Journal of Machine Learning Research 14 (2013) 2641-2687 Submitted 3/12; Revised 3/13; Published 9/13

Maximum Volume Clustering: A New Discriminative Clustering Approach*

Gang Niu GANG@SG.CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
2-12-1 O-okayama
Meguro-ku
Tokyo, 152-8552, Japan

Bo Dai BOHR.DAI@GMAIL.COM
College of Computing
Georgia Institute of Technology
801 Atlantic Drive
Atlanta, GA 30332, USA

Lin Shang SHANGLIN@NJU.EDU.CN
State Key Laboratory for Novel Software Technology
Nanjing University
163 Xianlin Avenue
Qixia District
Nanjing 210023, China

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
2-12-1 O-okayama
Meguro-ku
Tokyo, 152-8552, Japan

Editor: Ulrike von Luxburg

Abstract

The *large volume principle* proposed by Vladimir Vapnik, which advocates that hypotheses lying in an equivalence class with a larger volume are more preferable, is a useful alternative to the *large margin principle*. In this paper, we introduce a new discriminative clustering model based on the large volume principle called *maximum volume clustering* (MVC), and then propose two approximation schemes to solve this MVC model: A soft-label MVC method using *sequential quadratic programming* and a hard-label MVC method using *semi-definite programming*, respectively. The proposed MVC is theoretically advantageous for three reasons. The optimization involved in hard-label MVC is convex, and under mild conditions, the optimization involved in soft-label MVC is akin to a convex one in terms of the resulting clusters. Secondly, the soft-label MVC method pos-

*. A preliminary and shorter version has appeared in Proceedings of 14th International Conference on Artificial Intelligence and Statistics (Niu et al., 2011). The preliminary work was done when GN was studying at Department of Computer Science and Technology, Nanjing University, and BD was studying at Institute of Automation, Chinese Academy of Sciences. A Matlab implementation of maximum volume clustering is available from `http://sugiyama-www.cs.titech.ac.jp/~gang/software.html`.

sesses a *clustering error bound*. Thirdly, MVC includes the optimization problems of a spectral clustering, two relaxed k-means clustering and an information-maximization clustering as *special limit cases* when its regularization parameter goes to infinity. Experiments on several artificial and benchmark data sets demonstrate that the proposed MVC compares favorably with state-of-the-art clustering methods.

Keywords: discriminative clustering, large volume principle, sequential quadratic programming, semi-definite programming, finite sample stability, clustering error bound

1. Introduction

Clustering has been an important topic in machine learning and data mining communities. Over the past decades, a large number of clustering algorithms have been developed. For instance, *k-means clustering* (MacQueen, 1967; Hartigan and Wong, 1979; Girolami, 2002), *spectral clustering* (Shi and Malik, 2000; Meila and Shi, 2001; Ng et al., 2002), *maximum margin clustering* (MMC) (Xu et al., 2005; Xu and Schuurmans, 2005), *dependence-maximization clustering* (Song et al., 2007; Faivishevsky and Goldberger, 2010) and *information-maximization clustering* (Agakov and Barber, 2006; Gomes et al., 2010; Sugiyama et al., 2011). These algorithms have been successfully applied to diverse real-world data sets for exploratory data analysis.

To the best of our knowledge, MMC, which partitions the data samples into two clusters based on the *large margin principle* (LMP) (Vapnik, 1982), is the first clustering approach that is directly connected to the *statistical learning theory* (Vapnik, 1998). For this reason, it has been extensively investigated recently, for example, a generalization (Valizadegan and Jin, 2007) and many approximations for speedup (Zhang et al., 2007; Zhao et al., 2008b,a; Li et al., 2009; Wang et al., 2010).

