



HAL
open science

Constrained spectral embedding for K-way data clustering

Guillaume Wacquet, Emilie Poisson Caillault, Denis Hamad, Pierre-Alexandre Hébert

► **To cite this version:**

Guillaume Wacquet, Emilie Poisson Caillault, Denis Hamad, Pierre-Alexandre Hébert. Constrained spectral embedding for K-way data clustering. *Pattern Recognition Letters*, 2013, 34 (9), pp.1009-1017. 10.1016/j.patrec.2013.02.003 . hal-01536663

HAL Id: hal-01536663

<https://hal.science/hal-01536663>

Submitted on 15 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constrained Spectral Embedding for K-Way Data Clustering

G. Wacquet*, É. Poisson Caillault, D. Hamad, P.-A. Hébert

*LISIC - Lab. of Computing, Signal and Image Processing in Côte d'Opale
Université Lille Nord de France, ULCO
62228 Calais, France*

Tel: +33 (0)3 21 46 36 91, Fax: +33 (0)3 21 46 57 51

Abstract

Spectral clustering methods meet more and more success in machine learning community thanks to their ability to cluster data points of any complex shapes. The problem of clustering is addressed in terms of finding an embedding space in which the projected data are linearly separable by a classical clustering algorithm such as K-means algorithm. Often, spectral algorithm performances are significantly improved by incorporating prior knowledge in their design, and several techniques have been developed for this purpose. In this paper, we describe and compare some recent linear and non-linear projection algorithms integrating instance-level constraints ("must-link" and "cannot-link") and applied for data clustering. We outline a K-way spectral clustering algorithm able to integrate pairwise relationships between the data samples. We formulate the objective function as a combination of the original spectral clustering criterion and the penalization term based on the instance constraints. The optimization problem is solved as a standard eigensystem of a signed Laplacian matrix. The relevance of the proposed

*Corresponding author.

Email address: name@lisic.univ-littoral.fr (G. Wacquet)

algorithm is highlighted using six UCI benchmarks and two public face databases.

Keywords: Graph embedding, Spectral clustering, Pairwise constraints, Signed Laplacian.

1. Introduction

In many real-world applications, we are dealing with the problem of clustering of high dimensional databases for which we have little prior knowledge. Clustering aims to group data sharing similar properties to their respective categories. It was shown that, the introduction of domain knowledge in the clustering algorithms, may greatly improve their performances. Domain knowledge is generally provided in two forms: class labels (Chapelle et al., 2006) or instance constraints (Basu et al., 2008). Labelling data is a hard and long task for human experts while pairwise relationship between data is easier since it consists in simply indicating if two instances are similar (*must-link*) or dissimilar (*cannot-link*) (Wagstaff and Cardie, 2002).

Recently, spectral methods, based on graph concepts, have been developed for dimension reduction and data clustering (Saul et al., 2006; Shortreed and Meila, 2005; Von Luxburg, 2007). They meet more and more success in machine learning community thanks to their theoretical foundations and their practical applications. The problem of data clustering is considered in terms of finding an embedding space in which the projected data are linearly separable by a classical K-means algorithm. The data are represented in a graph where each vertex is associated with a data sample and the weighted edges encode the relationship between the underlying data. Usually, the embedding space is obtained by Laplacian Eigenmaps. This is carried out by selecting the eigenvectors of the graph Laplacian. Each eigen-

22 vector corresponds to non-linear projection of the data set. The performances
23 of spectral algorithms depend on the way they integrate the data constraints in
24 their design: (a) integration of constraints in the affinity matrix, (b) integration of
25 constraints in the optimization criterion (De Bie et al., 2004; Basu et al., 2008;
26 Wang and Davidson, 2010; Wang et al., 2012). Note that the graph built using
27 instance constraints may contain negatively weighted edges associated to cannot-
28 link constraints. In this situation, the obtained graph is called signed graph and
29 its associated Laplacian matrix is called signed Laplacian matrix (Kunegis et al.,
30 2010).

31 In this paper, we present and compare recent methods for data projection and
32 clustering, using pairwise relationships, in terms of spectral theory. Among the
33 spectral methods developed in the literature, some include the clustering step in
34 their algorithms (spectral clustering) and others are used for dimension reduction
35 (spectral embedding) (Saul et al., 2006). The latter can be easily used for data
36 clustering by applying a classical K-means algorithm on the projected data. We
37 briefly review the classical principal component analysis (PCA) and the locality
38 preserving projection LPP (He and Niyogi, 2002) as well as their constrained
39 variant. We develop a constrained spectral embedding algorithm for K-way data
40 clustering. The embedding obtained by our approach is closer to the constrained
41 Laplacian Eigenmaps (Chen et al., 2010). The algorithm optimizes an objective
42 function which is a combination of standard spectral clustering criterion and the
43 penalization term based on the instance constraints. The optimization problem
44 is solved as a standard eigensystem of a signed Laplacian matrix. We show the

45 relevance of the algorithm on many UCI¹ benchmark datasets and two well-known
46 face databases².

47 The paper is organized as follows. Section 2 presents some basic graph nota-
48 tions used for spectral methods. Section 3 describes two constrained linear projec-
49 tion methods: PCA and LPP methods. Spectral clustering approaches integrating
50 implicitly and explicitly the pairwise constraints are presented in Section 4. Sec-
51 tion 5 describes the proposed constrained K-way spectral clustering algorithm.
52 Section 6 presents some performance study of the proposed algorithm using six
53 UCI datasets and two public face databases. Finally, Section 7 shows some dis-
54 cussions and concluding remarks.

55 2. Basic notations of spectral methods

56 In this section, we present some basic notations used in the graph formalism
57 (Von Luxburg, 2007).

- 58 • $\mathcal{X} = \{x_1, \dots, x_i, \dots, x_N\}$ is a dataset of N instances, $x_i \in \mathfrak{R}^P$, $i = 1, \dots, N$;
- 59 • $G(V, E, W)$ is a weighted graph associated with \mathcal{X} : $V = \{v_1, \dots, v_i, \dots, v_N\}$
60 is the set of vertices corresponding to the N instances; E is the set of edges
61 and W is the weight matrix indicating the affinity or closeness of pairwise
62 instances x_i, x_j where $w_{ij} \geq 0$ and $w_{ij} = w_{ji}$;
- 63 • D is the degree matrix of graph G . D is a diagonal matrix where $d_{ii} =$
64 $\sum_{j=1}^N w_{ij}$ is the degree of the vertice v_i .

