Sparse Kernel Models for Spectral Clustering Using the Incomplete Cholesky Decomposition

Carlos Alzate, Johan A. K. Suykens

***Abstract*— A new sparse kernel model for spectral clustering is presented. This method is based on the incomplete Cholesky decomposition and can be used to efficiently solve large-scale spectral clustering problems. The formulation arises from a weighted kernel principal component analysis (PCA) interpretation of spectral clustering. The interpretation is within a constrained optimization framework with primal and dual model representations allowing the clustering model to be extended to out-of-sample points. The incomplete Cholesky decomposition is used to compute low-rank approximations of a modified affinity matrix derived from the data which contains cluster information. A reduced set method is also presented to compute efficiently the cluster indicators for out-of-sample data. Simulation results with large-scale toy datasets and images show improved performance in terms of computational complexity.**

I. Introduction

Spectral clustering methods [1], [2], [3], [4] correspond to relaxations of graph partitioning problems that are generally NP-hard. This family of techniques are known to perform successfully in difficult cases where classical clustering methods fail. An undirected graph is formed, where the vertices correspond to the data points and the edges have an associated weight measuring the affinity between each pair of points. The relaxed solutions can be obtained from the eigenvectors of the affinity matrix which contains pairwise similarities of the data. The simplest case consists of partitioning the graph into two disjoint subsets and the relaxed solution corresponds to a particular eigenvector. However, converting relaxed solutions to clusters when more than two groups are required is not straightforward and several approaches have been proposed [1], [5], [6], [7].

Spectral clustering is typically performed only on training data without clear extensions to new points. Approximation techniques based on the Nyström method have been devised to overcome this problem [8]. A different view on spectral clustering based on weighted kernel PCA was discussed in [9], [10]. 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) [11], [12]. 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. For large scale problems, the cost of storing and computing an eigendecomposition of the affinity matrix can be prohibitive. Another issue is that the cluster indicators are obtained through dense kernel expansions in which every training point contributes.

The incomplete Cholesky decomposition [13], [14] produces a low-rank approximation of a symmetric positive semidefinite matrix and has been used to reduce the computational burden of interior point methods in support vector machines (SVMs) [15] and kernel-based contrast functions for independent component analysis (ICA) [16], [17], [18].

In this paper, we propose a sparse kernel model for spectral clustering using the incomplete Cholesky decomposition. This sparse model is an extension of [10] and aims at approximating the eigenvectors of the modified affinity matrix by solving a smaller eigenvalue problem. The cluster indicators for out-of-sample points are also approximated via a reduced set method. This reduction allows the kernel model to handle large-scale data.

This paper is organized as follows: Section II contains a review of existing spectral clustering techniques. In Section III, we describe the multiway spectral clustering algorithm discussed in [10]. In Section IV, we propose a method to compute approximated eigenvectors and a method to efficiently compute cluster indicators for out-of-sample points using the incomplete Cholesky decomposition. Section V contains the empirical results and in Section VI we give conclusions.

II. Multiway Spectral Clustering

A. Preliminaries

In graph theory, a set of training data points $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and a similarity (also called affinity) measure $a_{ij} \in \mathbb{R}^+$ between each pair of points x_i and x_j define an undirected graph if the similarity measure is symmetric $a_{ij} = a_{ji}$. The vertices of the graph are represented by the data points and the weight of the edges is determined by a_{ij}. A classical problem is then to split the graph into two disjoint parts by cutting edges with low weight. This problem is called graph bipartioning and it has been studied in [19], [2]. Several cut criteria have been proposed to obtain meaningful partitions of the graph. Typical criteria include the normalized cut *NCut* [1] and the ratio cut *RCut* [20].

The degree of the i-th vertex is the sum of the edge weights connected to it:

$$\deg_i = \sum_{j=1}^{N} a_{ij},$$

the degree matrix D can then be defined as the diagonal matrix $D = \text{diag}(\deg_1, \ldots, \deg_N)$. The problem of minimizing

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).

the normalized cut criterion can be written as [1]:

$$\begin{array}{ll} \min\limits_{q} & \dfrac{q^T L q}{q^T D q} \\ \text{such that} & q \in \{-c_1, 1\}^N \end{array} \quad (1)$$

where $L = D - A$ is the graph Laplacian, A is the similarity matrix with ij-entry $A_{ij} = a_{ij}$, q is a clustering indicator vector and c_1 is a constant that depends of the number of data points assigned to each partition.

Minimizing the normalized cut is NP-hard. However, approximate solutions can be found by relaxing the discrete constraint on q. Letting q take real values, the relaxed solution is the eigenvector corresponding the second smallest eigenvalue (called the Fiedler vector) of the following problem:

$$L\tilde{q} = \nu D\tilde{q}$$

where $\tilde{q} \in \mathbb{R}^N$ and $\tilde{q}^T\tilde{q} = 1$. The Fiedler vector $\tilde{q}^\star$ can be binarized to obtain the clustering indicators:

$$q^\star = \text{sign}(\tilde{q}^\star - \theta)$$

where θ is a threshold to be determined.