However, LMP is not the only way to go in statistical learning theory. The *large volume principle* (LVP) was also introduced by Vapnik (1982) for *hyperplanes* and then extended by El-Yaniv et al. (2008) for *soft response vectors*. Roughly speaking, learning methods based on LVP should prefer hypotheses in certain large-volume equivalence classes. See Figure 1 as an illustrative comparison of two principles. Here, C_1, C_2 and C_3 represent three data clouds, and our goal is to choose a better hypothesis from two candidates h_1 and h_2. A hypothesis is a line (e.g., h_1), and an equivalence class is a set of lines which equivalently separate data samples (e.g., H_1). Hence, there are two equivalence classes H_1 and H_2. Given an equivalence class H_1 (or H_2), its margin is measured by the distance between two lines around it and its volume is measured by the area of the region around it in the figure. Though LMP prefers h_1 due to the larger margin of H_1 than H_2, we should choose h_2 when considering LVP since H_2 has a larger volume than H_1.

In this paper, we introduce a novel discriminative clustering approach called *maximum volume clustering* (MVC), which serves as a prototype to partition the data samples into two clusters based on LVP. We motivate our MVC as follows. Given the samples X_n, we construct an X_n-dependent hypothesis space $\mathcal{H}(X_n)$. If $\mathcal{H}(X_n)$ has a measure on it, namely the *power*, then we can talk about the *likelihood* or *confidence* of each equivalence class (Vapnik, 1998). Similarly to the *margin* used in MMC, the notion of *volume* (El-Yaniv et al., 2008) can also be regarded as an estimation of the power. Therefore, the larger the volume is, the more confident we are of the data partition, and we consider the partition lying in the equivalence class with the maximum volume as the best partition.

Similarly to the majority of clustering methods, the optimization problem involved in MVC is combinatorial and thus NP-hard, so we propose two approximation schemes:

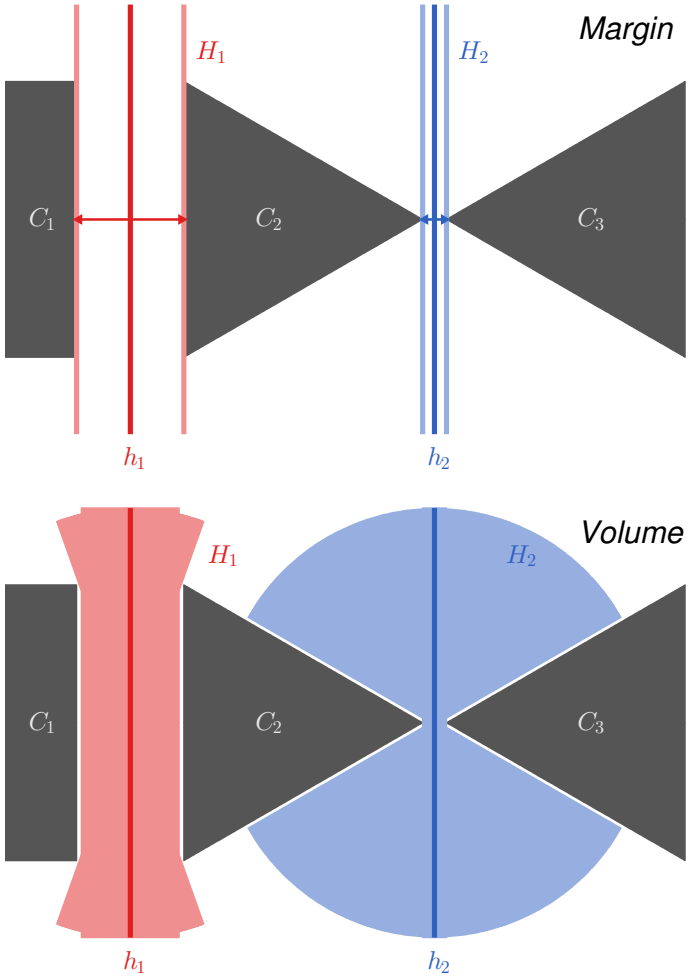


Figure 1: Large margin vs. large volume separation of three data clouds into two clusters. In this figure, the three data clouds are C_1, C_2 and C_3, and the two candidate hypotheses are h_1 and h_2. A hypothesis is a line (e.g., h_1), and an equivalence class is a set of lines which equivalently separate data samples (e.g., H_1). More specifically, we shape H_1 and H_2 by horizontally translating and rotating h_1 and h_2. Then given an equivalence class H_1 (or H_2), its margin is measured by the distance between two lines around it and its volume is measured by the area of the region around it (where we integrate all unit line segments and treat the resulting area as its volume). The large margin principle prefers h_1 and the large volume principle prefers h_2, since they consider different complexity measures.

- A soft-label MVC method that can be solved by *sequential quadratic programming* (Boggs and Tolle, 1995) in $O(n^3)$ time;

- A hard-label MVC method as a *semi-definite programming* problem (De Bie and Cristianini, 2004; Lanckriet et al., 2004) that can be solved in $O(n^{6.5})$ time.

Subsequently, we show that the primal problem of soft-label MVC can be reduced to the optimization problems of *unnormalized spectral clustering* (von Luxburg, 2007), plain and kernel *k-means clustering* after relaxations (Ding and He, 2004), and *squared-loss mutual information based clustering* (Sugiyama et al., 2011), as the regularization parameter of MVC approaches infinity. Hence, MVC might be regarded as a natural extension of many existing clustering methods. Moreover, we establish two theoretical results:

- A theory called *finite sample stability* for analyzing the soft-label MVC method. It suggests that under mild conditions, different locally optimal solutions to soft-label MVC would induce the same data partition, and thus the non-convex optimization of soft-label MVC seems like a convex one;
- A *clustering error bound* for the soft-label MVC method. It upper bounds the distance between the partition returned by soft-label MVC and any partially observed partition based on *transductive Rademacher complexity* (El-Yaniv and Pechyony, 2009).

Experiments on three artificial and fourteen benchmark data sets (i.e., ten IDA benchmarks, USPS, MNIST, 20Newsgroups and Isolet) demonstrate that the proposed MVC approach is promising.

The rest of this paper is organized as follows. First of all, we briefly review the large volume approximation in Section 2. Then, we propose the model and algorithms of MVC in Section 3, and show that they are closely related to several existing clustering methods in Section 4. In Section 5, we present the theory of finite sample stability. In Section 6, we derive the clustering error bound. Next, a comparison with related works is made in Section 7. Experimental results are reported in Section 8. Finally, we give concluding remarks and future prospects in Section 9.

2. Large Volume Approximation

Suppose that we are given a set of objects $X_n = \{x_1, \ldots, x_n\}$, where $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$, and most often but not necessarily, $\mathcal{X} \subset \mathbb{R}^d$ for some natural number d. We will construct a *hypothesis space* $\mathcal{H}(X_n)$ that depends on X_n, such that for any hypothesis $\boldsymbol{h} \in \mathcal{H}(X_n) \subset \mathbb{R}^n$, $[\boldsymbol{h}]_i$ stands for a *soft response* or *confidence-rated label* of x_i, where $[\cdot]_i$ means the i-th component of a vector. We will then pick a *soft response vector* $\boldsymbol{h}^*$ following the large volume principle and partition X_n into two clusters $\{x_i \mid [\boldsymbol{h}^*]_i > 0\}$ and $\{x_i \mid [\boldsymbol{h}^*]_i < 0\}$.[1]

As El-Yaniv et al. (2008), assume that we have a symmetric positive-definite matrix $Q \in \mathbb{R}^{n \times n}$ which contains the pairwise information about X_n. Consider the hypothesis space

$$\mathcal{H}_Q := \{\boldsymbol{h} \mid \boldsymbol{h}^\top Q \boldsymbol{h} \leq 1\},$$

which is geometrically an origin-centered ellipsoid $\mathcal{E}(\mathcal{H}_Q)$ in $\mathbb{R}^n$. The set of sign vectors

$$\{\mathrm{sign}(\boldsymbol{h}) \mid \boldsymbol{h} \in \mathcal{H}_Q\}$$

contains all 2^n possible dichotomies of X_n. In other words, the hypothesis space $\mathcal{H}_Q$ has been partitioned into a finite number of *equivalence classes* $H_1, \ldots, H_{2^n}$, such that for fixed $k \in \{1, 2, \ldots, 2^n\}$,

