WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia

IJCNN

A Semi-Supervised Formulation to Binary Kernel Spectral Clustering

Carlos Alzate and Johan A. K. Suykens

***Abstract*—A semi-supervised formulation to binary kernel spectral clustering is presented. The formulation fits in a constrained optimization setting with primal and dual model representations. The clustering model can be applied naturally to out-of-sample points allowing model selection and achieving good generalization capabilities. The proposed method incorporates labeled information into the core binary kernel spectral clustering by adding an extra term into the objective function together with a regularization constant. The resulting dual problem is no longer an eigenvalue problem as in the case of the original core model but a linear system. A model selection criterion combining a cluster distortion measure on the unlabeled part and the classification accuracy on the labeled part is also presented. This criterion can be used to obtain clustering parameters such that the clustering model evaluated at validation points display a desirable structure. Simulation results with toy data and real benchmark datasets show the applicability of the proposed method.**

I. Introduction

SPECTRAL clustering methods correspond to a family of unsupervised learning algorithms to group data points that are similar with respect to some given measure [1], [2], [3], [4], [5]. Their solutions can be obtained from an eigenvalue decomposition of a Laplacian matrix derived from the data. These method have been shown to outperform classical clustering algorithms such as k-means, especially in cases of high dimensionality and high nonlinearity. Classical spectral clustering has several issues concerned with the lack of an underlying model. Out-of-sample extensions do not follow naturally from the eigenvalue problem and should rely on approximations such as the Nyström method [6]. Model selection is also problematic and important clustering parameters such as the number of clusters, the similarity function used and its parameters are often set heuristically.

In order to overcome these issues, a different interpretation of spectral clustering called kernel spectral clustering (KSC) has been proposed in [7]. This formulation can be seen as a weighted version of kernel PCA in a constrained optimization setting with primal and dual model representations typical of least squares support vector machines (LS-SVM) [8]. The dual problem is related to the random walks algorithm for spectral clustering [9] for a particular choice of weights. By defining a clustering model, it is possible to obtain out-of-sample extensions and to perform model selection in a learning framework. A method to estimate eigenvectors for out-of-sample data is proposed in [10]. When the data display strong cluster structures, the eigenvectors show a structural property making them informative to obtain the underlying grouping of the data. This property is exploited to perform model selection by using unseen data and a model selection measure based on the Fisher criterion [11] has been proposed in [10].

Semi-supervised learning [12], [13] situates in between supervised and unsupervised learning. Most semi-supervised learning algorithms start by extending strategies either for unsupervised or for supervised learning towards the incorporation of information from the other learning scheme. In most cases, there is an abundance of unlabeled examples while the supervised information is scarce. For semi-supervised learning to work, several assumptions have to be made: *(i) smoothness:* if two points in a high-density zone are close to each other, the corresponding outputs should be close too; *(ii) low density separation:* the decision boundary should be in a low-density region; *(iii) manifold:* the input data lie roughly on a low-dimensional manifold.

In this paper, we propose a formulation for semi-supervised learning in the context of binary kernel spectral clustering. We start with the core binary clustering model and extend it towards the incorporation of labeled information which is assumed to be scarce. The binary core model is extended by adding a regularized term quantifying the squared distance between the projections of the labeled data points and their corresponding labels. The obtained dual problem is a linear system instead of a eigenvalue problem as in the original KSC core model. The incorporation of the labeled information transforms the eigenvalue problem into a linear system. This effect has also been discussed in [14] for visualization purposes. A model selection methodology called semi-supervised Fisher (SSF) criterion is also presented. The proposed measure is a weighted sum between the standard Fisher criterion [10] on the unlabeled part of the data and the classification accuracy for the labeled data. The objective is to obtain clustering parameters by maximizing the SSF criterion on validation data.

This paper is organized as follows: Section II summarizes kernel spectral clustering introduced in [7]. Section III contains the main contributions of this work: the proposed formulation to semi-supervised binary kernel spectral clustering together with a new model selection measure based on the Fisher criterion. An algorithm is also provided. In Section IV, we present the experimental results with toy, real and benchmark datasets. The proposed method is also compared with respect to the Laplacian SVM [15], [16] and in Section V we give concluding and future work comments.

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

II. Kernel Spectral Clustering

This Section summarizes the kernel spectral clustering framework introduced in [7]. The formulation puts spectral clustering in a constrained optimization setting allowing out-of-sample extensions. This method can be seen as a weighted version of kernel PCA where cluster decisions for unseen data points can be computed using the projections onto the eigenvectors solution. Consider a set of training data points $\mathcal{X} = \{x_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^d$. The objective of clustering is to partition the dataset $\mathcal{X}$ into k groups such that data points in the same group are more similar than data points belonging to different groups. The following clustering model can be adopted:

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

with unknowns $w^{(l)} \in \mathbb{R}^{d_h}, b_l \in \mathbb{R}$, $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$ which is a mapping to a high-dimensional feature space of dimension d_h. The projections $e_i^{(l)}$ can be written in compact form:

$$e^{(l)} = \Phi w^{(l)} + b_l 1_N = \left[e_1^{(l)}, \ldots, e_N^{(l)}\right]^T, l = 1, \ldots, n_e$$

where $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T], \Phi \in \mathbb{R}^{N \times d_h}$ and 1_N is a vector of N ones. The $e^{(l)}$ vectors represent the latent variables of a set of n_e binary clustering indicator vectors given by sign$(e^{(l)})$ which will be combined at a later stage in order to obtain the final k clusters. This model is inspired in multiclass kernel machines [8], [17].