B. Markov Random Walks

A random walk on a graph consists of random jumps from vertex to vertex. Many properties of spectral clustering methods can be expressed in terms of a transition matrix P obtained by normalizing the affinity matrix such that its rows sum to 1. The ij-th entry of P represents the probability of moving from node i to node j in one step. This transition matrix can be defined as $P = D^{-1}A$. The corresponding eigenvalue problem becomes

$$Pr = \xi r.$$

The eigenvectors of P are identical to the eigenvectors of the *NCut* algorithm and the eigenvalues are related by $\xi_i = 1 - \nu_i, i = 1, \ldots, N$ [9]. Therefore, the Fiedler vector in this case correspond to the eigenvector with second largest eigenvalue. Maximizing 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 the other cluster.

C. The k-way NCut *Relaxation*

A more general problem is the so called k-way partioning problem. Instead of partitioning the graph into two sets, k-way cuts aim to split the graph into k disjoint sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_k = \emptyset$ and $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_k = \mathcal{D}$.

Consider $f_l \in \{0,1\}^N$ as the cluster indicator vector for the l-th cluster such that f_l has a 1 in the entries corresponding to the data points in the l-th cluster. The cluster indicator matrix becomes $F = [f_1, \ldots, f_k] \in \{0,1\}^{N\times k}$. Hence, the k-way *NCut* can be expressed as [21]:

$$\text{NCut}(\mathcal{A}_1, \ldots, \mathcal{A}_k) \quad = \quad k - \left(\frac{f_1^T W f_1}{f_1^T D f_1} + \ldots + \frac{f_k^T W f_k}{f_k^T D f_k} \right).$$

Defining

$$g_l = \frac{D^{1/2} f_l}{||D^{1/2} f_l||_2}, l = 1, \ldots, k$$

and $G = [g_1, \ldots, g_k]$ leads to

$$\text{NCut}(\mathcal{A}_1, \ldots, \mathcal{A}_k) = k - \text{tr}(G^T \hat{L} G)$$

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

$$\begin{array}{rcl} \underset{G}{\text{argmin}}\ \text{NCut}(\mathcal{A}_1, \ldots, \mathcal{A}_k) & = & \underset{G}{\text{argmax}}\ \text{Tr}(G^T \hat{L} G) \\ \text{such that} & & G^T G = I_k. \end{array}$$

This problem is NP-hard due to the fact that G only takes discrete values. Defining $\tilde{G} \in \mathbb{R}^{N\times k}$ as the relaxed clustering indicator matrix leads to the k-way *NCut* relaxation [21]:

$$\begin{array}{rcl} \underset{\tilde{G}}{\text{argmin}}\ \text{NCut}(\mathcal{A}_1, \ldots, \mathcal{A}_k) & = & \underset{\tilde{G}}{\text{argmax}}\ \text{Tr}(\tilde{G}^T \hat{L} \tilde{G}) \\ \text{such that} & & \tilde{G}^T \tilde{G} = I_k. \end{array}$$

This maximization problem is a special form of Fan's theorem [22], [23], [24]. The optimal solution of the relaxed problem is given by:

$$\tilde{G}^\star = B R_2$$

where $B = [b^{(1)}, \ldots, b^{(k)}] \in \mathbb{R}^{N\times k}$ is any orthonormal basis of the k-th principal subspace of $\hat{L}$ and $R_2 \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix [22], [5]. In other words, the relaxed solution is contained in the eigenvectors corresponding to the k largest eigenvalues of the normalized Laplacian $\hat{L} = D^{-1/2} A D^{-1/2}$ but the eigenvectors can have arbitrary rotations.

To obtain cluster indicators from the eigenvectors of the normalized Laplacian, several methods have been proposed. Typical approaches include recursive binary cuts [1], reclustering [1] and rounding [25]. The most used approach is reclustering which consists of applying k-means on the eigenvectors. This technique works well only if the clusters in the new space represented by the eigenvectors are spherical and well-separated.

III. Spectral Clustering through Weighted Kernel PCA

A multiway spectral clustering method with out-of-sample extensions was introduced in [10]. This spectral method is based on a weighted kernel PCA technique in which the clustering model can be extended to out-of-sample points. The formulation can be interpreted as an unsupervised version of multiclass LS-SVM models [11] and as a extension of [9]. Additional score variables are introduced as constraints and each score variables vector provides a binary clustering.