¹<http://archive.ics.uci.edu/ml/>

²<http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html/>

65 Spectral methods consider the K-way clustering problem as a K-way graph-
 66 cut into K non-empty groups denoted by $\{V_1, \dots, V_k, \dots, V_K\}$ such as $\cup_{k=1}^K V_k = V$
 67 and $V_k \cap V_l = \emptyset, k \neq l$:

- 68 • Volume of a set V_k : $vol(V_k) = \sum_{v_i \in V_k} d_{ii}$;
- 69 • Similarity within cluster V_k : $Cut(V_k, V_k) = \sum_{v_i \in V_k} \sum_{v_j \in V_k} w_{ij}$;
- 70 • Similarity between V_k and its complement \bar{V}_k : $Cut(V_k, \bar{V}_k) = \sum_{v_i \in V_k} \sum_{v_j \in \bar{V}_k} w_{ij}$.

71 Let $u_k = (u_{1k}, \dots, u_{ik}, \dots, u_{Nk})^T$ be the indicator vector of V_k :

$$u_{ik} = \begin{cases} 1 & \text{if } v_i \in V_k, \\ 0 & \text{if } v_i \notin V_k. \end{cases} \quad (1)$$

72 Using u_k , the above graph characteristics can be defined more consistently:

$$vol(V_k) = u_k^T D u_k, \quad (2)$$

$$Cut(V_k, V_k) = u_k^T W u_k, \quad (3)$$

$$Cut(V_k, \bar{V}_k) = u_k^T (D - W) u_k = u_k^T L u_k. \quad (4)$$

73 where:

$$L = D - W, \quad (5)$$

74 is called graph Laplacian matrix. L is symmetric and positive semi-definite.

75

76 Usually, in addition to unlabeled dataset, we have some kind of knowledge
 77 known as instance-level constraints: two instances are similar and their vertices
 78 "must be linked" or dissimilar and therefore the corresponding vertices "cannot
 79 be linked":

- 80 • the set of must-links is defined by: $\mathcal{M} = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ are similar}\}$.
- 81 • the set of cannot-links is defined by: $\mathcal{C} = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ are dissimilar}\}$.

82 In the context of spectral theory, must-link and cannot-link graphs are built
 83 from the pairwise constraints. Usually, the edges of must-link graph have positive
 84 weights while the edges of cannot-link graph may have negative weights. There-
 85 fore, the graph associated to the unlabelled data and pairwise constraints may
 86 contain negative weights. It is called signed graph and its associated Laplacian
 87 matrix is called signed Laplacian matrix: $\bar{L} = \bar{D} - W$ where $\bar{d}_{ii} = \sum_{j=1}^N |w_{ij}|$ is the
 88 degree of the vertex v_i .

89 In the next section, we present two existing methods for dimensionality reduc-
 90 tion based on pairwise constraints: the constrained principal component analysis
 91 and the constrained locality preserving projection.

92 **3. Constrained linear projection approaches**

93 In many domains, we often deal with high dimensional datasets. However, all
 94 dimensions are not necessary and reducing the input space to a lower space will
 95 make the clustering problem not only computationally easier, but also allow to
 96 discover the data structure. The performances of projection methods are highly
 97 conditioned by the way they integrate the instance-level constraints in their de-
 98 sign. In the sequel, we will briefly describe two existing constrained projection
 99 techniques based on the classical principal component analysis (PCA) and the
 100 more recent locality preserving projection (LPP) (He and Niyogi, 2002).

101 *3.1. Constrained principal component analysis*

102 PCA method performs dimensionality reduction by projecting the input data
 103 onto a lower dimensional space spanned by the largest eigenvectors of the data
 104 covariance matrix. The problem is to find a linear function between the input data
 105 space and the projected data space of the form:

$$y = a^T x, \text{ with } a^T a = 1, \quad (6)$$

106 which maximizes the objective criterion:

$$J_{PCA} = \frac{1}{N} \sum_{i=1}^N (y_i - m)^2, \quad (7)$$

107 where $m = \frac{1}{N} \sum_{i=1}^N y_i$. Equation (7) can also be written as:

$$\begin{aligned} J_{PCA} &= \frac{1}{N^2} \sum_{i,j} (y_i - y_j)^2, \\ J_{PCA} &= \frac{1}{N^2} \sum_{i,j} (a^T x_i - a^T x_j)^2. \end{aligned} \quad (8)$$

108 The solution a is the eigenvector associated with the largest eigenvalue of the
 109 data covariance matrix. Therefore, the PCA space is spanned by the top eigenvec-
 110 tors in which the data are best spread.

111

112 The constrained PCA (cPCA) takes into consideration the instance-level con-
 113 straints sets \mathcal{M} and \mathcal{C} . Indeed, the main idea is to look for a direction a such as
 114 the projected points $y_i = a^T x_i$ ($a^T a = 1$) satisfies PCA criterion in Equation (8) as
 115 well as the instance-level constraints (Zhang et al., 2007). The constrained PCA
 116 criterion is defined by:

$$J_{cPCA} = J_{PCA} + \underbrace{\frac{1}{|\mathcal{C}|} \sum_{(x_i, x_j) \in \mathcal{C}} (a^T x_i - a^T x_j)^2}_{\text{cannot-link constraints}} - \underbrace{\frac{1}{|\mathcal{M}|} \sum_{(x_i, x_j) \in \mathcal{M}} (a^T x_i - a^T x_j)^2}_{\text{must-link constraints}}, \quad (9)$$

117 with $|\mathcal{C}|$ and $|\mathcal{M}|$ are the cardinals of constraint sets \mathcal{C} and \mathcal{M} respectively.

118 Equation (9) can also be written as:

$$J_{cPCA} = \frac{1}{2} \sum_{i,j} (a^T x_i - a^T x_j)^2 \tilde{w}_{ij}, \quad (10)$$

119 with:

$$\tilde{w}_{ij} = \begin{cases} \frac{1}{N^2} + \frac{1}{|\mathcal{C}|} & \text{if } (x_i, x_j) \in \mathcal{C}, \\ \frac{1}{N^2} - \frac{1}{|\mathcal{M}|} & \text{if } (x_i, x_j) \in \mathcal{M}, \\ \frac{1}{N^2} & \text{else.} \end{cases} \quad (11)$$

120 The development of Equation (10) leads to:

$$\begin{aligned} J_{cPCA} &= \sum_{i,j} (a^T x_i \tilde{w}_{ij} x_j^T a - a^T x_i \tilde{w}_{ij} x_j^T a) \\ J_{cPCA} &= a^T X (\tilde{D} - \tilde{W}) X^T a. \end{aligned} \quad (12)$$

121 where $\tilde{D} \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix of \tilde{W} .