A. Primal and Dual Formulation

The primal problem of kernel spectral clustering is defined as follows [7]:

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

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

where γ_l are regularization parameters and $V = \text{diag}([v_1, \ldots, v_N]), v_i \in \mathbb{R}^+$ is a user-defined weighting matrix. Using the Lagrangian and the Karush-Kuhn-Tucker (KKT) optimality conditions leads to the following dual eigenvalue problem:

$$V M_V \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, \quad (3)$$

where $\lambda_l = N/\gamma_l$ are the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{n_e}$, $\alpha^{(l)} \in \mathbb{R}^N$ are corresponding eigenvectors, M_V is a centering matrix $M_V = I_N - (1/1_N^T V 1_N) 1_N 1_N^T V$, Ω is the training kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j)$, $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a kernel function satisfying Mercer's condition, thus, $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j), i, j = 1, \ldots, N$. The bias terms b_l becomes $b_l = -(1/1_N^T V 1_N) 1_N^T V \Omega \alpha^{(l)}$. The clustering model evaluated at the training data points can now be written in terms of the eigenvectors (dual variables):

$$e^{(l)} = \Phi w^{(l)} + b_l 1_N = M_V \Omega \alpha^{(l)}, l = 1, \ldots, n_e.$$

B. Relation to Spectral Clustering and the Choice of n_e

It was shown in [7] that if $V = D^{-1} = \text{diag}(1/d_1, \ldots, 1/d_N)$ where $d_i = \sum_{j=1}^N K(x_i, x_j)$ is the degree of the i-th data point, then the eigenvectors with large eigenvalue of the dual problem (3) contain information about the underlying grouping of the data. It is assumed here that the kernel function $K(x_i, x_j)$ acts as a non-negative similarity function[1]. Let us consider an ideal clustering scenario:

(a) The data contain k clusters denoted as $\Delta = \{\mathcal{A}_p\}_{p=1}^k, k > 1$.
(b) $K(x_i, x_j) > 0$ if x_i and x_j belong to the same cluster.
(c) $K(x_i, x_j) = 0$ if x_i and x_j are in different clusters.

In this situation, the properties of the eigenspectrum of $D^{-1} M_D \Omega$ relevant to clustering are summarized as follows:

1) The geometric multiplicity of the maximal eigenvalue (eigenvalue 1) is $k - 1$.
2) The $k-1$ eigenvectors with eigenvalue 1 can be written as linear combinations of the k indicator vectors of Δ.

These two properties mean that the eigenvectors with eigenvalue 1 are *piecewise constant* on the partitioning, that is, if $\alpha^{(l)} \in \mathbb{R}^N, l = 1, \ldots, k-1$ is a piecewise constant eigenvector then $\alpha_i^{(l)} = \alpha_j^{(l)} = c_p^{(l)}$ when $x_i, x_j \in \mathcal{A}_p$ and $c_p^{(l)}$ is the constant value of the l-th eigenvector associated to the p-th cluster. Due to the fact that the eigenspace of 1 is spanned by piecewise constant eigenvectors, the clusters have a localized representation in this new space: data points in the same cluster will be mapped to exactly the same point in the eigenspace. However, the condition stating that $K(x_i, x_j) = 0$ when x_i and x_j are in different clusters is not fulfilled in practical applications. Moreover, kernel functions used in clustering such as the RBF kernel do not become exactly zero. In this case, if the kernel between points in different clusters is non-zero but small enough, the eigenvectors will be *approximately* piecewise constant still preserving the localized representation of the clusters [5], [9]. Since only the $k-1$ top eigenvectors of $D^{-1} M_D \Omega$ contain information useful for clustering, the number of additional projections n_e defined in the clustering model (1) is equal to $k-1$. At this point, classical and kernel spectral clustering algorithms share the approximately piecewise constant property of their eigenvectors. However, there also exist fundamental differences which are summarized as follows.

C. Differences with Classical Spectral Clustering

Kernel spectral clustering can be seen as a weighted version of kernel PCA where the weighting matrix is set to D^{-1}. The clustering model is defined in the dual as projections onto the eigenvectors. Having a primal and dual model for performing spectral clustering is the main difference with respect to classical algorithms where no model is defined. The clustering model allows out-of-sample extensions and model selection criteria, which are important for obtaining

[1] Popular kernel functions such as the RBF kernel fulfill the non-negativity condition.

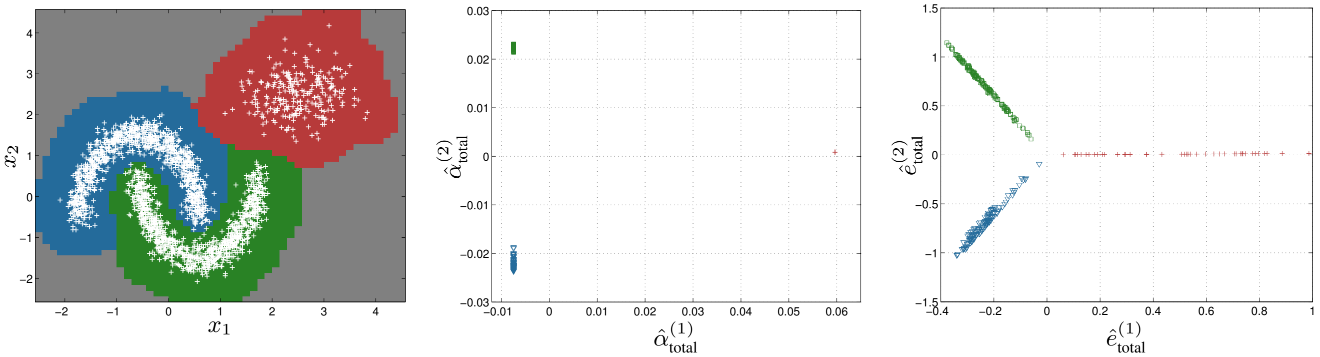

Fig. 1. Main idea behind kernel spectral clustering. **Left:** Clustering results and boundaries. A random subsample of the full dataset ($N_{\text{total}} = 2,250$) was used to create the training set ($N = 600$). The clustering parameters $k = 3$ and $\sigma^2 = 0.08$ were determined using grid search and the Fisher criterion on validation data ($N_{\text{val}} = 800$). The boundaries display good generalization capabilities together with non-linear decision regions. **Center:** Estimated out-of-sample eigenvectors of the full dataset showing a localized representation of the clusters. The Fisher criterion value is high in this case. **Right:** Estimated projections of the full dataset showing cluster collinearity. The BLF criterion is high when the projections display collinearity.

