{k-1}\big(\frac{\gamma_l}{N}{e^{(l)}}^T D^{-1} e^{(l)} - {w^{(l)}}^T w^{(l)} - \rho_l r_l^2\big) - \sum_{l=1}^{k-1}\alpha^{(l)^T}(e^{(l)} - \Phi w^{(l)} - b_l 1_N) - \sum_{l=1}^{k-1}\beta_l\big({w^{(l)}}^T(\varphi(x_a) - c\varphi(x_d)) - r_l\big),$$

with KKT conditions for optimality:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w^{(l)}} = 0 \rightarrow w^{(l)} = \Phi^T\alpha^{(l)} - \beta_l(\varphi(x_a) - c\varphi(x_d)) \\ \frac{\partial\mathcal{L}}{\partial e^{(l)}} = 0 \rightarrow \alpha^{(l)} = (\gamma_l/N)D^{-1}e^{(l)} \\ \frac{\partial\mathcal{L}}{\partial b_l} = 0 \rightarrow 1_N^T\alpha^{(l)} = 0 \\ \frac{\partial\mathcal{L}}{\partial r_l} = 0 \rightarrow \beta_l = \rho r_l \\ \frac{\partial\mathcal{L}}{\partial \alpha^{(l)}} = 0 \rightarrow e^{(l)} = \Phi w^{(l)} + b_l 1_N \\ \frac{\partial\mathcal{L}}{\partial \beta_l} = 0 \rightarrow {w^{(l)}}^T(\varphi(x_a) - c\varphi(x_d)) = r_l. \end{cases}$$

Eliminating the primal variables $w^{(l)}, e^{(l)}, b_l, r_l$ and expressing the KKT conditions in terms of the dual variables $\alpha^{(l)}$ leads to:

$$\lambda_l\alpha^{(l)} = D^{-1}M_D\Omega^{\star\star}\alpha^{(l)}, \quad (13)$$

$\lambda_l = N/\gamma_l$, $\Omega^{\star\star}$ is the matrix:

$$\Omega^{\star\star} = \Omega - k_2(\Omega(:,a) - c\Omega(:,d))(\Omega(:,a) - c\Omega(:,d))^T, \quad (14)$$

where

$$k_2 = \frac{\rho}{1+\rho k_1},$$

$l = 1,\ldots,k-1$. The equivalent kernel for this soft constrained problem becomes:

$$K^{\star\star}_{\text{eq}}(x_i,x_j) = K(x_i,x_j) - k_2\big(K(x_i,x_a) - cK(x_i,x_d)\big)\big(K(x_j,x_a) - cK(x_j,x_d)\big). \quad (15)$$

Note that incorporating a single pairwise constraint into the formulation leads to a modified metric induced by the equivalent kernel function. This means that multiple pairwise constraints can be incorporated by iteratively adding one constraint at a time and successively building the equivalent kernel matrix.

The bias terms become

$$b_l = -\frac{1}{1_N^T D^{-1} 1_N} 1_N D^{-1}\Omega^{\star\star}\alpha^{(l)} \quad (16)$$

and the projected variables for an out-of-sample point x:

$$z^{(l)}(x) = {w^{(l)}}^T \varphi(x) + b_l = \sum_{j=1}^{N} \alpha_j^{(l)} K^{\star\star}_{\text{eq}}(x,x_j) + b_l, \quad (17)$$

for $l = 1,\ldots,k-1$.

Note that, the equivalent kernel of the hard constrained formulation (10) is a particular case of the soft constrained formulation when $\rho \rightarrow \infty$. Moreover, if $\rho \rightarrow 0$ then $k_2 \rightarrow 0$ and the soft constrained formulation becomes the standard multiway spectral clustering formulation (1). Therefore, the regularization term ρ controls the importance of the available prior knowledge in terms of pairwise constraints.

An implementation of the proposed regularized approach (12) is shown in Algorithm 1.

V. A Constrained Model Selection Criterion

In this Section, we extend an existing model selection criterion for spectral clustering in order to incorporate pairwise constraints.