122 The constrained optimization criterion of Equation 12 is:

$$J_{cPCA} = a^T X \tilde{L} X^T a, \quad \mathbf{s.t.} \quad a^T a = 1. \quad (13)$$

123 The *cPCA* projection space is obtained from the top eigenvectors of the eigen-
124 system:

$$X \tilde{L} X^T a = \lambda a. \quad (14)$$

125 where $\tilde{L} = \tilde{D} - \tilde{W}$.

126 (Zhang et al., 2007) used a semi-supervised dimensionality reduction method
 127 (SSDR) which introduces penalty terms α and β in order to balance "cannot-link"
 128 and "must-link" contributions in the optimization criterion (Equation 9). It is easy
 129 to show that, $\alpha = \beta = 0$ lead to a classical PCA criterion. In their experiments,
 130 the authors proposed to choose $\alpha = 1$ and $\beta \geq 1$ in order to favour the "must-link"
 131 constraints. In (Davidson, 2009; Tang and Zhong, 2007), only pairwise constraints
 132 are used to guide the dimensionality reduction for clustering.

133 3.2. Constrained locality preserving projection

134 The locality preserving projection (denoted LPP) is a dimensionality reduction
 135 method recently used in the literature (He and Niyogi, 2002). LPP constructs the
 136 affinity matrix W using a Gaussian kernel:

$$w_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} & \text{if } x_i \text{ (respectively } x_j) \text{ is among the} \\ & \text{nearest neighbors of } x_j \text{ (respectively } x_i), \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

137 with σ is a scale parameter. Note that there are several ways for setting the affinity
 138 matrix W . LPP criterion is defined by:

$$J_{LPP} = \sum_{i,j} (a^T x_i - a^T x_j)^2 w_{ij}, \quad \mathbf{s.t.} \quad a^T XDX^T a = 1. \quad (16)$$

139 In the objective function J_{LPP} , the penalty contribution w_{ij} is high if neighbor-
 140 ing points x_i and x_j are projected far apart. Therefore, minimizing J_{LPP} attempts
 141 to ensure that if x_i and x_j are close then $y_i = a^T x_i$ and $y_j = a^T x_j$ are close as well.

142 A compact form of J_{LPP} using the Laplacian matrix is given by:

$$J_{LPP} = a^T XLX^T a, \quad \mathbf{s.t.} \quad a^T XDX^T a = 1. \quad (17)$$

143 The LPP projection space is obtained from the eigenvectors associated with
 144 the smallest eigenvalues of the generalized eigensystem:

$$XLX^T a = \lambda XDX^T a. \quad (18)$$

145 The constrained locality preserving projection method (denoted cLPP) inte-
 146 grates the constraints sets in the objective criterion (Cevikalp and Verbeek, 2008;
 147 Yu et al., 2010):

$$J_{cLPP} = J_{LPP} + \underbrace{\sum_{(x_i, x_j) \in \mathcal{M}} (a^T x_i - a^T x_j)^2}_{\text{must-link constraints}} - \underbrace{\sum_{(x_i, x_j) \in \mathcal{C}} (a^T x_i - a^T x_j)^2}_{\text{cannot-link constraints}}. \quad (19)$$

148 Equation (19) can also be written as:

$$J_{cLPP} = \sum_{i,j} (a^T x_i - a^T x_j)^2 \tilde{w}_{ij}, \quad (20)$$

149 with:

$$\tilde{w}_{ij} = \begin{cases} w_{ij} + 1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ w_{ij} - 1 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ w_{ij} & \text{else.} \end{cases} \quad (21)$$

150 The constrained optimization criterion can be written using the Laplacian ma-
 151 trix:

$$J_{cLPP} = a^T X \tilde{L} X^T a, \text{ s.t. } a^T X \tilde{D} X^T a = 1. \quad (22)$$

152 where $\tilde{L} = \tilde{D} - \tilde{W}$.

153 The cLPP algorithm attempts to preserve the locality of data and to satisfy the
 154 space-level constraints at the same time.

155 In the context of clustering, cPCA and cLPP are usually followed by a K-
 156 means algorithm. In (Zheng et al., 2004), the concept of locality preservation is
 157 used for data clustering.

158 **4. Constrained spectral clustering approaches**

159 In the literature, a number of algorithms have been proposed in order to in-
 160 corporate instance-level constraints into spectral clustering. They can be grouped
 161 into two categories:

- 162 • direct integration of pairwise constraints in the affinity matrix (Kamvar et
 163 al., 2003; Xu et al., 2005).
- 164 • integration of pairwise constraints in the optimization criterion (Wang and
 165 Davidson, 2010; Wang et al., 2012).

166 In the following, we briefly describe the spectral clustering approach.

167 *4.1. Spectral clustering*

168 Spectral clustering method (SC) is usually used in its normalized form (Meila
 169 and Shi, 2000; Ng et al., 2002; Shi and Malik, 2000; Shortreed and Meila, 2005;
 170 Von Luxburg, 2007). The goal is to use the graph-cut in order to partition the data
 171 into K clusters. The objective function of spectral clustering is to find a vector u
 172 which minimizes the following criterion:

$$J_{SC} = \sum_{i,j} (u_i - u_j)^2 w_{ij}, \quad (23)$$

173 which can be written as:

$$J_{SC} = u^T L u, \quad (24)$$

174 under the constraints: $u^T D u = 1$ and $D u \perp 1$ where u is the relaxed cluster indicator
 175 vector (i.e, the components of u can have real values). For simplicity of notation,
 176 we substitute u by $D^{-\frac{1}{2}} z$ and the Equation becomes:

$$J_{SC} = z^T \mathcal{L} z, \text{ s.t. } z^T z = 1. \quad (25)$$

177 where $\mathcal{L} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$ is the normalized Laplacian matrix.

178 The problem solution is given by:

$$\mathcal{L}z = \lambda z, \quad (26)$$

179 The spectral space associated to the K smallest positive eigenvalues is similar
180 to the one obtained by the Laplacian eigenmap developed in (Belkin and Niyogi,
181 2002).

182 4.2. Integration of constraints in the affinity matrix

183 In (Kamvar et al., 2003), the authors adapt the spectral clustering algorithm to
184 constrained classification problem: spectral learning algorithm (SL). They incor-
185 porate pairwise constraints into the affinity matrix:

$$\tilde{w}_{ij} = \begin{cases} 0 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ w_{ij} & \text{otherwise.} \end{cases} \quad (27)$$

186 The spectral learning algorithm proceeds just as any other standard spectral
187 clustering algorithm. However, the main weakness of this algorithm is that it
188 implicitly encodes the constraints by modifying the graph Laplacian matrix. A
189 more natural approach is to preserve the original graph Laplacian and to explicitly
190 encode the constraints.