good generalization capabilities and reducing the computational complexity as it has been shown in [7]. Another important difference comes from the use of bias terms in the clustering model. The bias terms induce a special centering matrix leading to piecewise constant eigenvectors with zero mean. This property together with the orthogonality of the eigenvectors lead to data points in the same cluster mapped to the same coordinates in the eigenspace and different clusters mapped to different orthants. This additional result allows us to obtain the final k groups by using the sign patterns of the projections since they will be different for every cluster. This step is typically done in classical spectral clustering by using k-means.

D. *Out-of-Sample Extensions and Model Selection*

Given an out-of-sample data point x, the clustering model becomes:

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

$l = 1, \ldots, k-1$. The projections for out-of-sample points drawn i.i.d. from the same probability distribution as the training points display a special structure when the eigenvectors $\alpha^{(l)}$ are (approximately) piecewise constant. Namely, data points belonging to the same cluster are collinear in the projection space [7]. This structural property of the projections allows model selection by selecting clustering parameters leading to cluster collinearity in the projection space. A criterion to select good values for the clustering parameters was introduced in [7]. The Balanced Line Fit (BLF) criterion measures average collinearity and balanced of the obtained clusters on validation data. Another model selection method based on the Fisher criterion was first discussed in [10]. This criterion is applied on estimated out-of-sample eigenvectors which display a localized structure (instead of collinearity) when the clusters are well-formed. The main idea behind kernel spectral clustering is shown in Figure 1.

III. Semi-Supervised Formulation for Binary Kernel Spectral Clustering

Consider now a set of training data points $\mathcal{X} = \{x_1, \ldots, x_N, x_{N+1}, \ldots, x_M\}$ where $x_i \in \mathbb{R}^d$ and $i = 1, \ldots, M$. Classification labels are available for the last $N_{\text{L}} = M - N$ data points: $\mathcal{Y} = \{y_{N+1}, \ldots, y_M\}$, $y_m \in \{-1, 1\}, m = N+1, \ldots, M$. The objective is to incorporate the labeled information into the clustering problem. From now on, we consider the core binary kernel spectral clustering model, that is, equation (2) with $n_e = 1$ to go inline with the binary labeled information. This binary core model can be extended by adding a term and an extra regularization constant $\rho \in \mathbb{R}^+$ into the objective function.

A. *Primal and Dual Formulation*

The proposed primal problem becomes:

$$\min_{w,e,b} \frac{1}{2} w^T w - \frac{\gamma}{2} e^T D^{-1} e + \frac{\rho}{2} \sum_{m=N+1}^{M} (e_m - y_m)^2 \quad (4)$$
$$\text{such that } e = \Phi w + b 1_M.$$

This extra term quantifies the squared distance between the projections of the labeled data points and their corresponding labels. The primal problem can be written in matrix form as:

$$\min_{w,e,b} \frac{1}{2} w^T w - \frac{\gamma}{2} e^T D^{-1} e + \frac{\rho}{2} \left(e^T A e - 2 y^T e + y^T y \right) \quad (5)$$
$$\text{such that } e = \Phi w + b 1_M,$$

where

$$A = \left[\begin{array}{c|c} 0_{N \times N} & 0_{N \times N_{\text{L}}} \\ \hline 0_{N_{\text{L}} \times N} & I_{N_{\text{L}}} \end{array} \right] \in \mathbb{R}^{M \times M},$$

$y = [0, \ldots, 0, y_{N+1}, \ldots, y_M]^T \in \{-1, 0, 1\}^M$ and $I_{N_{\text{L}}}$ is the $N_{\text{L}} \times N_{\text{L}}$ identity matrix. The Lagrangian of the constrained optimization problem (5) becomes:

$$\begin{aligned} \mathcal{L}(w, e, b; \alpha) = & \frac{1}{2} w^T w - \frac{1}{2} e^T (\gamma D^{-1} - \rho A) e - \rho y^T e \\ & + \frac{\rho}{2} y^T y + \alpha^T (e - \Phi w - b 1_M), \end{aligned} \quad (6)$$

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

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \Phi^T \alpha \\ \dfrac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \alpha = (\gamma D^{-1} - \rho A) e + \rho y \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow 1_M^T \alpha = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \alpha} = 0 \rightarrow e = \Phi w + b 1_M, \end{cases} \tag{7}$$

after eliminating the primal variables w and e, the following dual problem is obtained:

$$\Big(I_M - \big(\gamma D^{-1} - \rho A\big) M_S \Omega \Big) \alpha = \rho M_S^T y, \tag{8}$$

where I_M is the $M \times M$ identity matrix and $M_S \in \mathbb{R}^{M \times M}$ is a centering matrix defined as:

$$M_S = I_M - \frac{1}{c} 1_M 1_M^T (\gamma D^{-1} - \rho A),$$

and $c = 1_M^T (\gamma D^{-1} - \rho A) 1_M$. The dual problem (8) corresponds to a linear system to be solved in α. Note that, the primal problem (5) is in general non-convex due to the minus sign in the objective function. This means that the KKT optimality conditions characterize the stationary points of the Lagrangian of (5). The linear system (8) has a unique solution when the left-hand size matrix is full-rank which is dependent on the regularization parameters γ and ρ. The choice of these parameters will be discussed later. The bias term becomes:

$$b = -\frac{1}{c} \big(1_M^T (\gamma D^{-1} - \rho A) \Omega \alpha + \rho 1_M^T y \big). \tag{9}$$

The latent clustering model for training points can now be written in terms of the dual variables α:

$$e = \Phi w + b 1_M = M_S \Omega \alpha - \frac{\rho}{c} 1_M 1_M^T y, \tag{10}$$

and the binary cluster membership is determined by sign(e).