A. Multiway Formulation

Given training data $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of clusters k, the following primal problem can be formulated [10]:

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

$$\text{such that} \quad \begin{cases} e^{(1)} = \Phi w^{(1)} + b_1 1_N \\ e^{(2)} = \Phi w^{(2)} + b_2 1_N \\ \vdots \\ e^{(k-1)} = \Phi w^{(k-1)} + b_{k-1} 1_N \end{cases}$$

where $e^{(l)} = [e_1^{(l)}, e_2^{(l)}, \ldots, e_N^{(l)}]$ is the compact form of the projected variables $e_i^{(l)} = w^{(l)^T}\varphi(x_i) + b_l$, $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the mapping to high-dimensional feature space (which can be infinite dimensional), γ_l are regularization parameters, b_l are bias terms and Φ is the $N \times n_h$ feature matrix:

$$\Phi = \begin{bmatrix} \varphi(x_1)^T \\ \varphi(x_2)^T \\ \vdots \\ \varphi(x_N)^T \end{bmatrix},$$

$i = 1, \ldots, N, l = 1, \ldots, k-1$. Each projected variables vector $e^{(l)}$ (also called score variables) provides a binary clustering with cluster indicator $q^{(l)}$:

$$q_i^{(l)} = \text{sign}(e_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1. \quad (3)$$

The Lagrangian of this constrained optimization problem is:

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

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

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \to w^{(l)} = \Phi^T \alpha^{(l)} \\ \frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \to \alpha^{(l)} = (\gamma_l/N)D^{-1}e^{(l)} \\ \frac{\partial \mathcal{L}}{\partial b_l} = 0 \to 1_N^T \alpha^{(l)} = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \to e^{(l)} = \Phi w^{(l)} + b_l 1_N, \end{cases} \quad (4)$$

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

$$b_l = -\frac{1}{1_N^T D^{-1} 1_N} 1_N^T D^{-1} \Omega \alpha^{(l)}, l = 1, \ldots, k-1.$$

Eliminating the primal variables $w^{(l)}, e^{(l)}, b_l$ leads to

$$\frac{\gamma_l}{N} M\Phi\Phi^T\alpha^{(l)} = \alpha^{(l)}$$

where

$$M = D^{-1} - \frac{1}{1_N^T D^{-1} 1_N} D^{-1} 1_N 1_N^T D^{-1}. \quad (5)$$

Applying the kernel trick $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$, and defining $\lambda_l = N/\gamma_l$ leads to the following eigenvalue problem:

$$M\Omega\alpha^{(l)} = \lambda_l \alpha^{(l)}, j = 1, \ldots, k-1 \quad (6)$$

where Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j)$. The primal problem (2) is in general a non-convex problem therefore the KKT conditions are necessary but not sufficient. In this case, the eigenvectors of $M\Omega$ are stationary points of the Lagrangian. Note that the matrix $M\Omega$ is not symmetric but it is the product of two symmetric positive semidefinite matrices therefore its eigenvalues are real [13]. The projections of the training points onto the eigenvectors become:

$$e_i^{(l)} = w^{(l)^T}\varphi(x_i) + b_l = \sum_{j=1}^{N}\alpha_j^{(l)}K(x_i, x_j) + b_l, \quad (7)$$

and the clustering indicators are given by (3).

In classical spectral clustering, the grouping is obtained only for training data with no clear extension to new points except for approximation schemes based on the Nyström method [8]. In the proposed approach, the clustering can be extended to new points due to the clear primal/dual model in an optimization framework. Given a set of N_{test} test points $\{x_t^{\text{test}}\}_{t=1}^{N_{\text{test}}}$, the score variables for these points become:

$$z_t^{(l)} = w^{(l)^T}\varphi(x_t^{\text{test}}) + b_l = \sum_{j=1}^{N}\alpha_j^{(l)}K(x_t^{\text{test}}, x_j) + b_l, \quad (8)$$

and the cluster indicators are given by

$$\breve{q}_t^{(l)} = \text{sign}(z_t^{(l)}), \quad (9)$$

$l = 1, \ldots, k-1, t = 1, \ldots, N_{\text{test}}$.

B. Encoding / Decoding

Once the eigenvectors $\alpha^{(l)}$ have been computed, a codebook $\mathcal{C} = \{c_p\}_{p=1}^k$ can then be obtained from the rows of the binarized eigenvector matrix $\hat{A} = [\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \ldots, \hat{\alpha}^{(k-1)}]$ where $\hat{\alpha}^{(l)} = \text{sign}(\alpha^{(l)}), l = 1, \ldots, k-1$. The codebook is formed by the k encodings (rows of $\hat{A}$) with most occurrences. For test data the cluster indicator matrix is $\breve{Q} \in \{-1, 1\}^{N_{\text{test}} \times (k-1)}$ with tl-entry:

$$\breve{Q}_{tl} = \breve{q}_t^{(l)}, l = 1, \ldots, k-1, t = 1, \ldots, N_{\text{test}}. \quad (10)$$

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 [26].

IV. Reduced Set Spectral Clustering via the Incomplete Cholesky Decomposition

For large-scale problems, the cost of storing the kernel matrix Ω and computing the eigenvalue decomposition of $M\Omega$ can be prohibitive. Moreover, the clustering indicators can be obtained through the score variables which are dense expansions of kernel evaluations. In this paper, we propose a method to compute the eigenvectors of $M\Omega$ by solving a

reduced eigenvalue problem and a method to approximate clustering indicators via sparse expansions using a reduced set method. Both methods are based on the incomplete Cholesky decomposition.