191 4.3. Integration of constraints in the optimization criterion

192 In (Wang and Davidson, 2010), the authors combine spectral clustering and
193 pairwise constraints criteria in a flexible manner. The flexible constrained spectral
194 clustering (FCSC) preserves the original graph Laplacian matrix and explicitly

195 encodes the constraints. FCSC is solved by a generalized eigenvalue system. This
 196 approach includes a user-specified parameter α which serves as a tradeoff factor
 197 between the structure defined by the graph Laplacian and that by the constraint
 198 matrix.

199 The FCSC algorithm is detailed with $K = 2$ (Wang and Davidson, 2010). The
 200 constraints matrix is defined by:

$$q_{ij} = \begin{cases} -1 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ 0 & \text{else.} \end{cases} \quad (28)$$

201 In order to measure how well the constraints are satisfied by the cluster assign-
 202 ment, Wang and Davidson used:

$$u^T Q u = \sum_{i,j} u_i u_j q_{ij}. \quad (29)$$

203 where $u \in \{-1, +1\}^N$ is the cluster indicator vector.

204 The problem is then formulated as a constrained optimization problem, letting
 205 $z = D^{\frac{1}{2}} u$ and $Q_n = D^{-\frac{1}{2}} Q D^{-\frac{1}{2}}$:

$$\arg \min_z z^T \mathcal{L} z, \text{ s.t. } z^T Q_n z \geq \alpha, z^T z = \text{vol}(G), z \neq D^{\frac{1}{2}} \mathbf{1}, \quad (30)$$

206 where $\text{vol}(G) = \sum_{i=1}^N d_{ii}$. Recently, the authors generalized the above bipartition
 207 method to a K-way constrained clustering by selecting not only the first, but the
 208 top- K generalized eigenvectors corresponding to positive eigenvalues (Wang et
 209 al., 2012).

210 **5. Constrained spectral clustering**

211 In this section, we develop our constrained spectral clustering algorithm (de-
 212 noted cSC). Here, the objective function J_{cSC} consists in the combination of the
 213 classical spectral clustering criterion (J_{SC}) and a penalization term based on the
 214 instance constraints (J_{CM}):

$$J_{cSC} = \gamma \cdot J_{SC} + (1 - \gamma) \cdot J_{CM}, \quad (31)$$

215 The regularization coefficient γ has to be adjusted in order to balance the contri-
 216 bution of J_{SC} and J_{CM} , where:

$$J_{SC} = \sum_{i,j} (u_i - u_j)^2 w_{ij} = \mathbf{u}^T L \mathbf{u}, \quad (32)$$

217 and

$$J_{CM} = - \sum_{(x_i, x_j) \in \mathcal{C}} (u_i - u_j)^2 + \sum_{(x_i, x_j) \in \mathcal{M}} (u_i - u_j)^2. \quad (33)$$

218 J_{CM} can be written as:

$$J_{CM} = \sum_{i,j} (u_i - u_j)^2 q_{ij}, \quad (34)$$

219 with:

$$q_{ij} = \begin{cases} -1 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ 0 & \text{else.} \end{cases} \quad (35)$$

220 We can rewrite J_{CM} as:

$$J_{CM} = \mathbf{u}^T L_Q \mathbf{u}, \quad (36)$$

221 where L_Q is the Laplacian matrix of constraints graph:

$$L_Q = D_Q - Q, \quad (37)$$

222 with D_Q is the degree matrix of constraints graph.

223 Using Equations (32) and (36), Equation (31) becomes:

$$J_{cSC} = \mathbf{u}^T (\gamma.L + (1 - \gamma).L_Q) \mathbf{u} = \mathbf{u}^T L_{cSC} \mathbf{u}. \quad (38)$$

224 where:

$$L_{cSC} = D_{cSC} - W_{cSC}, \quad (39)$$

225 with:

$$D_{cSC} = \gamma.D + (1 - \gamma).D_Q, \quad (40)$$

226 and

$$W_{cSC} = \gamma.W + (1 - \gamma).Q. \quad (41)$$

227 The constrained spectral space is obtained from the eigenvectors of the Lapla-
228 cian matrix $L_{cSC} = \gamma L + (1 - \gamma)L_Q$.

229 Note that the eigenvalues of L_{cSC} may have negative sign due to the negative
230 edges weights of cannot-links constraints. In order to overcome this limitation,
231 we use the signed Laplacian matrix defined by (Kunegis et al., 2010):

$$\bar{L}_{cSC} = \bar{D}_{cSC} - W_{cSC}, \quad (42)$$

232 where \bar{D}_{cSC} is the signed degree matrix given by:

$$\bar{d}_{cSC}(i, i) = \sum_{j=1}^N |w_{cSC}(i, j)|, \quad (43)$$

233 Note that the signed Laplacian matrix is semi-definite positive. By substituting
234 \mathbf{u} by $\bar{D}_{cSC}^{-\frac{1}{2}} \mathbf{z}$ to relax the discreteness condition, Equation (38) becomes:

$$J_{cSC} = \mathbf{z}^T \bar{L}_{cSC} \mathbf{z}, \quad (44)$$

235 where $\bar{\mathcal{L}}_{cSC}$ is the normalized signed Laplacian matrix defined as $\bar{\mathcal{L}}_{cSC} = \bar{D}_{cSC}^{-\frac{1}{2}} \bar{\mathcal{L}}_{cSC} \bar{D}_{cSC}^{-\frac{1}{2}}$.

236 The constrained spectral space is obtained from the K lowest eigenvectors of $\bar{\mathcal{L}}_{cSC}$.

237 It is interesting to know that, in case $K = 2$, the retained solution is the second
238 smallest eigenvector. Indeed, the first vector (z_1) is constant and represents a
239 trivial solution for $\lambda = 0$. The final partition is then obtained by partitioning the
240 projected data thanks to the sign of values in z_2 .

241 In case $K > 2$, we maintain the usage of K eigenvectors, considering that the
242 constant vector $z_1 = (1, \dots, 1)^T$ has no impact on the obtained spectral subspace.
243 These K eigenvectors are then used in order to cluster the data thanks to the K-
244 means algorithm. Input instances are assigned to their corresponding clusters ob-
245 tained in the constrained spectral space.

246 In (De Bie et al., 2004), the authors proposed a softly constrained spectral clus-
247 tering using a regularization term similar to the one of equation (31). In this work,
248 the constraint matrix is used in order to constrain the projection of data according
249 to their labels. However, the constraints matrix used in multiclass learning do not
250 include the cannot link constraints.