B. Out-of-Sample Extension

The clustering membership of an arbitrary data point x can be determined by first computing its out-of-sample extension:

$$\begin{aligned} \hat{e}(x) &= w^T \varphi(x) + b \\ &= \big(\theta(x)^T - \frac{1}{c} 1_M^T (\gamma D^{-1} - \rho A) \Omega \big) \alpha - \frac{\rho}{c} 1_M^T y, \end{aligned} \tag{11}$$

where $\theta : \mathbb{R}^d \rightarrow \mathbb{R}^M$, $\theta(\cdot) = [K(x_1, \cdot), \ldots, K(x_M, \cdot)]^T$ and then binarizing it: sign($\hat{e}(x)$).

C. Properties of the solution vector α and choice of γ, ρ

Consider now for simplicity, the proposed formulation (5) without a bias term. In this case, the simplified dual becomes:

$$\Big(I_M - \big(\gamma D^{-1} - \rho A\big) \Omega \Big) \alpha = \rho y.$$

Rewriting it leads to:

$$\alpha = \gamma D^{-1} \Omega \alpha - \rho A \Omega \alpha + \rho y,$$

assuming that α is piecewise constant with respect to the underlying partitioning of the data allows us to write:

$$\alpha = \gamma \alpha - \rho A e + \rho y$$

since any piecewise constant vector is an eigenvector of $D^{-1}\Omega$ with eigenvalue 1. Writing the last equation in terms of block matrices:

$$\left[\frac{\alpha_{\mathrm{U}}}{\alpha_{\mathrm{L}}} \right] = \gamma \left[\frac{\alpha_{\mathrm{U}}}{\alpha_{\mathrm{L}}} \right] - \rho \left[\frac{0}{e_{\mathrm{L}}} \right] + \rho \left[\frac{0}{y_{\mathrm{L}}} \right],$$

which leads to:

$$\begin{cases} \alpha_{\mathrm{U}} = \gamma \alpha_{\mathrm{U}} \\ \alpha_{\mathrm{L}} = \gamma \alpha_{\mathrm{L}} - \rho (e_{\mathrm{L}} - y_{\mathrm{L}}), \end{cases} \tag{12}$$

where α_{U} is the block of α corresponding to the unlabeled data points: $\alpha_{\mathrm{U}} = [\alpha_1, \ldots, \alpha_N]^T$. Likewise, $\alpha_{\mathrm{L}} = [\alpha_{N+1}, \ldots, \alpha_M]^T$ corresponds to the labeled information of α, $e_{\mathrm{L}} = [e_{N+1}, \ldots, e_M]^T$ and $y_{\mathrm{L}} = [y_{N+1}, \ldots, y_M]^T$. This result means that α is a piecewise constant vector and a solution to the simplified dual linear system, if $\gamma = 1$ and $e_{\mathrm{L}} = y_{\mathrm{L}}$. For this reason, we fix the value of γ to 1 and do not tune this parameter at the model selection stage. We also restrict the range of ρ to $0 < \rho \leq 1$ to keep in line with equation (12) discouraging a change of sign in α_{L}. This analysis also holds for the original dual problem with bias (5), since the only effect of the bias term is to induce a special centering matrix causing that the solution vector α has zero mean as can be seen in the second KKT optimality condition (7). Thus, to summarize, γ is fixed to 1 and the user-defined ρ ranges between 0 and 1. An algorithm for the proposed method is presented in Algorithm 1.

Algorithm 1 A semi-supervised formulation to kernel spectral clustering

```
Input: Training dataset $\mathcal{X}$, labels $\mathcal{Y}$, RBF kernel function $K(x_i, x_j)$, clustering parameters $\sigma^2, \rho$, validation set $\mathcal{D}^{\mathrm{val}} = \{x_t^{\mathrm{val}}\}_{t=1}^{N_{\mathrm{v}}}$.
Output: Partition $\{\mathcal{A}_-, \mathcal{A}_+\}$
1: Solve the dual linear system (8) to obtain $\alpha$ and compute the bias term $b$ by (9).
2: Compute the projections for training data $e$ using (10).
3: Binarize $e$ to find the cluster membership for training data points.
4: $\forall i$, assign $x_i$ to $\mathcal{A}_-, \mathcal{A}_+$ depending on the sign of $e_i$.
5: Compute the projections for validation data $\hat{e}_{\mathrm{val}}$ using (13).
6: $\forall t$, assign $x_t^{\mathrm{val}}$ to $\mathcal{A}_-, \mathcal{A}_+$ depending on the sign of $\hat{e}_{\mathrm{val},t}$
```

D. Out-of-Sample Localized Solution

Since the solution to the dual linear system (8) has piecewise constant properties when there is an underlying cluster structure in the data, it becomes important for generalization, model selection and visualization purposes to extend these properties to out-of-sample data points. Estimation of out-of-sample eigenvectors in kernel spectral clustering has been discussed in [10] and we adopt a similar strategy here. Consider a set of N_{v} validation data points $\mathcal{D}^{\text{val}} = \{x_t^{\text{val}}\}_{t=1}^{N_{\text{v}}}$. The latent variables of the clustering model are:

$$\hat{e}_{\text{val}} = \Phi_{\text{val}} w + b_l 1_{N_{\text{v}}} = \Omega_{\text{val}}\alpha + b 1_{N_{\text{v}}}, \qquad (13)$$