Algorithm 1 Regularized Spectral Clustering with Pairwise Constraints

Input: Training set $\mathcal{D} = \{x_i\}_{i=1}^N$, positive definite kernel function $K(x_i, x_j)$, number of clusters k, ρ_m, ρ_c regularization terms for *must-link* and *cannot-link* constraints respectively, pairwise constraint index set $\mathcal{J} = \{(a_n, b_n)\}_{n=1}^{N_{\text{const}}}, a_n, b_n \in \mathbb{N}$, where N_{const} is the total number of pairwise constraints, test set $\{x_t^{\text{test}}\}_{t=1}^{N_{\text{test}}}$.
Output: Partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, cluster codeset $\mathcal{C} = \{c_p\}_{p=1}^k, c_p \in \{-1, 1\}^{k-1}$.

1: Compute the kernel matrix Ω, $\Omega_{ij} = K(x_i, x_j), i, j = 1, \ldots, N$.
2: **for** $n = 1$ to N_{const} **do**
3: $x_a \leftarrow x_{a_n}, x_b \leftarrow x_{b_n}$ and let $c = -1$ for *cannot-link* and $c = 1$ for *must-link*.
4: Compute the equivalent kernel $\Omega^{\star\star}$ using Ω, ρ_m, ρ_c and (14).
5: $\Omega \leftarrow \Omega^{\star\star}$
6: **end for**
7: Compute the eigenvectors $\alpha^{(l)}, l = 1, \ldots, k-1$ corresponding to the largest $k-1$ eigenvalues of $D^{-1}M_D\Omega^{\star\star}$ where D^{-1} is the inverse graph degree matrix of $\Omega^{\star\star}$.
8: Binarize the eigenvectors: $\text{sign}(\alpha_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$, and let $\text{sign}(\alpha_i) \in \{-1, 1\}^{k-1}$ be the encoding vector for the training data point x_i, $i = 1, \ldots, N$.
9: Count the occurrences of the different encodings and find the k encodings with most occurrences. Let the codeset be formed by these k encodings: $\mathcal{C} = \{c_p\}_{p=1}^k, c_p \in \{-1, 1\}^{k-1}$
10: $\forall i$, 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.
11: Binarize the test data projections $\text{sign}(z_m^{(l)}), m = 1, \ldots, N_{\text{test}}, l = 1, \ldots, k-1$ using (17) and let $\text{sign}(z_m) \in \{-1, 1\}^{k-1}$ be the encoding vector of x_m^{test}, $m = 1, \ldots, N_{\text{test}}$.
12: $\forall m$, assign x_m^{test} to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\text{sign}(z_m), c_p)$.

A. The Balanced Line Fit

The Balanced Line Fit criterion (BLF) was discussed in [7]. This model selection criterion can be used to obtain kernel parameters and number of clusters k such that the clusters are represented as lines in the score variables space. This structural property of the projections has been shown to appear when the clusters are well-formed. The BLF can be applied to out-of-sample data which becomes important in terms of generalization and predictive purposes. Consider the following average measure of collinearity of a validation dataset with respect to a clustering $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}, k > 2$

$$\text{linefit}(k) = \frac{1}{k}\sum_{p=1}^{k} \frac{k-1}{k-2}\left(\frac{\zeta_1^{(p)}}{\sum_l \zeta_l^{(p)}} - \frac{1}{k-1}\right) \tag{18}$$

where $\zeta_1^{(p)} \geq \ldots \geq \zeta_{k-1}^{(p)}$ are the ordered eigenvalues of the sample covariance matrix:

$$C_{\tilde{Z}}^{(p)} = \frac{1}{|\mathcal{A}_p|}\tilde{Z}^{(p)^T}\tilde{Z}^{(p)}, p = 1, \ldots, k, \tag{19}$$

$\tilde{Z}^{(p)} \in \mathbb{R}^{|\mathcal{A}_p| \times (k-1)}$ is the matrix representing the zero mean score variables for validation data assigned to the p-th cluster. If the validation score variables for the p-th cluster are collinear, then all variance is contained on the first eigenvector and $\zeta_1^{(p)}/\sum_l \zeta_l^{(p)}$ equals 1. On the other hand, if the total variance is evenly distributed on the eigenvectors then $\zeta_1^{(p)}/\sum_l \zeta_l^{(p)}$ equals $1/(k-1)$.