251 In the context of kernel machines, Alzate and Suykens (Alzate and Suykens,
252 2009, 2010, 2012) revised the spectral clustering in terms of weighted kernel PCA
253 using the least square support vector machines developed by (Suykens and Van-
254 dewalle, 1999). Indeed, Alzate et al. present interesting weighted kernel PCA
255 approaches to deal with the framework of spectral clustering. Furthermore, in
256 (Alzate and Suykens, 2009) the authors formulate weighted kernel PCA with pair-
257 wise constraints which leads to a constrained spectral algorithm. Their contribu-
258 tion is of utmost importance since it integrates, using the kernel concept, the
259 out-of-sample extension in a natural way allowing model selection and general-

260 ization capabilities. However, the regularization term of cannot-link constraints is
261 chosen much smaller than the regularization term of must-link constraints in order
262 to avoid negative entries in the equivalent kernel matrix due to rank-1 downdates.
263 The constrained spectral clustering approach, presented in our paper, does not
264 require positive eigenvalues of the graph Laplacian and can deal with negative
265 eigenvalues, which may occur due to cannot-link constraints. Moreover, the regu-
266 larization coefficient in equation (31), weights the contribution of spectral cluster-
267 ing and must-link and cannot-link constraints. It can take any value in its interval.
268 Thus, our approach can be applied using only pairwise constraints (regularization
269 coefficient = 0) without the use of unlabelled data. Note that, for a regularization
270 term equal to 1, we obtain the classical spectral clustering. It should be noted that,
271 our approach is closer to Constrained Laplacian Eigenmap (Chen et al., 2010)
272 while that of (Alzate and Suykens, 2009) is derived from kernel PCA and there-
273 fore, the out of sample is obtained naturally. We stress the fact that the objective
274 of our presented work is the use constrained spectral clustering without seeking a
275 solution for the out-of-sample problem.

276 **6. Experiments**

277 We propose to compare our constrained spectral clustering algorithms with
278 cPCA (Zhang et al., 2007), cLPP (Cevikalp and Verbeek, 2008), SL (Kamvar et
279 al., 2003) and FCSC (Wang and Davidson, 2010) algorithms described in Sections
280 3 and 4.

281 *6.1. UCI databases*

282 In this section, we evaluate the performances of the proposed constrained spec-
283 tral clustering (cSC) algorithm and compare it with the presented algorithms on

284 six UCI databases ("Hepatitis", "Ionosphere", "Wine", "Dermatology", "Glass"
285 and "Ecoli"). These databases were chosen because they represent data of differ-
286 ent sizes and different densities. Table 1 summarizes the characteristics of each
287 database.

288 In our experiments, we evaluate the performances of the several clustering al-
289 gorithms described in the Sections 3, 4 and 5. For linear methods (cPCA and
290 cLPP) we used 95% of the total variance which made the maximum dimension
291 for each dataset as follows: Hepatitis (3), Ionosphere (24), Wine (2), Dermatol-
292 ogy (3), Glass (5) and Ecoli (5). For non linear methods, we adopted the classi-
293 cal strategy by keeping the first K eigenvectors as projection coordinates. In our
294 experiments, cPCA and cLPP are followed by K-means algorithm for data clus-
295 tering in the projection space. For cLPP, we set the number of neighbors to 10.
296 For FCSC, the parameter measuring the constraints satisfaction and denoted by α
297 is set to $0.5 \times \lambda_{\max} \text{vol}(G)$ (Wang and Davidson, 2010). Moreover, the projected
298 data are normalized to have unit-length. For our cSC, we simply set the balancing
299 parameter γ to 0.5.

300 The affinity matrix terms are of Gaussian form with a scale parameter σ^2 equal
301 to the average of the variances of database features. The generation procedure of
302 pairwise constraints is as follows: we randomly select pairs of instances from the
303 dataset and create "must-link" or "cannot-link" constraints depending on whether
304 the two instances belong or not to the same class. We iterate this scheme and
305 enrich the generated constraint sets from 0% constraints (unlabelled data) to 20%
306 constraints. The performances of algorithms are averaged over 10 repetitions of
307 the constraints generation process.

308 Each clustering algorithm generates cluster label for each data instance and the

Table 1: UCI datasets used for experiments.

	Nb. Objects	Nb. Features	Nb. Classes (K)
Hepatitis	80	19	2
Ionosphere	351	34	2
Wine	178	13	3
Dermatology	366	34	6
Glass	214	9	6
Ecoli	336	7	8

309 clustering performance of the algorithm is evaluated by comparing the generated
 310 class label and the ground-truth label. Given a data point x_i , let \hat{k}_i be the obtained
 311 cluster label and k_i the ground-truth label respectively. The agreement between the
 312 algorithm decision and the ground-truth is measured by $\delta(k_i, \text{map}(\hat{k}_i))$ (Carpaneto
 313 and Toth, 1980):

$$\delta(k_i, \text{map}(\hat{k}_i)) = \begin{cases} 1 & \text{if } k_i = \text{map}(\hat{k}_i), \\ 0 & \text{otherwise.} \end{cases} \quad (45)$$

314 In order to evaluate the performance of the presented spectral algorithms, we
 315 propose to use two criteria: the Accuracy and the Rand Index (Rand, 1971).

316 The Accuracy criterion is defined by:

$$\text{Accuracy} = \sum_{i=1}^N \frac{\delta(k_i, \text{map}(\hat{k}_i))}{N}, \quad (46)$$

317 The Rand Index is given by:

$$\text{Rand Index} = \frac{\text{number of correct decisions}}{\text{number of total decisions}}. \quad (47)$$

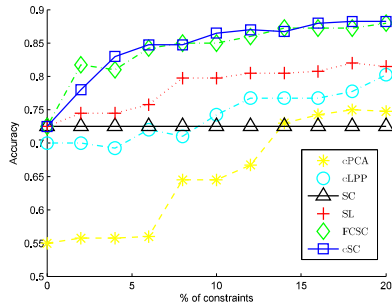
318 A decision is considered correct if the proposed clustering obtained by an algo-
319 rithm agrees with the target clustering. More specifically, a decision is considered
320 as correct if: two instances are in the same cluster and the algorithm partitioned
321 them into a same cluster, or they are in different clusters and partitioned into dif-
322 ferent clusters. Rand Index values ranges from 0 to 1.

323 Figure 1 shows the Accuracy of the spectral algorithms depending on the con-
324 straints rates applied on the six UCI datasets. In multiclass problems, FCSC is not
325 represented because it does not allow to obtain a feasible solution with the fixed
326 value of θ . We can see that cSC algorithm outperforms all algorithms followed
327 by SL algorithm. Moreover, for all databases ($K = 2$ and $K > 2$), their accuracies
328 increase as the constraints rates increase. Globally, cPCA and cLPP produce the
329 worst performances, specifically for the "Wine" database.