1. Due to our clustering model that will be defined as optimization (2) in page 2646, $[\boldsymbol{h}^*]_i = 0$ hardly happens in practice, and we simply assume $[\boldsymbol{h}^*]_i \neq 0$ in our problem setting.

all hypotheses in H_k will generate the same dichotomy of X_n. The *power* of an equivalence class H_k is defined as a probability mass

$$\mathcal{P}(H_k) := \int_{H_k} p(\boldsymbol{h}) \mathrm{d}\boldsymbol{h}, \quad k = 1, \ldots, 2^n,$$

where $p(\boldsymbol{h})$ is the underlying probability density of $\boldsymbol{h}$ over $\mathcal{H}_Q$. The hypotheses in H_k with a large power $\mathcal{P}(H_k)$ are preferred according to statistical learning theory (Vapnik, 1998).

When no specific domain knowledge is available (i.e., $p(\boldsymbol{h})$ is unknown), it would be natural to assume the continuous uniform distribution $p(\boldsymbol{h}) = 1/\sum_{k=1}^{2^n} \mathcal{V}(H_k)$, where

$$\mathcal{V}(H_k) := \int_{H_k} \mathrm{d}\boldsymbol{h}, \quad k = 1, \ldots, 2^n,$$

is the *volume* of H_k as well as the geometric volume of the k-th quadrant of $\mathcal{E}(\mathcal{H}_Q)$. Consequently, $\mathcal{P}(H_k)$ is proportional to $\mathcal{V}(H_k)$, and the larger the value of $\mathcal{V}(H_k)$ is, the more confident we are of the data partition $\mathrm{sign}(\boldsymbol{h}^*)$ where $\boldsymbol{h}^*$ is chosen from H_k.

However, it is very hard to accurately compute the geometric volume of a single n-dimensional convex body let alone for all 2^n convex bodies, so we employ an efficient approximation introduced by El-Yaniv et al. (2008) as follows. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of Q, and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the associated normalized eigenvectors. Then, $\boldsymbol{v}_i$ and $1/\sqrt{\lambda_i}$ are the direction and length of the i-th principal axis of $\mathcal{E}(\mathcal{H}_Q)$. Note that a small angle from some $\boldsymbol{h} \in H_k$ to $\boldsymbol{v}_i$ with a small/large index i (i.e., a long/short principal axis) implies that $\mathcal{V}(H_k)$ is large/small. Based on this key observation, we define

$$V(\boldsymbol{h}) := \sum_{i=1}^{n} \lambda_i \left(\frac{\boldsymbol{h}^\top \boldsymbol{v}_i}{\|\boldsymbol{h}\|_2} \right)^2 = \frac{\boldsymbol{h}^\top Q \boldsymbol{h}}{\|\boldsymbol{h}\|_2^2}, \tag{1}$$

where $\boldsymbol{h}^\top \boldsymbol{v}_i / \|\boldsymbol{h}\|_2$ means the cosine of the angle between $\boldsymbol{h}$ and $\boldsymbol{v}_i$. We subsequently expect $V(\boldsymbol{h})$ to be small when $\boldsymbol{h}$ lies in a large-volume equivalence class, and conversely to be large when $\boldsymbol{h}$ lies in a small-volume equivalence class.

3. Maximum Volume Clustering

In this section, we define our clustering model and propose two approximation algorithms.