When $k = 2$ only one eigenvector is needed to obtain a binary clustering. Consider the matrix $\breve{Z}^v \in \mathbb{R}^{N_v \times 2}$ where the m-th entry in the first column corresponds to the score variables $z_m = \sum_{i=1}^N \alpha_i K(x_i, x_m^v)$ and the m-th entry in the second column to $\sum_{i=1}^N K(x_i, x_m^v)$. This modification allows the application of the linefit criterion when $k = 2$:

$$\text{linefit}(k) = \begin{cases} \sum_{p=1}^2 \left(\frac{\breve{\zeta}_1^{(p)}}{\breve{\zeta}_1^{(p)} + \breve{\zeta}_2^{(p)}} - \frac{1}{2}\right) & \text{if } k = 2 \\ \frac{1}{k}\sum_{p=1}^k \frac{k-1}{k-2}\left(\frac{\zeta_1^{(p)}}{\sum_l \zeta_l^{(p)}} - \frac{1}{k-1}\right) & \text{if } k > 2 \end{cases} \tag{20}$$

where $\breve{\zeta}_l^{(p)}$ are the ordered eigenvalues of the matrix $\breve{Z}^{(p)} \in \mathbb{R}^{|\mathcal{A}_p| \times 2}$ representing the zero mean score variables for validation data assigned to the p-th cluster when $k = 2$.

In real-life problems, the balance of the obtained clusters also becomes important. Consider the following clustering balance measure:

$$\text{balance}(k) = \frac{\min\{|\mathcal{A}_1|, \ldots, |\mathcal{A}_k|\}}{\max\{|\mathcal{A}_1|, \ldots, |\mathcal{A}_k|\}}. \tag{21}$$

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:

$$\text{BLF}(k) = \eta\text{linefit}(k) + (1-\eta)\text{balance}(k), \tag{22}$$

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

B. The Constrained BLF

Introducing an additional term measuring the accuracy of fulfilling the pairwise constraints into the BLF formula leads to the Constrained BLF (cBLF):

$$\text{cBLF}(k) = \eta_l\text{linefit}(k) + \eta_b\text{balance}(k) + \eta_c\text{cfit}(k), \tag{23}$$

where $\text{cfit}(k)$ is the ratio of fulfilled constraints with respect to the total number of constraints, η_l, η_b, η_c are given parameters controlling the importance of the linefit, balance and constraint fit respectively and $\eta_l + \eta_b + \eta_c = 1$.

VI. Experimental Results

In this section we report experimental results. The assessment of the clustering is done using the adjusted Rand index [21]. This external clustering performance measures the degree of mismatch between some known cluster indicators and the results of a clustering algorithm. The adjusted Rand index ranges from 0 in the case of complete mismatch to 1 in the case of a perfect match. The regularization term ρ can be different for *must-link* and *cannot-link* constraints and can be also incorporated into a model selection procedure. However, in this work we fixed this parameter to $\rho_m = 10^{-1}$ for *must-link* constraints and $\rho_c = 10^{-3}$ for *cannot-link* constraints. We choose ρ_c much smaller than ρ_m to avoid negative entries in the equivalent kernel matrix due to rank-1 downdates. For the BLF, η was fixed to 0.75 and for the cBLF, $\eta_l = 0.5, \eta_b = 0.25, \eta_c = 0.25$. All experiments were performed on a MacBook Intel Core 2 Duo 2.4 GHz, 4 GB RAM in Matlab 7.7.

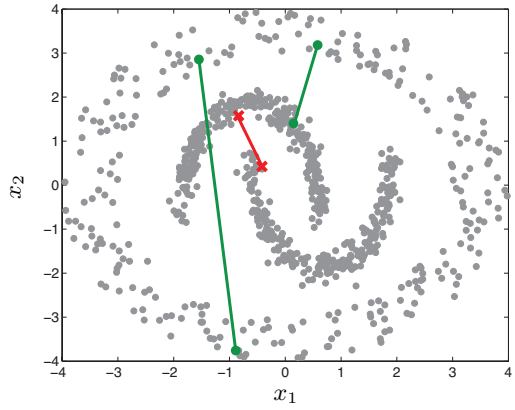


Fig. 1. Toy problem 1 - Training set with constraints. *Must-link* constraints are shown in green with circles and *cannot-link* constraints are shown in red with crosses.