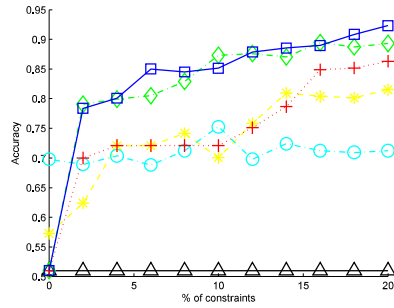
330 Figure 2 summarizes the Rand Index of the six spectral algorithms according
331 to the constraints rates applied on the six databases. It is easy to see that cSC
332 algorithm outperforms all other methods, followed by SL algorithm. Moreover,
333 their Rand Indices increase as the constraints rates increase. The Rand Index of
334 SC has a constant value since it does not depend on the constraints rates. cSC
335 obtains the better Rand Indices which can be explained by the fact that, unlike
336 SL algorithm, cSC takes into account the spatial position of data (w_{ij}) and their
337 constraints $(-1, +1)$. Finally, Rand Indices of cPCA and cLPP which are linear
338 projection algorithms are lower than the Rand Indices of constrained non-linear
339 projection algorithms.

340 6.2. *ORL and Yale Face databases*

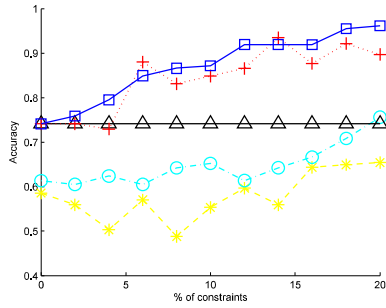
341 We compare the performances of the proposed algorithm with that of the pre-
342 sented algorithms on two well-known databases used in face recognition domain:



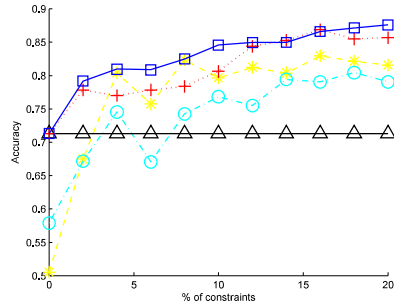
Hepatitis ($K = 2$)



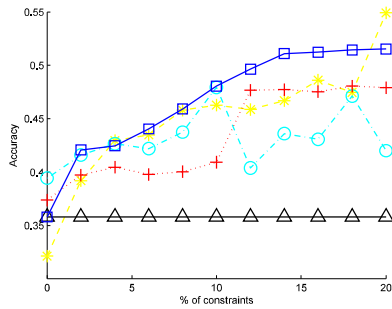
Ionosphere ($K = 2$)



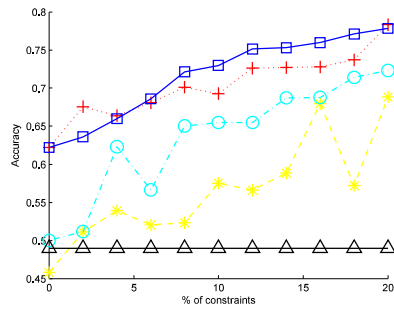
Wine ($K = 3$)



Dermatology ($K = 6$)



Glass ($K = 6$)



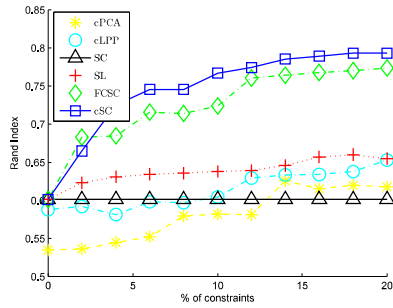
Ecoli ($K = 8$)

Figure 1: Average Accuracy according to the constraints rates, on UCI datasets.

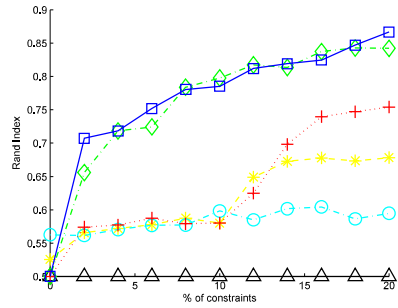
343

- "Yale Face database" contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configura-

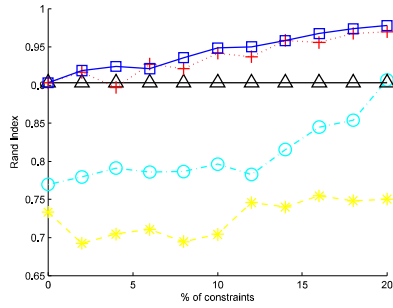
344



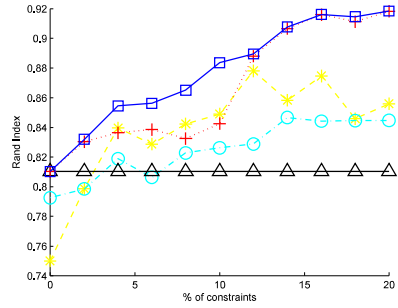
Hepatitis ($K = 2$)



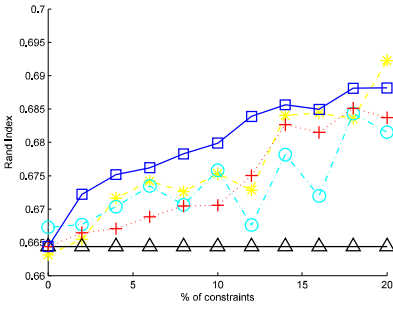
Ionosphere ($K = 2$)



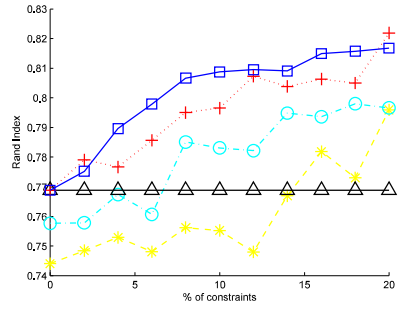
Wine ($K = 3$)



Dermatology ($K = 6$)



Glass ($K = 6$)



Ecoli ($K = 8$)

Figure 2: Average Rand Index according to the constraints rates, on UCI datasets.

- "ORL database" contains 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details.

Table 2 summarizes the characteristics of each database. The affinity matrix terms are of Gaussian form with a scale parameter σ^2 computed in a local way (Zelnik-Manor and Perona, 2004). The experimental protocol used is the same as in Section 6.1. For linear methods (cPCA and cLPP) we used 95% of the total variance which made the maximum dimension for each dataset as follows: Yale Face (71) and ORL (115). For non linear methods, we adopted the classical strategy by keeping the first K eigenvectors as projection coordinates.