where $\Omega_{\text{val}} = \Phi_{\text{val}}\Phi^T$ is the $N_{\text{v}} \times M$ validation kernel matrix with tj-th entry $\Omega_{\text{val},tj} = K(x_t^{\text{val}}, x_j), t = 1, \ldots, N_{\text{v}}, j = 1, \ldots, M$, $\Phi_{\text{val}} = [\varphi(x_1^{\text{val}})^T; \ldots; \varphi(x_{N_{\text{v}}}^{\text{val}})^T]$ and $1_{N_{\text{v}}}$ is a vector of N_{v} ones. The second KKT condition in (7):

$$\alpha = (\gamma D^{-1} - \rho A)e + \rho y,$$

links the solution vector α and the projections for training data e. The main idea is to extend this link to out-of-sample projections, such that we obtain an out-of-sample solution with localized (piecewise-constant) properties. The estimated out-of-sample solution $\hat{\alpha}_{\text{val}} \in \mathbb{R}^{N_{\text{val}}}$ becomes:

$$\hat{\alpha}_{\text{val}} = (\gamma D_{\text{val}}^{-1} - \rho A_{\text{val}})\hat{e}_{\text{val}} + \rho y_{\text{val}}, \qquad (14)$$

where $D_{\text{val}}^{-1} = \text{diag}([1/\text{deg}(x_1^{\text{val}}), \ldots, 1/\text{deg}(x_{N_{\text{v}}}^{\text{val}})]) \in \mathbb{R}^{N_{\text{v}} \times N_{\text{v}}}$ is the inverse degree matrix for validation data and $\text{deg}(x) = \sum_{j=1}^{M} K(x, x_j)$ extends the concept of degree to out-of-sample data. Since $\mathcal{D}^{\text{val}}$ is unlabeled[2], the matrix A_{val} and the validation labels y_{val} equal zero leading to

$$\hat{\alpha}_{\text{val}} = \gamma D_{\text{val}}^{-1}\big(\Omega_{\text{val}}\alpha + b 1_{N_{\text{v}}}\big). \qquad (15)$$

If the validation dataset is sampled i.i.d. from the same distribution as the training data points, then the estimated out-of-sample solution $\hat{\alpha}_{\text{val}}$ will display localized cluster structures.

E. Model Selection

Having a localized representation for out-of-sample solutions allows us to perform model selection in a learning setting. The clustering parameters can be estimated by optimizing a criterion on validation data. The proposed criterion combines the Fisher criterion [11], [10] adapted to binary clustering for the unlabeled part and the classification accuracy on the labeled part of the dataset. The Fisher criterion measures how localized the clusters appear in the out-of-sample solution and is defined for binary clustering as follows. Given $\hat{\alpha}_{\text{val}}$ and the clusters $\mathcal{A}_-, \mathcal{A}_+$ with membership given by $\text{sign}(\hat{\alpha}_{\text{val}})$, the Fisher criterion $F(\sigma^2, \rho)$ can be defined as:

$$F(\sigma^2, \rho) = \frac{s_{\text{B}}}{s_{\text{B}} + s_{\text{W}}}, \qquad (16)$$

where $s_{\text{B}} = \zeta_-(\hat{\mu}_- - \hat{\mu})^2 + \zeta_+(\hat{\mu}_+ - \hat{\mu})^2$, $\hat{\mu}_-$ and $\hat{\mu}_+$ are the cluster mean values for the negative and positive cluster respectively, $\hat{\mu}$ is the mean of $\hat{\alpha}_{\text{val}}$, $s_{\text{W}} = \varsigma_- + \varsigma_+$, ς_- and ς_+ are the cluster variance values for the negative and positive cluster respectively. Note that, s_{B} is the weighted sum of the squared distances from each cluster mean to the global mean with weights[3] given by ζ_- and $\zeta_+, \zeta_- + \zeta_+ = 1$. ($s_W$) is a measure of cluster compactness with value 0 when the clusters in the out-of-sample solution are perfectly localized (piecewise constant). Thus, the idea behind the Fisher criterion is to look for clustering parameters σ^2, ρ, γ such that s_B is maximized and s_W is minimized. The Fisher criterion is bounded between 0 and 1, taking its maximal value in the clusters appear well-separated and localized in the out-of-sample solution. The proposed criterion called semi-supervised Fisher (SSF) then becomes a weighted sum of the Fisher criterion and the classification accuracy for the labeled data points used during training / validation:

$$\begin{aligned}\text{SSF}(\sigma^2, \rho) =& \eta_{\text{U}} F(\sigma^2, \rho) \\ &+ (1 - \eta_{\text{U}})\text{accuracy}(\text{sign}(e_{\text{L}}), y_{\text{L}}), \qquad (17)\end{aligned}$$

where $0 \leq \eta_{\text{U}} \leq 1$, is a user-defined weight controlling the importance given to the unlabeled measure, e_{L} are the projections of the labeled data and y_{L} are the corresponding labels.

IV. Experimental Results

Simulation results are presented in this Section. All experiments reported are performed in MATLAB 7.13 on a 2.2 GHz quad-core Intel i7, 4 GB, Mac OS X. For the real-life and benchmark datasets, we used the Laplacian SVM MATLAB code[4] provided by the authors of [16] to solve LapSVM in the primal. The original dual LapSVM problem is solved via a MEX interface to libSVM [18]. In all cases, we subsample the full dataset to obtaining training and validation sets and perform model selection using the validation set and a grid search over the parameters. For the proposed approach, we used the semi-supervised Fisher criterion with weight $\eta_{\text{U}} = 0.25$ (thus, giving more importance to classification accuracy) to obtain the clustering parameters σ^2 and ρ. We fixed γ to 1. For the Laplacian SVMs, we tuned the kernel parameter and γ_A with respect to the accuracy on the validation set. The remaining parameters are set to their default values ($\gamma_I = 1$, $NN = 6$).