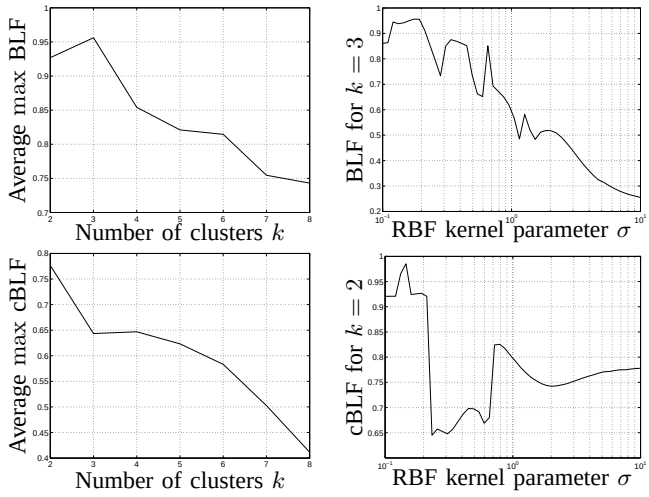


Fig. 2. Toy problem 1 - Model selection on validation data using the BLF and the cBLF. **Top row:** Tuning without constraints (BLF). **Bottom row:** Tuning with 3 pairwise constraints (cBLF). Note that the constraints shift the cBLF to be maximal at 2 clusters instead of 3 clusters like in the unconstrained case.

A. Toy Problems

The first toy problem of two moons and a ring for a total of $1,500$ data points. The training scenario consists of 500 data points for training and 500 data points for validation. Figure 1 shows the training set and the pairwise constraints. *Must-link* constraints are shown in green with circles and *cannot-link* constraints are shown in red with crosses. Figure 2 shows the model selection plots. The number of clusters k and the RBF kernel parameter σ^2 are obtained through the BLF using the validation set. For the original unconstrained problem, we used the BLF and for the problem with constraints, we used the cBLF. Figure 3 shows the clustering results.

The second toy problem consists of four Gaussian clouds in $2D$ for a total of 900 data points. Model selection was performed in a similar way as in the toy problem 1. Figure 4 shows the complete set with constraints. The constraints impose a "XOR"-like solution. Note that only two constraints were required to solve this problem.

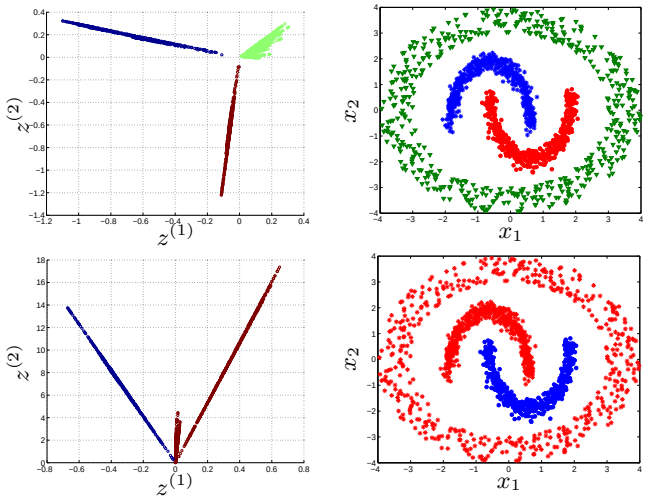


Fig. 3. Toy problem 1 - Clustering results after tuning. **Top row:** Original problem without constraints. **Bottom row:** Clustering problem with pairwise constraints. The first column shows the score variables space. Note the strong line structures detected by the model selection criteria. The constraints cause that the line structure corresponding to the ring moves to the same quadrant as the upper moon and therefore belonging to the same cluster.