A. The Cholesky Decomposition

A symmetric positive definite matrix $A \in \mathbb{R}^{N\times N}$ can be decomposed as $A = LL^T$ where $L \in \mathbb{R}^{N\times N}$ is a lower triangular matrix [13], [14]. This factorization is called the Cholesky decomposition and it is widely used for solving linear systems. If the matrix A is positive semidefinite, it is still possible to compute an *incomplete* Cholesky decomposition where some columns of L are zero. Symmetric permutations of A (also called symmetric pivoting) are needed to ensure numerical stability of the decomposition [14]. Therefore, given an error threshold η, the incomplete Cholesky decomposition can be written as $A \approx GG^T, G \in \mathbb{R}^{N\times R}$ with $||A - GG^T||_2^2 \leq \eta$ and $R \leq N$. The η parameter controls the rank of the approximation. For large values of η, the approximation gets worse and R is small. If the eigenvalues of A decay rapidly then incomplete Cholesky leads to small numerical error and $R \ll N$ [15]. This factorization has been used to decrease the storage requirement and computational complexity of interior point point methods in SVMs [27] and to approximate contrast functions for ICA in an efficient way [18], [16], [17], [28].

B. Approximated Eigenvectors

Consider the incomplete Cholesky decomposition of Ω:

$$\Omega \approx GG^T, \tag{11}$$

where $G \in \mathbb{R}^{N\times R}$, the singular value decomposition of G gives:

$$G = U\Lambda V^T \tag{12}$$

where $U \in \mathbb{R}^{N\times R}$ is the matrix of left singular vectors, $\Lambda \in \mathbb{R}^{R\times R}$ is the diagonal matrix of singular values and $V \in \mathbb{R}^{N\times R}$ is the matrix of right singular vectors. Substituting (12) into (11) gives $\Omega \approx U\Lambda^2 U^T$. Hence, (6) can be written as

$$MU\Lambda^2 U^T\alpha^{(l)} = \lambda_l\alpha^{(l)}, l = 1,\ldots,k-1,$$

premultiplying by U^T gives:

$$\begin{aligned} U^TMU\Lambda^2U^T\alpha^{(l)} &= \lambda_l U^T\alpha^{(l)} \\ U^TMU\Lambda^2\rho^{(l)} &= \lambda\rho^{(l)} \end{aligned} \tag{13}$$

where $\rho^{(l)} = U^T\alpha^{(l)}$ and $l = 1,\ldots,k-1$. Note that (13) involves the eigendecomposition of a $R \times R$ matrix which can be much smaller than the $N \times N$ matrix in (6). This approximation scheme can also be applied to kernel PCA and to other spectral clustering techniques such as the normalized cut, the random walks model and the NJW algorithm [3].

C. Reduced Set Method

Note that the primal projection vectors $w^{(l)}$ are expressed in terms of dense kernel expansions:

$$w^{(l)} = \sum_{i=1}^{N}\alpha_i^{(l)}\varphi(x_i) = \Phi^T\alpha^{(l)}.$$

The objective of a reduced set method is to approximate $w^{(l)}$ by

$$\tilde{w}^{(l)} = \sum_{m=1}^{N_s}\beta_m^{(l)}\varphi(\tilde{x}_m) = \Psi^T\beta^{(l)}$$

where $l = 1,\ldots,k-1$, $\Psi = [\varphi(\tilde{x}_1^T);\ldots;\varphi(\tilde{x}_{N_s}^T)]$, $\Psi \in \mathbb{R}^{N_s\times n_h}$, $\beta = [\beta_1,\ldots,\beta_{N_s}]$ is the vector of reduced set coefficients, $\mathcal{R} = \{\tilde{x}_m\}_{m=1}^{N_s}$ is the reduced set and $N_s < N$. This scheme was used in [29] to approximate support vector decision rules for classification and in [30] to sparsify kernel PCA expansions. The reduced set problem consists of two subproblems: first, finding the reduced set coefficients β and second finding the reduced set $\mathcal{R}$. One approach to solve the first subproblem is to minimize the squared two-norm of the residuals with respect to $\beta^{(l)}$:

$$\min_{\beta^{(l)}}||w^{(l)} - \tilde{w}^{(l)}||_2^2 = \min_{\beta} {w^{(l)}}^T w^{(l)} - 2{\tilde{w}^{(l)}}^T w + {\tilde{w}^{(l)}}^T\tilde{w}^{(l)}.$$

Using $w^{(l)} = \Phi\alpha^{(l)}$ and $\tilde{w}^{(l)} = \Psi^T\beta^{(l)}$ gives:

$$\min_{\beta^{(l)}} {\alpha^{(l)}}^T\Phi\Phi^T\alpha^{(l)} - 2\beta^{(l)}\Psi\Phi^T\alpha^{(l)} + {\beta^{(l)}}^T\Psi\Psi^T\beta^{(l)}$$

and applying the kernel trick leads to:

$$\min_{\beta^{(l)}} {\alpha^{(l)}}^T\Omega\alpha^{(l)} - 2\beta^{(l)}\Omega^{\Psi\Phi} + {\beta^{(l)}}^T\Omega^{\Psi\Psi},$$

where $l = 1,\ldots,k-1$, $\Omega^{\Psi\Phi}$ is a $N_s \times N$ matrix with mi-th entry $\Omega_{mi}^{\Psi\Phi} = K(\tilde{x}_m, x_i), m = 1,\ldots,N_s, i = 1,\ldots,N$ and $\Omega^{\Psi\Psi}$ is a $N_s \times N_s$ matrix with mn-th entry $\Omega_{mn}^{\Psi\Psi} = K(\tilde{x}_m, \tilde{x}_n), n = 1,\ldots,N_s$.