A. Toy Problems

The first toy problem consists of three well-separated Gaussian clouds in $2D$. The full dataset consist of $2,000$ data points. A training set of $N = 600$ data points is subsampled randomly. The validation set is composed of $N_{\text{v}} = 800$ points. We artificially introduced binary labels for two data points. The proposed model was trained with tuned parameters $\sigma^2 = 2.5, \rho = 0.9$ using the semi-supervised Fisher (SSF) criterion on validation data. The model was tested on a uniform grid of data points to determine the decision boundaries. The negative class is depicted in blue and the positive class is depicted in green. The two labeled data points are depicted as a blue square and a green circle for negative and positive class respectively. Figure 2 shows the simulation results for completely unsupervised kernel spectral clustering (KSC) and the semi-supervised approach (Semi-KSC). The eigenvector solution of KSC and the dual

[2] If there are labels available for some validation points, equation (14) is applied instead.

[3] The weights can be set to $\zeta_- = |\mathcal{A}_-|/N_{\text{v}}$ and $\zeta_+ = |\mathcal{A}_+|/N_{\text{v}}$ to give preference to balanced clusters.

[4] Available at http://www.dii.unisi.it/ melacci/lapsvmp/

linear system solution of Semi-KSC are also depicted. The simulation results show good generalization performance with only one labeled data point of each class. The out-of-sample solution of Semi-KSC also displays a multi-cluster structure which is not visible in the KSC eigenvector. A piecewise constant solution is clearly visible in both methods but Semi-KSC correctly detect the three clusters in data, in spite of the fact that it is formulated to perform binary clustering.

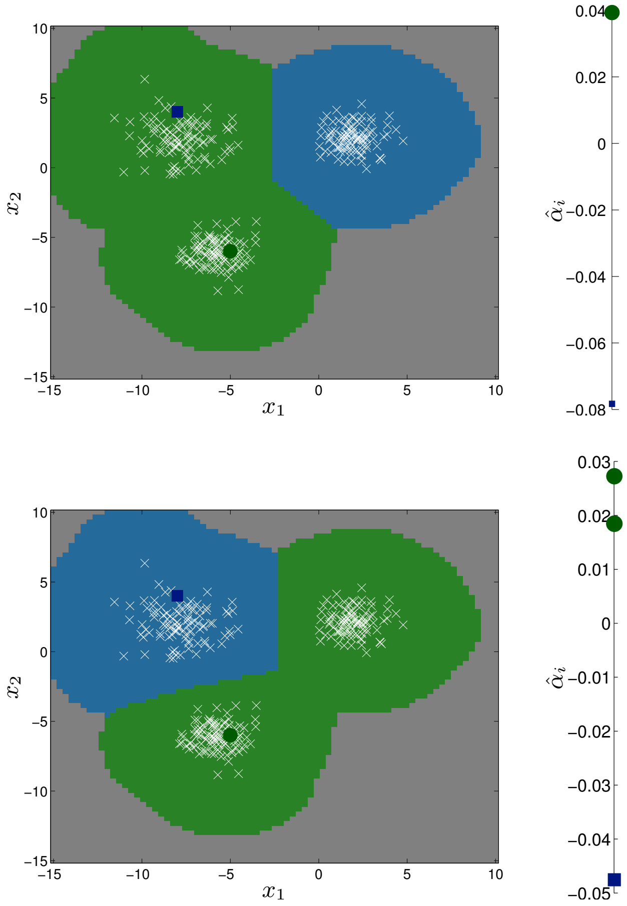

Fig. 2. Toy problem 1 - Three Gaussians. The training scenario consists of $N = 600$ data points for training and $N_{\text{v}} = 800$ for validation. The blue square (negative class) and the green circle (positive class) represent the two data points with labels. **Top:** Unsupervised results using KSC. The Gaussian could on the right is discriminated with respect to the other two. The out-of-sample eigenvector α display a strong piecewise constant structure of two clusters. **Bottom:** Results with the semi-supervised proposed approach. The labeled information propagates through the data points altering its cluster/class membership in order to fit the imposed labels. The out-of-sample localized solution displays a strong piecewise constant structure together with correctly detecting the three Gaussian clouds.

The second toy problem consists of four Gaussian clouds in $2D$ with some overlapping regions. The idea is to test the proposed approach with respect to cluster/class overlap. The training scenario is the same as in the first toy problem. The model parameters were tuned using the SSF criterion on validation data. In this case we used three labeled datapoints. Figure 3 shows the results displaying good generalization capabilities as in the previous toy example. The incorporation of few labeled information propagates in the Gaussian clouds fulfilling the desired requirements. The out-of-sample solution of Semi-KSC also displays a multi-cluster structure not present on KSC. The four clouds are correctly detected by Semi-KSC.