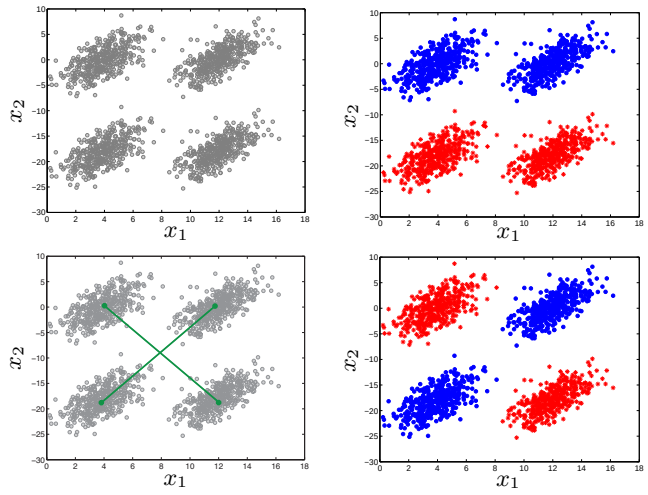


Fig. 4. Toy problem 2 - Clustering results. **Top row:** Original problem without constraints. **Bottom row:** Clustering problem with pairwise constraints. The clustering parameters (k, σ^2) were obtained through the BLF and cBLF in a similar fashion as the toy problem 1. Note that only two constraints were required to solve this problem.

The third toy problem consists of the Ripley dataset. The goal is to obtain a classification decision by iteratively increasing the number of pairwise constraints in a semi-supervised way. In this case, we compared with the constrained spectral clustering method described in [14]. However, since this method does not have out-of-sample extensions, we combine it with the Nyström method discussed in [5] to be able to use a smaller set of data points for training and then approximate the full eigenvectors. The dataset set was divided into 250 points for training, 500 for validation and 500 for testing. The RBF kernel parameter σ^2 was

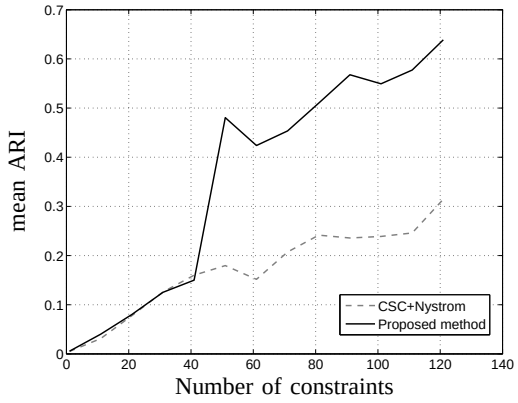


Fig. 5. Toy problem 3 - Ripley dataset. Mean adjusted Rand index (ARI) of the full dataset ($1,250$ points) with respect of numbers of pairwise constraints over 10 randomizations of the constraints. The training set was fixed to 250 points. **Dashed line:** Constrained spectral clustering algorithm of [14] together with the Nyström approximation [5]. **Solid line:** Proposed method. The proposed method outperforms [14] in terms of agreement with the true labels.

obtained through the cBLF each time a pairwise constraint was added. The pairwise constraint set was randomized 10 times. Figure 5 shows the average adjusted Rand index (ARI) with respect to the true labels in terms of increasing number of constraints. The proposed method achieves an ARI of 0.63 with 125 pairwise constraints while [14] achieves an ARI of 0.31. Figure 6 shows the results of the full Ripley dataset after model selection on validation data. The proposed method outperforms [14] in terms of agreement with the true labels.

B. Image Segmentation

Image segmentation is a difficult problem for spectral clustering due to the large number of points and the size of the matrices involved in the eigendecomposition. In this experiment, we used an image from the Berkeley image database [22]. 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 $h^{(j)}$ [23] $\chi^2_{ij} = 0.5\sum_{b=1}^{B}(h_b^{(i)} - h_b^{(j)})^2/(h_b^{(i)} + h_b^{(j)})$, where B is total number of quantization levels. The histograms are assumed to be 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 a robust and efficient way of comparing the similarity between two histograms for color discrimination and image segmentation [5]. The image is 321×481 pixels wide giving a total number of data points of $154,401$. The training scenario is 600 pixel histograms for training, $3,000$ for model selection using the cBLF and the remaining pixel histograms are the test set. Note that [14] alone cannot cope with the full image since it requires storing and computing the eigendecomposition of a $154,401 \times 154,401$ kernel matrix. Hence, we use the Nyström method for spectral clustering described in [5] to approximate the eigenvectors and be able to compare with