The first-order optimality condition

$$\frac{\partial||w^{(l)} - \tilde{w}^{(l)}||_2^2}{\partial\beta^{(l)}} = 0$$

gives:

$$\Omega_{\Psi\Psi}\beta^{(l)} = \Omega_{\Psi\Phi}\alpha^{(l)} \tag{14}$$

which is a linear system[1] to be solved in $\beta^{(l)}, l = 1,\ldots,k-1$. The second subproblem is automatically solved using the pivots found in the incomplete Cholesky decomposition of Ω. The pivots represent the subset $\mathcal{R} = \{\tilde{x}_m\}_{m=1}^{N_s} \subset \mathcal{D}$ for which the mapped points $\varphi(\tilde{x}_m)$ are linearly independent. In this case, the matrix $\Omega_{\Psi\Psi}$ is full rank and $N_s = R$.

The approximated score variables for training data be-

[1] A different approach for sparse expansions was also mentioned in [30]. This approach uses L_1 shrinkage penalizers and the reduced set coefficients can be obtained by solving a QP problem.

come:

$$e_i^{(l)} \approx \sum_{m=1}^{R} \beta_m^{(l)} K(x_i, \tilde{x}_m), i = 1, \ldots, N \tag{15}$$

and for test data:

$$z_t^{(l)} \approx \sum_{m=1}^{R} \beta_m^{(l)} K(x_t^{\text{test}}, \tilde{x}_m), t = 1, \ldots, N_{\text{test}}. \tag{16}$$

D. Algorithm

An implementation of the proposed sparse approach is shown in Algorithm 1. Note that the symmetric matrix $M \in \mathbb{R}^{N \times N}$ does not need to be stored. The ij-th entry of M is:

$$M_{ij} = \begin{cases} -c_2/(\text{deg}_i \text{deg}_j) & \text{if } j \neq i \\ (\text{deg}_i - c_2)/\text{deg}_i^2 & \text{if } j = i \end{cases}$$

where $c_2 = \sum_{f=1}^{N} (1/\text{deg}_f)$ and $i, j = 1, \ldots, N$. Therefore in step 3, the rows of M are computed on demand to form the $R \times R$ matrix $U^T M U \Lambda^2$.

Algorithm 1 Sparse Kernel Model for Spectral Clustering Using the Incomplete Cholesky Decomposition

Input: Training set $\{x_i\}_{i=1}^N$, positive definite kernel function $K(x_i, x_j)$, number of clusters k, Cholesky error tolerance η, test set $\{x_t^{\text{test}}\}_{t=1}^{N_{\text{test}}}$.
Output: Partition $\Delta = \{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, approximated eigenvectors α, cluster codeset $\mathcal{C} = \{c_p\}_{p=1}^k, c_p \in \{-1, 1\}^{k-1}$, reduced set $\mathcal{R}$.

1: Compute the incomplete Cholesky factor $G \in \mathcal{R}^{N \times R}$ of the kernel matrix Ω such that $||\Omega - GG^T||_2^2 \leq \eta$ and obtain the reduced set $\mathcal{R} = \{\tilde{x}_m\}_{m=1}^R$ from the pivots.
2: Compute the left singular vectors U and the singular values Λ of G.
3: Compute the eigenvectors ρ corresponding to the $k-1$ largest eigenvalues of $U^T M U \Lambda^2$.
4: Obtain the approximated eigenvectors via:

$$\alpha^{(l)} = U \rho^{(l)}, l = 1, \ldots, k-1.$$

5: Binarize the eigenvectors:

$$\hat{\alpha}_i^{(l)} = \text{sign}(\alpha_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1.$$

6: Let $\hat{\alpha}_i \in \mathbb{R}^{k-1}$ be the encoding vector for the training data point x_i, $i = 1, \ldots, N$.
7: Count the occurrences of the different encodings and find the k encodings with most occurrences. Let the codeset be formed by these k encodings: $\mathcal{C} = \{c_p\}_{p=1}^k, c_p \in \{-1, 1\}^{k-1}$
8: For all i, assign x_i to $\mathcal{A}_{p^\star}$ where $p^\star = \text{argmin}_p d_H(\hat{\alpha}_i, c_p)$ and $d_H(\cdot, \cdot)$ is the Hamming distance.
9: Compute the approximated score variables $z_t^{(l)}$ for test data using (16).
10: Compute $\breve{q}_t^{(l)}$ using (9) and let $\breve{q}_t \in \mathbb{R}^{k-1}$ be the encoding vector for the test data point $x_t^{\text{test}}, t = 1, \ldots, N^{\text{test}}, l = 1, \ldots, k-1$.
11: For all t, assign x_t^{test} to $\mathcal{A}_{p^\star}$ where

$$p^\star = \text{argmin}_p d_H(\breve{q}_t, c_p)$$

V. Empirical Results

In this section we report empirical results. The assessment of the clustering for the toy examples is done using the adjusted Rand index [31]. 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. All experiments were performed on an Intel Core Duo 2.0 GHz, 2 GB RAM in Matlab 7.5. We used the incomplete Cholesky routine described in [16] which computes efficiently the Cholesky factor.