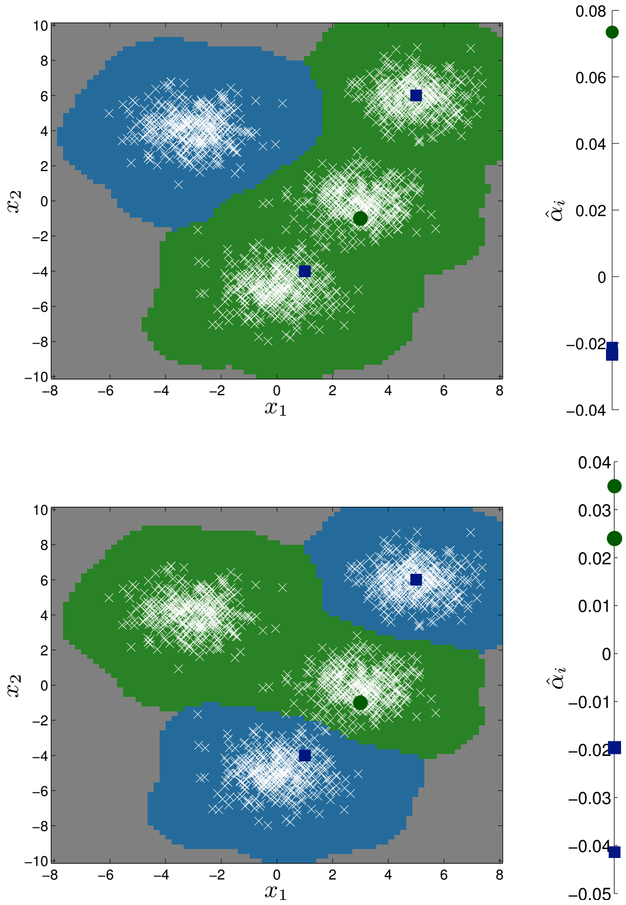

Fig. 3. Toy problem 2 - Four Gaussians with some overlap. The training scenario consists of $N = 600$ data points for training and $N_{\text{v}} = 800$ for validation. The blue square (negative class) and the green circle (positive class) represent the two data points with labels. **Top:** Unsupervised results using KSC. The out-of-sample eigenvector α display a strong piecewise constant structure of two clusters. **Bottom:** Results with the Semi-KSC. The cluster/class membership is altered accordingly in order to fulfill the few imposed labels. The out-of-sample localized solution displays a strong piecewise constant structure and correctly detects the four clouds.

B. Real and Benchmark Problems

We used the benchmark datasets `g241c`, `g241d`, `Digit1`, `USPS`, `BCI` and `Text` for semi-supervised learning described in [12]. All datasets have a total of $1,500$ data points and the dimensionality is 241 (except for `Text` which has $11,960$ sparse discrete dimensions). The first three datasets were artificially generated while `USPS`, `BCI` and `Text` were derived from real data. All datasets have binary labels. We compare the proposed approach with respect to the Laplacian support vector machine (LapSVM) introduced in [15] and its more recent extension to provide a solution in the

TABLE I

Average test error ×100% of 4 benchmark datasets described in [12]. The number of unlabeled training points is set to $N = 600$ except for BCI where $N = 150$ was used. N_{L} indicates the number of labeled examples. The methods are then evaluated on the full datasets to calculate the test error. The proposed method is compared against Laplacian SVM solved in the dual [15] and solved in the primal [16]. Semi-KSC compares favorably with respect to the Laplacian SVM with results significantly better on the Text dataset.

	g241c	g241d	BCI	Text	
LapSVM	0.48 ± 0.02	$\mathbf{0.42 \pm 0.03}$	0.48 ± 0.03	0.37 ± 0.04	$N_{\mathrm{L}} = 10$
LapSVMp	0.49 ± 0.01	0.43 ± 0.03	0.48 ± 0.02	0.40 ± 0.05	
Semi-KSC	$\mathbf{0.42 \pm 0.03}$	0.43 ± 0.04	$\mathbf{0.46 \pm 0.03}$	$\mathbf{0.29 \pm 0.06}$	
LapSVM	0.40 ± 0.06	0.31 ± 0.03	0.37 ± 0.04	0.27 ± 0.02	$N_{\mathrm{L}} = 100$
LapSVMp	0.36 ± 0.07	0.31 ± 0.02	0.32 ± 0.02	0.32 ± 0.02	
Semi-KSC	$\mathbf{0.29 \pm 0.05}$	$\mathbf{0.28 \pm 0.05}$	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.22 \pm 0.02}$	