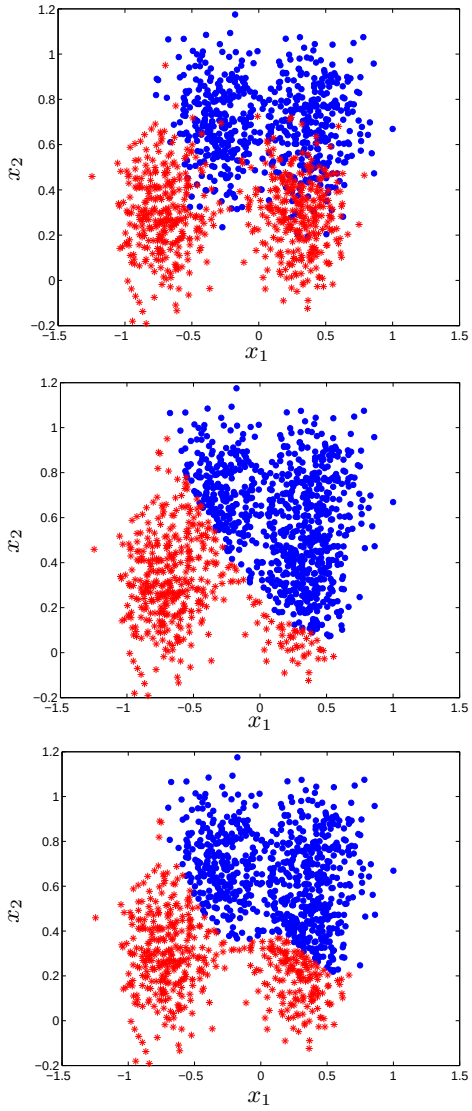


Fig. 6. Toy problem 3 - Ripley dataset results after model selection using the cBLF on validation data. **Top:** True classification. **Center:** Results of [14] together with the Nyström approximation [5] using 125 constraints. The ARI with respect to the true labels is 0.31. **Bottom:** Proposed method using 125 constraints. The ARI is 0.63. The proposed method outperforms [14] in terms of agreement with the classification labels.

the proposed method. The computation time of the proposed method was in average 15 seconds while [14] combined with [5] was in average 170 seconds. Figure 7 shows the original image with constraints and the results of [14] and the proposed method. The proposed method produces a visually appealing clustering with respect to the given constraints.

VII. Conclusions

A regularized formulation to incorporate pairwise constraints into spectral clustering is proposed. This formulation is based on a weighted kernel PCA framework in which the clustering model can be extended to new points. The proposes approach in cast in a constrained optimization framework allowing pairwise constraints to be naturally



Fig. 7. Image Segmentation - Clustering results. The image is 321×481 pixels wide, for a total of $154,401$ data points. The training set consists of 600 randomly selected pixel histograms and the cluster model was extended to the remaining data points using the out-of-sample extension. **Left:** Original image with constraints. **Center:** Clustering results of [14] together with the Nyström approximation [5]. Note that [14] alone cannot cope with the large number of data points of this image. **Right:** Proposed method. The proposed method produces a visually appealing clustering with respect to the given constraints.

added in the primal problem. The corresponding dual problem is an eigenvalue problem of a modified kernel matrix where the constraints correspond to rank-1 downdates of the original affinity. The proposed method can be trained, validated and tested in a learning framework using a model selection criterion. Simulations results with toy examples and an image segmentation problem show the applicability of the proposed method.

ACKNOWLEDGMENT

This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD / Postdocs & 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, ANMMM; 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 (ISMC/IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). 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] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[3] 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.

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

[5] 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.

[6] Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[7] 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. In Press*, 2009.

[8] ——, "A weighted kernel PCA formulation with out-of-sample extensions for spectral clustering methods," in *Proceedings of the 2006 International Joint Conference on Neural Networks (IJCNN'06)*, 2006, pp. 138–144.

[9] 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.

[10] J. A. K. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor, "A support vector machine formulation to PCA analysis and its kernel version," *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 447–450, Mar. 2003.

[11] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004, pp. 59–68.

[12] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 577–584.

[13] S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral learning," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003, pp. 561–566.

[14] Z. Lu and M. Carreira-Perpiñán, "Constrained spectral clustering through affinity propagation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008.

[15] 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.

[16] 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.

[17] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations." *Proc. Natl. Acad. Sci. USA*, vol. 35, no. 11, pp. 652–655, 1949.

[18] 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.

[19] M. Hein, J.-Y. Audibert, and U. von Luxburg, "Graph Laplacians and their convergence on random neighborhood graphs," *Journal of Machine Learning Research*, vol. 8, pp. 1325–1370, 2007.

[20] 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.

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

[22] 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.

[23] 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.