A. Three Gaussian clouds in 3D

This experiment consists of three Gaussian clouds in a 3D space. The number of training point is equal to $6,000$. The RBF kernel parameter σ^2 is fixed to 15. The Cholesky error tolerance η is fixed within the range $10^{-5} \leq \eta \leq 10^{-1}$. The adjusted Rand index between the approximated cluster indicators and the true labels is equal to 1 for the whole range. For $\eta > 0.5$, the adjusted Rand index starts to decrease (i.e. some points are assigned to the wrong Gaussian cloud). The computation times with respect to η are shown in Figure 1. The dashed line corresponds to the *full* approach which is computing the eigendecomposition of (6) directly without approximations. The solid line shows the time spent to solve the proposed approach. Note that for $\eta > 10^{-3}$ the reduction in time is more than one order of magnitude. Figure 2 shows the training points in gray and the reduced set points in black. The obtained reduced set size is $R = 78$ and the points are distributed on the outskirts of the clouds.

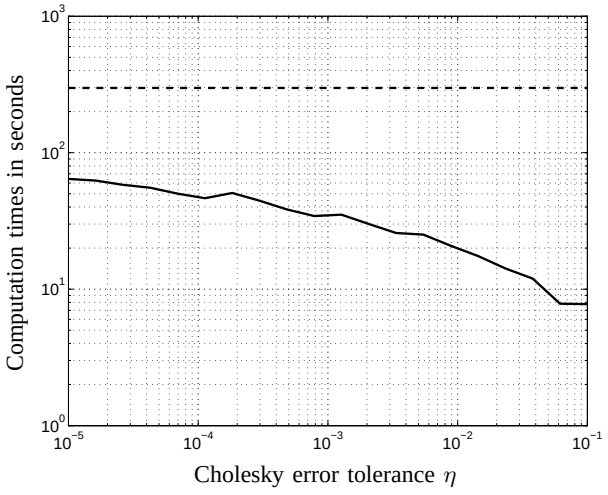

Fig. 1. Three Gaussian clouds experiment. Computation times in seconds of the sparse approach (solid line) with respect to the Cholesky error tolerance η. The dashed line corresponds to the full approach. The reduction in time is more than one order of magnitude for $\eta > 10^{-3}$.

B. Large-scale intertwined spirals problem

This experiment consists of two intertwined spirals in 2D. The RBF kernel parameter was fixed to $\sigma^2 = 0.5$ and the

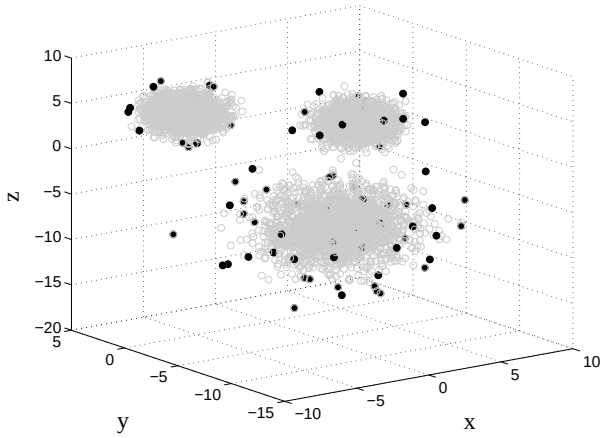

Fig. 2. Three Gaussian clouds experiment. The reduced set ($R = 78$) is depicted in black and the training points ($N = 6,000$) in black. The reduced set points are mainly distributed on the borders of the clouds.

Cholesky error tolerance to $\eta = 0.5$. The computation times of the proposed approach with respect to the number of data points are shown in Figure 3 and reduced set sizes in Table I. The proposed sparse approach takes about 8 hours to compute the approximated eigenvectors of the dataset when $N = 100,000$ and the obtained reduced set size is 261 points (0.21%). Figure 4 shows the clustering results for the given σ^2 and $N = 40,000$. The two spiral clusters are correctly detected. We did not perform comparisons with the full approach because in that case the matrix $M\Omega$ does not fit into memory.