Table 2: Other datasets used for experiments.

	Nb. Objects	Nb. Features	Nb. Classes (K)
Yale Face	165	1024	15
ORL	400	1024	40

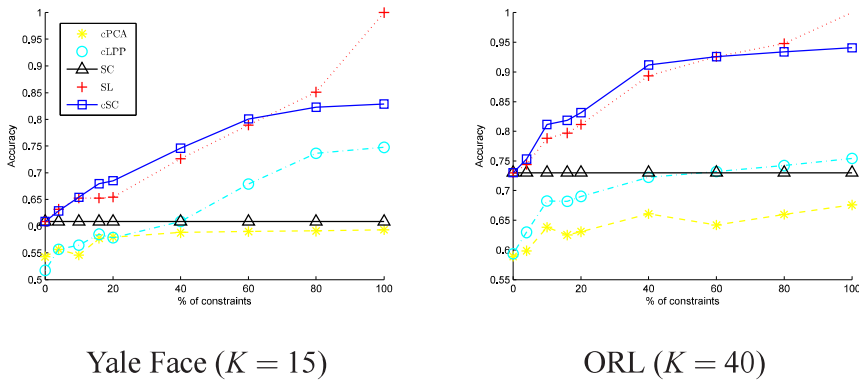


Figure 3: Average Accuracy according to the constraints rates, on other datasets.

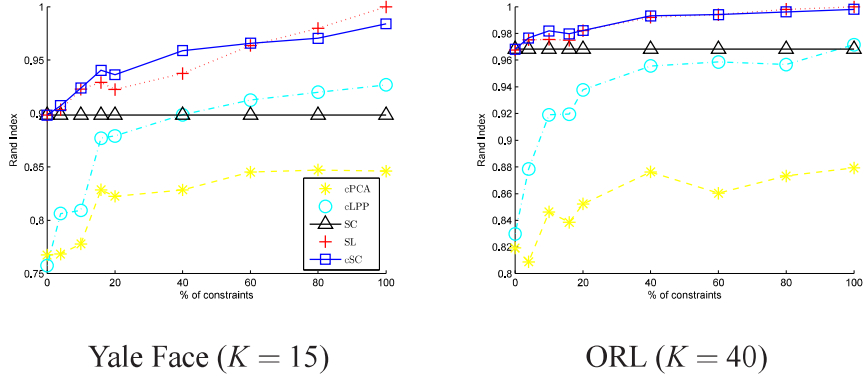


Figure 4: Average Rand Index according to the constraints rates, on other datasets.

356 Figure 3 and 4 show respectively the Accuracy and the Rand Index of the spec-
 357 tral algorithms according to the constraints rates, applied on the two databases. As
 358 for UCI datasets, when the number of constraints is low, it is easy to see that cSC
 359 algorithm outperforms all algorithms, followed by SL algorithm. Moreover, for
 360 this two examples, Rand Indices and Accuracies of the linear projection algo-
 361 rithms (cPCA and cLPP) are lower than the performance scores obtained by the
 362 classical spectral clustering even with 20% of constraints. When the percentage
 363 of constraints is high, SL algorithm obtains the best scores. One possible expla-
 364 nation is that, in our experiments, the chosen γ was arbitrarily set to 0.5, so that
 365 label information was not fully exploited in the global criterion.

366 7. Discussions and conclusions

367 In this work, we presented recent linear and non-linear projection algorithms
 368 integrating constraints and applied for data clustering. PCA and LPP are linear
 369 dimension reduction methods. They find linear relationship between input data
 370 space and output projection space. PCA approach is sensitive to global data dis-

371 persion, while LPP is sensitive to local data variation. Constrained PCA (cPCA)
372 and constrained LPP (cLPP) inherit the same properties of PCA and LPP meth-
373 ods. However, contrary to cPCA, the cLPP encourages nearby data input to be
374 projected to nearby outputs, and penalizes neighboring points if they are projected
375 far apart.

376 It would be important to note that, the main step in spectral clustering method
377 is to project the data in a non linear manner and therefore the resulting spectral
378 space is similar to the one obtained by Laplacian eigenmap method (Belkin and
379 Niyogi, 2002). Indeed, spectral clustering and LPP methods optimize the same
380 objective function. However, unlike SC, the LPP method finds a linear relation-
381 ship between input data space and output projected data space. Thus, the proposed
382 method and cLPP methods inherit the same properties of SC and LPP methods.

383 Spectral clustering approaches encode instance-level constraints implicitly in
384 the affinity matrix (Kamvar et al., 2003) or explicitly in the optimization criterion
385 (Wang and Davidson, 2010). The main weakness of the spectral learning is that
386 it implicitly encodes "must-link" and "cannot-link" constraints by modifying the
387 Laplacian matrix. A more natural approach is to preserve the original Laplacian
388 matrix and to explicitly encode the constraints in the optimization criterion.

389 (Wang et al., 2012) proposed a smart approach which combines spectral clus-
390 tering and pairwise constraints in a flexible manner. Their FCSC algorithm allows
391 the violation of some constraints by introducing a lower-bound of satisfied con-
392 straints in the optimization criterion. However, the associated Laplacian matrix
393 may contain negative eigenvalues and lead sometimes to no solution.

394 In this work, we proposed an efficient constrained spectral clustering algo-
395 rithm which balances the unlabelled data contribution with the pairwise relation-

396 ships and compared its performance to the recent constrained clustering algo-
397 rithms. Many UCI benchmark databases and face recognition databases (ORL
398 and Yale Face datasets) have been used to demonstrate the relevance of the pro-
399 posed algorithm compared to the most known algorithms.

400 In our experiments, the constraints have been randomly generated based on
401 labelled data. However, it may happen that some generated constraints are re-
402 dundant or inconsistent which may deteriorate the performance of classification
403 algorithms (Davidson et al., 2006). Therefore, the constraints should be generated
404 in an intelligent manner, and guided by a human expert.

405

406 **Acknowledgement.** The authors are very grateful to the editors and reviewers
407 for their valuable comments and suggestions.

408 **References**

409 Alzate C., Suykens J.A.K., 2009. A Regularized Formulation for Spectral Clus-
410 tering with Pairwise Constraints. Proceedings of the 2009 International Joint
411 Conference on Neural Networks (IJCNN'09), Atlanta, U.S.A, pp. 141-148.