3.1 Basic Formulation

Motivated by Xu et al. (2005), we think of the binary clustering problem from a regularization viewpoint. If we had labels $Y_n = \{y_1, \ldots, y_n\}$ where $y_i \in \{-1, +1\}$, we could find a certain classification method to compute

$$\vartheta(X_n, Y_n) := \min_{\boldsymbol{h} \in \mathcal{H}(X_n, Y_n)} \Delta(Y_n, \boldsymbol{h}) + \gamma W(X_n, \boldsymbol{h}),$$

where $\mathcal{H}(X_n, Y_n)$ is a hypothesis space (which depends upon X_n and Y_n), $\Delta(Y_n, \boldsymbol{h})$ is an overall loss function, $W(X_n, \boldsymbol{h})$ is a regularization function, and $\gamma > 0$ is a regularization parameter. The value of $\vartheta(X_n, Y_n)$ is generally a measure of *classification quality*.

When the labels Y_n are absent, a clustering method tries to minimize $\vartheta(X_n, \boldsymbol{y})$ over all possible assignments $\boldsymbol{y} \in \{-1, +1\}^n$ for given X_n, that is, to solve the problem

$$\boldsymbol{y}^* = \operatorname*{arg\,min}_{\boldsymbol{y} \in \{-1, +1\}^n} \vartheta(X_n, \boldsymbol{y}).$$

Generally speaking, $\vartheta(X_n, \boldsymbol{y}^*)$ can be regarded as a measure of *clustering quality*. The smaller the value of $\vartheta(X_n, \boldsymbol{y}^*)$ is, the more satisfied we are with the resulting data partition $\boldsymbol{y}^*$.

In our discriminative clustering model, we hope to use $V(\boldsymbol{h})$ in Equation (1) as our regularization function. Formally speaking, given the matrix Q, by instantiating $\Delta(\boldsymbol{y}, \boldsymbol{h}) = -2\boldsymbol{h}^\top \boldsymbol{y}$, we define the basic model of *maximum volume clustering* (MVC) as

$$\min_{\boldsymbol{y} \in \{-1,+1\}^n} \min_{\boldsymbol{h} \in \mathcal{H}_Q} -2\boldsymbol{h}^\top \boldsymbol{y} + \gamma \cdot \frac{\boldsymbol{h}^\top Q \boldsymbol{h}}{\|\boldsymbol{h}\|_2^2}, \tag{2}$$

where $\mathcal{H}_Q = \{\boldsymbol{h} \mid \boldsymbol{h}^\top Q \boldsymbol{h} \leq 1\}$ is the hypothesis space mentioned in Section 2, and $\gamma > 0$ is a regularization parameter. Optimization problem (2) is computationally intractable, due to not only the non-convexity of $V(\boldsymbol{h})$ but also the integer feasible region of $\boldsymbol{y}$ which makes (2) combinatorial. In the next two subsections, we will discuss two approximation schemes of (2) in detail.

3.2 Soft-Label Approximation

We now try to optimize $\boldsymbol{h}$ alone by removing $\boldsymbol{y}$. After exchanging the order of the minimizations of $\boldsymbol{y}$ and $\boldsymbol{h}$ in optimization (2), it is easy to see that the optimal $\boldsymbol{y}$ should be $\operatorname{sign}(\boldsymbol{h})$, since the second term is independent of $\boldsymbol{y}$ and the first term is minimized when $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{h})$ for fixed $\boldsymbol{h}$. Therefore, (2) becomes

$$\min_{\boldsymbol{h} \in \mathcal{H}_Q} -2\|\boldsymbol{h}\|_1 + \gamma \cdot \frac{\boldsymbol{h}^\top Q \boldsymbol{h}}{\|\boldsymbol{h}\|_2^2}. \tag{3}$$

Similarly to El-Yaniv et al. (2008), we replace the feasible region $\mathcal{H}_Q$ with $\mathbb{R}^n$, and relax (3) into

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} -2\|\boldsymbol{h}\|_1 + \gamma \boldsymbol{h}^\top Q \boldsymbol{h} \qquad \text{s.t. } \|\boldsymbol{h}\|_2 = 1. \tag{4}$$

Although the optimization is done in $\mathbb{R}^n$, the regularization is done relative to $\mathcal{H}_Q$. Optimization (4) is the primal problem of *soft-label MVC* (MVC-SL).