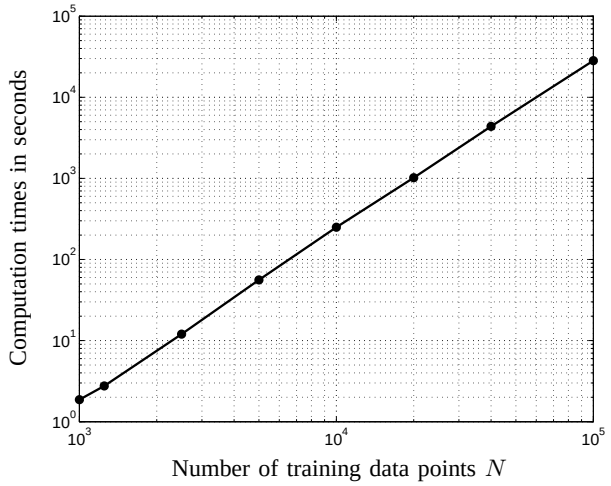

Fig. 3. Large-scale intertwined spirals. Computation times of the proposed approach with respect to N with η fixed to 0.5. Note that for the given range of N, the full approach does not fit into memory.

C. Image Segmentation

For this experiment we use a 400×280 pixels image from the Berkeley image dataset [32]. The RBF kernel parameter was tuned using the model selection method described in [10]. This criterion uses validation data to select parameters

Training set size N	Reduced set size R
$1,000$	168 (83.2%)
$5,000$	210 (95.8%)
$10,000$	223 (97.8%)
$50,000$	242 (99.5%)
$100,000$	261 (99.7%)

TABLE I

LARGE-SCALE INTERTWINED SPIRALS. REDUCED SET SIZE R WITH RESPECT TO N. THE PERCENTAGES INDICATE THE DEGREE OF SPARSENESS. THE CHOLESKY ERROR TOLERANCE η IS FIXED TO 0.5.

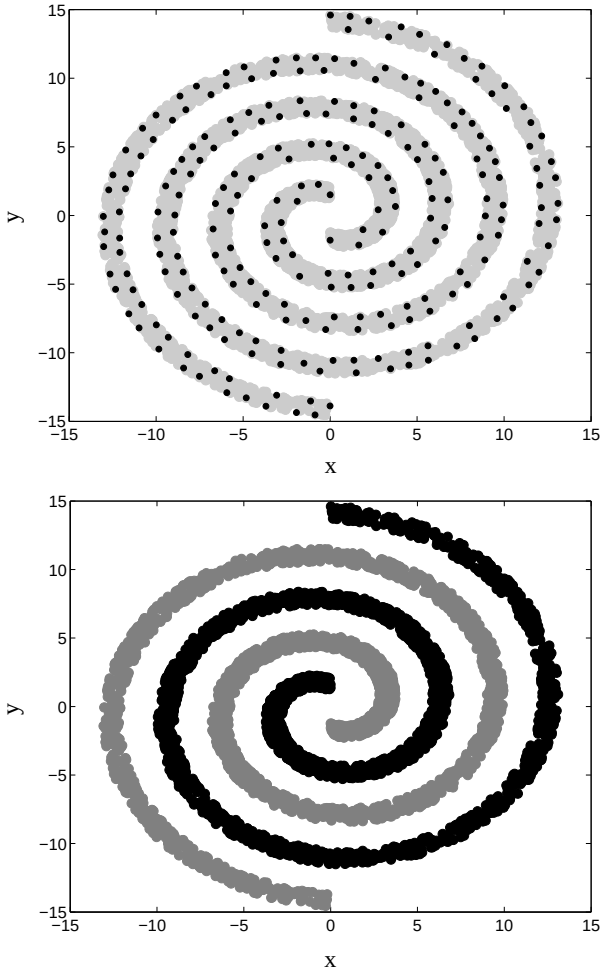

Fig. 4. Large-scaled intertwined spirals with $N = 40,000$. *Top:* Training set shown in gray. The reduced set points (shown in black) are distributed on the borders of the spirals. *Bottom:* Clustering results. Cluster 1 is shown in gray, cluster 2 in black. The proposed sparse approach correctly indicate the two spiral clusters.

in such a way that the projected variables are as collinear as possible for points in the same cluster. This geometrical property is linked with the piecewise constant property of the eigenvectors of some spectral clustering methods in the case of compact and well-separated clusters. The tuned parameter is $\sigma^2 = 0.15$. The image was subsampled to create training, validation and test set and each pixel is represented as a three-

dimensional vector with the levels of red, blue and green using 8 bits. The training set consists of a 200×140 pixels image therefore $N = 28,000$. The validation set is used to find the kernel parameters and the cluster indicators of the test sets are inferred using the out-of-sample extension. The Cholesky error tolerance η is fixed to 10^{-2} with an obtained reduced set of 124 pixels (99.6% sparseness). The total computation time was 403 seconds. Figure 5 shows the segmentation results. The clusters found are visually appealing.

VI. Conclusions

A new sparse kernel model for computing efficiently and approximating the solutions of spectral clustering problems is proposed. This method is based on the incomplete Cholesky decomposition and also allows the efficient computation of cluster indicators for out-of-sample points via a reduced set technique. The formulation is cast in an optimization framework where the eigenvectors of a modified affinity matrix are the dual solutions of a primal problem. The framework allows extending the clustering model to out-of-sample points. The proposed approximations can handle large-scale data with a reduced computational complexity and with similar performance compared to the original method without approximations. With the out-of-sample extension the clustering model can be trained, validated and tested in a learning frawework. Simulations with large-scale toy datasets and images 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.