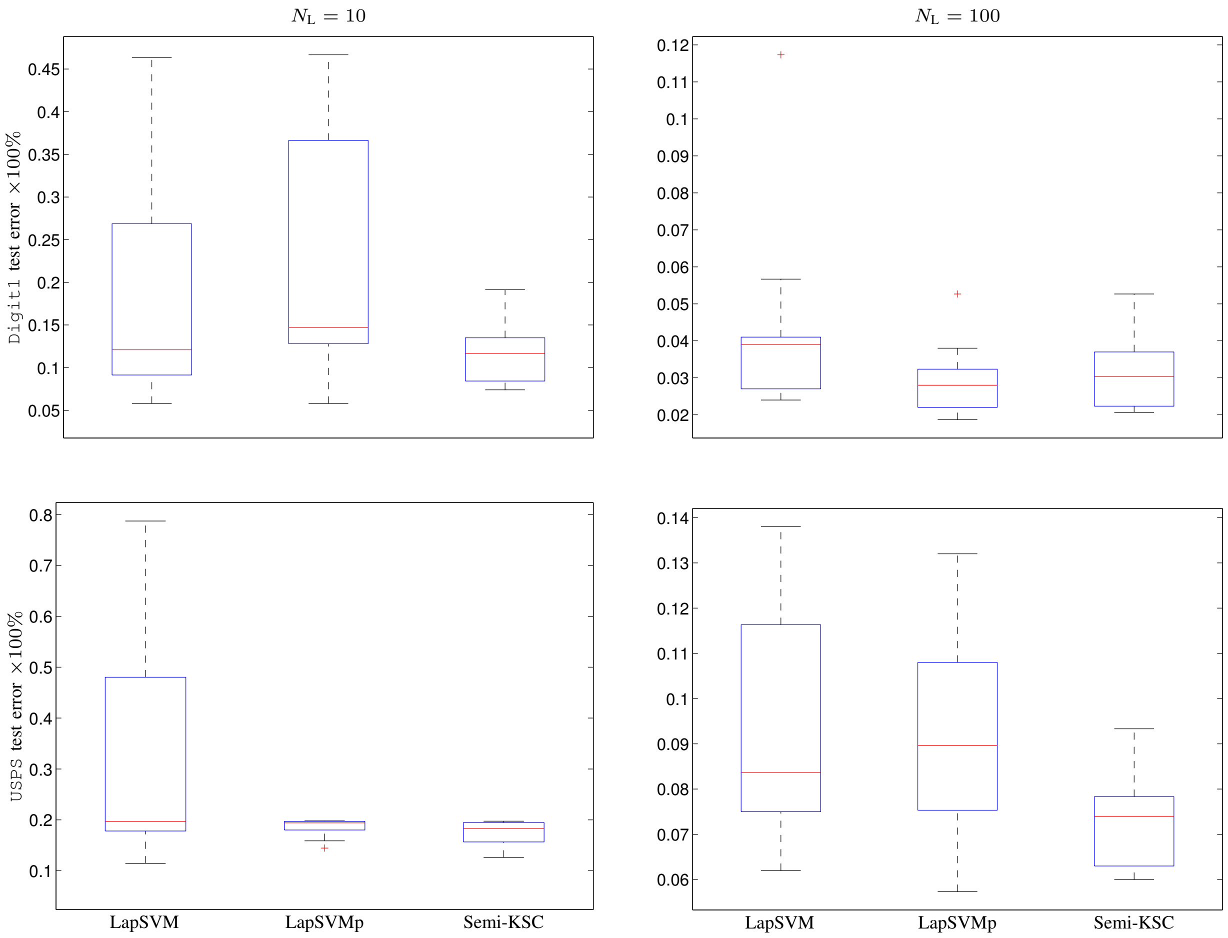

Fig. 4. Test error for benchmark datasets Digit1 (top row) and USPS (bottom row) of [12]. The number of unlabeled training points is $N = 600$. The proposed method performs consistently good with reduced variability with respect to the number of labeled examples N_{L}.

primal [16]. The datasets contain already 12 randomizations of the labeled data points and all reported results indicate the variability with respect to these randomizations. For each randomization of the labels, a training subset of $N = 600$ unlabeled data points is selected for all datasets (except for Text for which $N = 150$). A validation set of $N_{\mathrm{v}} = 600$ unlabeled examples is also drawn at random. The parameters of the Laplacian SVMs are determined by maximizing the classification accuracy on 10 randomizations of the validation set. The parameters of the proposed approach are calculated in a similar fashion but maximizing the semi-supervised Fisher criterion instead. The test results are shown in Table I and in Figure 4. The proposed approach has a consistently good performance, outperforming in most cases the Laplacian SVMs. Figure 5 shows the training computation times with respect to an increasing number of training points. Although all methods perform quite fast, Semi-KSC showed a considerably reduced computation times. These results are expected since the Laplacian SVMs lead to solving more complex optimization problem than Semi-KSC which is a linear system.

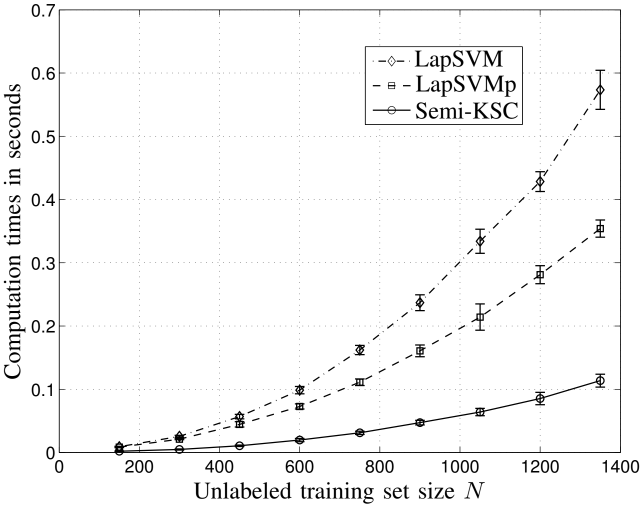

Fig. 5. Training time in seconds for the `Text` dataset with an increasing number of unlabeled training points and fix $N_{\mathrm{L}} = 100$. All methods perform quite fast but Semi-KSC takes less time than the Laplacian SVMs.

V. Conclusions and Future Work

A semi-supervised formulation to binary kernel spectral clustering called Semi-KSC is presented. The primal formulation takes the binary core model for kernel spectral clustering and incorporates the available labeled information by an additional term in the objective function and a regularization constant. This extension transform the original dual eigenvalue problem of kernel spectral clustering into a linear system. The proposed formulation allows out-of-sample extensions giving the possibility to obtain cluster/class membership for unseen data. A model selection criterion suitable for semi-supervised problems is also proposed. The proposed criterion extends the Fisher criterion used for measuring cluster by adding an extra term measuring classification accuracy on the available labeled information. The applicability of the proposed framework is shown on illustrative toy examples and on real and benchmark datasets popular in semi-supervised learning applications. Semi-KSC is also compared with respect to the Laplacian SVM in its primal and dual versions. The proposed approach outperforms Laplacian SVM in most cases in term of classification accuracy together with reduced training times. As future work, we will further investigate the intriguing results shown in the toy examples where the out-of-sample dual solution vector display a multi-cluster structure, despite the fact than the formulation is for binary clustering. Another direction worth investigating is a formulation for multi-class classification and multi-group clustering and the role of encoding/decoding schemes.

Acknowledgement

This work was supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet , CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants;Flemish Government:FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); IBBT; EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other:Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. Carlos Alzate is a postdoctoral fellow of the Research Foundation - Flanders (FWO). Johan Suykens is a professor at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

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

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

[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] 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.

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

[6] 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.

[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*, vol. 32, no. 2, pp. 335–347, February 2010.

[8] 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.

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

[10] C. Alzate and J. A. K. Suykens, “Out-of-sample eigenvectors in kernel spectral clustering,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN’11)*, 2011, pp. 2349–2356.

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

[12] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[13] X. Zhu and A. Goldberg, *Introduction to Semi-Supervised Learning*. Morgan and Claypool, 2009.

[14] J. A. K. Suykens, “Data visualization and dimensionality reduction using kernel maps with a reference point,” *IEEE Transactions on Neural Networks*, vol. 19, no. 9, pp. 1501–1517, 2008.

[15] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[16] S. Melacci and M. Belkin, “Laplacian support vector machines trained in the primal,” *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[17] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[18] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.