412 Alzate C., Suykens J.A.K., 2010. Multiway Spectral Clustering with Out-of-
413 Sample Extensions through Weighted Kernel PCA. IEEE Transactions on Pat-
414 tern Analysis and Machine Intelligence, vol. 32, pp. 335-347.

415 Alzate C., Suykens J.A.K., 2012. A Semi-Supervised Formulation to Binary Ker-
416 nel Spectral Clustering. Proceedings of the 2012 IEEE World Congress on
417 Computational Intelligence (IEEE WCCI/IJCNN 2012), Brisbane, Australia.

418 Basu S., Davidson I., Wagstaff K., 2008. Constrained Clustering: Advances in

- 419 Algorithms, Theory and Applications. Chapman and Hall, CRC Press, part of
420 the Data Mining and Knowledge Discovery Series.
- 421 Belkin M., Niyogi P., 2002. Laplacian Eigenmaps for Dimensionality Reduction
422 and Data Representation. *Neural Computation* (vol. 15), pp. 1373-1396.
- 423 Cai D., He X., Han J., 2005. Document Clustering Using Preserving Indexing.
424 *IEEE Transactions on Knowledge and Data Engineering* (vol. 17), pp. 1624-
425 1637.
- 426 Carpaneto G., Toth P., 1980. Algorithm 548: solution of assignment problem.
427 *ACM Transactions on Mathematical Software*.
- 428 Cevikalp H., Verbeek J., 2008. Semi-supervised dimensionality reduction using
429 pairwise equivalence constraints. *International Conference on Computer Vision*
430 *Theory and Applications*, pp. 489-496.
- 431 Chapelle O., Scholkopf B., Zien A., 2006. *Semi-Supervised Learning*. MIT Press,
432 Cambridge MA.
- 433 Chen C., Zhang L., Bu J., Wang C., Chen W., 2010. Constrained laplacian eigen-
434 map for dimensionality reduction. *Neurocomputing Journal*, pp. 951-958.
- 435 Davidson I., Wagstaff K., Basu S., 2006. Measuring constraint set utility for parti-
436 tional clustering algorithms. *Proceedings of the European Conference on Prin-
437 ciples and Practice of Knowledge Discovery in Databases*, pp. 115-126.
- 438 Davidson I., 2009. Knowledge Driven Dimension Reduction for Clustering. *Inter-
439 national Joint Conference on Artificial Intelligence*, pp. 1034-1039.

- 440 De Bie T., Suykens J.A.K., De Moor B., 2004. Learning from general label con-
441 straints. Proceedings of the joint IAPR international workshops on Syntactical
442 and Structural Pattern Recognition and Statistical Pattern Recognition (SSSPR
443 2004), Lisbon, Portugal, vol. 3138.
- 444 Han J., Kamber M., Pei J., 2011. Data Mining: Concepts and Techniques. Morgan
445 Kaufmann Publishers.
- 446 He X., Niyogi P., 2002. Locality preserving projections. Computer and Informa-
447 tion Science, pp. 153-160.
- 448 He X., Yan S., Hu Y., Niyogi P., Zhang H., 2005. Face Recognition Using Lapla-
449 cianfaces. IEEE Transactions on Pattern Analysis and Machine Inetlligence
450 (vol. 27), pp. 328-340.
- 451 Kamvar S., Klein D., Manning C., 2003. Spectral Learning. International Joint
452 Conference on Artificial Intelligence, pp. 561-566.
- 453 Kunegis J., Schmidt S., Lommatzsch A., Lerner J., De Luca E., Albayrak S., 2010.
454 Spectral analysis of Signed Graphs for Clustering, Prediction and Visualization.
455 SIAM, pp. 559-570.
- 456 Meila M., Shi J., 2000. Learning segmentation by random walks. Neural Informa-
457 tion Processing Systems NIPS12, pp. 873-879.
- 458 Ng A., Jordan M., Weiss Y., 2002. On spectral clustering: Analysis and an algo-
459 rithm. Neural Information Processing Systems NIPS14, pp. 849-856.
- 460 Rand W., 1971. Objective criteria for the evaluation of clustering methods. Journal
461 of the American Statistical Association, pp. 846-850.

- 462 Saul L., Weinberger K., Sha F., Ham J., Lee D., 2006. Spectral Methods for Di-
463 mensionality Reduction. Semi-Supervised Learning book, MIT Press, pp. 267-
464 282.
- 465 Shi J., Malik J., 2000. Normalized cuts and image segmentation. IEEE Transac-
466 tions on Pattern Analysis and Machine Intelligence, pp. 888-905.
- 467 Shortreed S., Meila M., 2005. Unsupervised spectral learning. Proceedings of the
468 Twenty-First Conference Annual on Uncertainty in Artificial Intelligence, pp.
469 543-416.
- 470 Suykens J.A.K., Vandewalle J., 1999. Least squares support vector machine clas-
471 sifiers. Neural Processing Letters, vol. 9(3), pp. 293-300.
- 472 Tang W., Zhong S., 2007. Pairwise Constraints-Guided Dimensionality Reduc-
473 tion. Computational Methods of Feature Selection, Chapman and Hall, CRC
474 2007, pp. 295-312.
- 475 Von Luxburg U., 2007. A Tutorial on Spectral Clustering. Statistics and Comput-
476 ing (Vol.17, Issue 4), pp. 395-416.
- 477 Wagstaff K., Cardie C., 2002. Clustering with Instance-level Constraints. Interna-
478 tional Conference on Machine Learning, pp. 1103-1110.
- 479 Wang X., Davidson I., 2010. Flexible Constrained Spectral Clustering. Interna-
480 tional Conference on Knowledge Discovery and Data Mining, pp. 563-572.
- 481 Wang X., Qian B., Davidson I., 2012. On Constrained Spectral Clustering and Its
482 Applications. arXiv: 1201.5338.

- 483 Xu Q., DesJardins M., Wagsatff K., 2005. Constrained spectral clustering under a
484 local proximity structure assumption. FLAIRS Conference, pp. 866-867.
- 485 Yu G., Peng H., Wei J., Ma Q., 2010. Robust locality preserving projections with
486 pairwise constraints. Joint Symposium on Information Systems, pp. 1631-1636.
- 487 Zelnik-Manor L., Perona P., 2004. Self tuning spectral clustering. Advances in
488 Neural Information Processing Systems, pp. 1601-1608.
- 489 Zhang D., Zhou Z.-H., Chen S., 2007. Semi-supervised dimensionality reduction.
490 Seventh International Conference on Data Mining, pp. 629-634.
- 491 Zheng X., Cai D., He X., Ma W.-Y., Lin X. 2004. Locality preserving clustering
492 for image database. Proceedings of the ACM Conference on Multimedia. ACM
493 Press, pp 885-891.