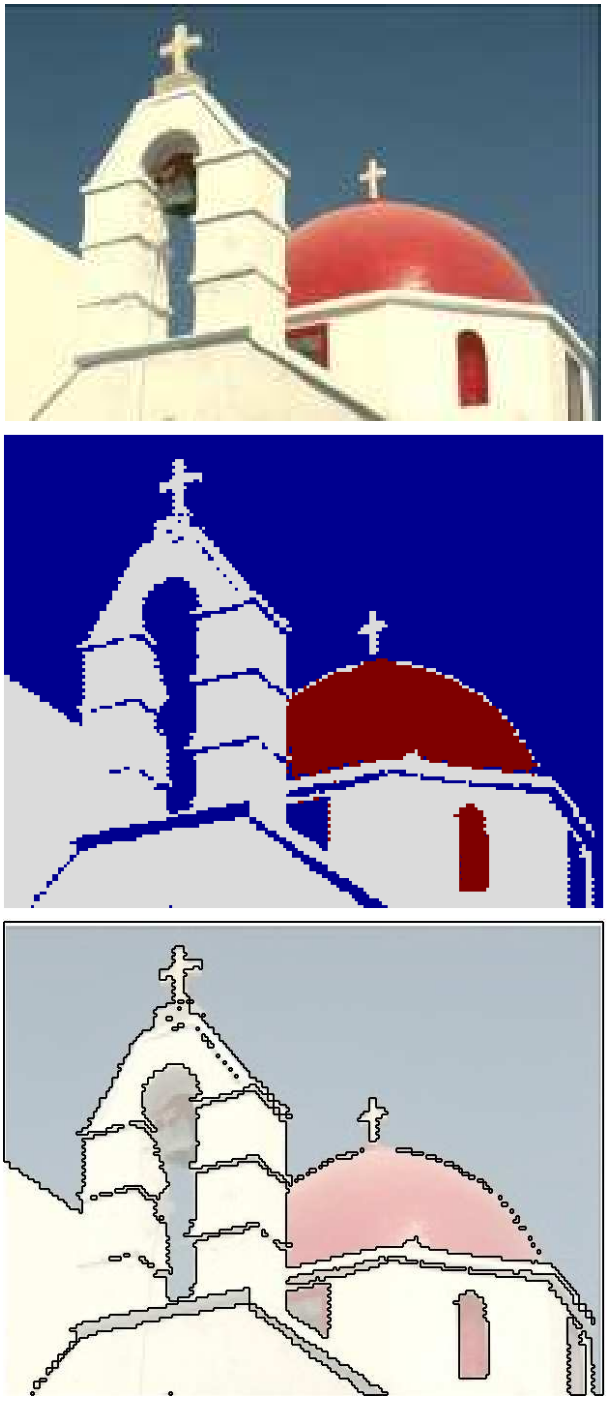

Fig. 5. Image segmentation. Visually appealing clusters found by the proposed sparse method. *Top:* Original 400×280 image. Training, validation and test sets are obtained from this image using subsampling. *Center:* Obtained clusters shown in different colors. *Bottom:* Sobel edge detector on the obtained clusters. The original image is shown faded.

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] 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.

[6] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 1601–1608.

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

2004, p. 30.
[8] 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.
[9] C. Alzate and J. A. K. Suykens, "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.
[10] ——, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," ESAT-SISTA, K.U.Leuven, Internal Report 07-229, 2007.
[11] 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.
[12] 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, March 2003.
[13] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.
[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.
[15] S. J. Wright, "Modified Cholesky factorizations in interior-point algorithms for linear programming," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1159–1191, 1999.
[16] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.
[17] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, 2005.
[18] C. Alzate and J. A. K. Suykens, "A regularized kernel CCA contrast function for ICA," ESAT-SISTA, K.U.Leuven, Internal Report 07-149. Submitted for publication, 2007.
[19] 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.
[20] L. Hagen and A. Kahng, "New spectral methods for ratio cut partitioning algorithms," *IEEE Trans. Computer-Aided Design*, vol. 11, no. 9, pp. 1074–1085, 1992.
[21] 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.
[22] 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.
[23] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Trans. Computer-Aided Design*, vol. 13, no. 9, pp. 1088–1096, 1994.
[24] C. Ding and X. He, "K-means clustering via principal component analysis," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM Press, 2004, p. 29.
[25] F. R. Bach and M. I. Jordan, "Learning spectral clustering," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.
[26] 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.
[27] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, December 2001.
[28] C. Alzate and J. A. K. Suykens, "ICA through an LS-SVM based kernel CCA measure for independence," in *Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN'07)*, 2007, pp. 2920–2925.
[29] C. J. C. Burges, "Simplified support vector decision rules," in *Proceedings of the 13th Intl. Conf. on Machine Learning*, 1996, pp. 71–77.
[30] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, M. Scholz, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, September 1999.
[31] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, pp. 193–218, 1985.
[32] 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.