Optimization (4) is non-convex mainly attributed to the minimization of negative ℓ_1-norm rather than the equality constraint of ℓ_2-norm. In order to solve this optimization, we resort to *sequential quadratic programming* (SQP) (Boggs and Tolle, 1995). The basic idea of SQP is modeling a non-convex problem by a sequence of convex subproblems: At each step, it uses a quadratic model for the objective function and linear models for the constraints. A nonlinear optimization problem with a quadratic objective function and linear constraints is known as *quadratic programming* (QP). An SQP constructs and solves a local QP at each iteration, yielding a step toward the optimum.

More specifically, let us include a class balance constraint $-b \leq \boldsymbol{h}^\top \mathbf{1}_n \leq b$ with a user-specified class balance parameter $b > 0$ to prevent skewed clustering sizes. Denote the objective function of optimization (4) by

$$f(\boldsymbol{h}) := -2\boldsymbol{h}^\top \operatorname{sign}(\boldsymbol{h}) + \gamma \boldsymbol{h}^\top Q \boldsymbol{h},$$

and the auxiliary functions by

$$f_1(\boldsymbol{h}) := \boldsymbol{h}^\top \boldsymbol{h} - 1,$$
$$f_2(\boldsymbol{h}) := \boldsymbol{h}^\top \mathbf{1}_n,$$

where $\mathbf{1}_n$ means the all-one vector in $\mathbb{R}^n$. Subsequently, let λ_1 be the smallest eigenvalue of Q, the corresponding Lagrange function should be[2]

$$L(\boldsymbol{h},\eta,\mu,\nu) = f(\boldsymbol{h}) - \eta f_1(\boldsymbol{h}) - \mu(f_2(\boldsymbol{h}) - b) + \nu(f_2(\boldsymbol{h}) + b),$$

where $\eta < \gamma\lambda_1$ is the Lagrangian multiplier for the constraint $f_1(\boldsymbol{h}) = 0$, and $\mu, \nu \geq 0$ are the Lagrangian multipliers for the constraint $-b \leq f_2(\boldsymbol{h}) \leq b$. Then, given constant $\boldsymbol{h}$ and variable $\boldsymbol{p}$ with a tiny norm, the auxiliary functions can be approximated by

$$\begin{aligned} f_1(\boldsymbol{h}+\boldsymbol{p}) &\approx \boldsymbol{p}^\top \nabla f_1(\boldsymbol{h}) + f_1(\boldsymbol{h}), \\ f_2(\boldsymbol{h}+\boldsymbol{p}) &= \boldsymbol{p}^\top \nabla f_2(\boldsymbol{h}) + f_2(\boldsymbol{h}), \end{aligned}$$

so the constraints are replaced with

$$\begin{aligned} \boldsymbol{p}^\top \nabla f_1(\boldsymbol{h}) + f_1(\boldsymbol{h}) &= 0, \\ -b \leq \boldsymbol{p}^\top \nabla f_2(\boldsymbol{h}) + f_2(\boldsymbol{h}) &\leq b. \end{aligned}$$

Nevertheless, it would be incorrect to adopt the second-order Taylor expansion of $f(\boldsymbol{h}+\boldsymbol{p})$ as our new objective function, since we need to capture the curvature of $f_1(\boldsymbol{h}+\boldsymbol{p})$. The correct way is to use the quadratic model[3]

$$L(\boldsymbol{h}+\boldsymbol{p},\eta) \approx \frac{1}{2}\boldsymbol{p}^\top \nabla^2 L(\boldsymbol{h},\eta)\boldsymbol{p} + \boldsymbol{p}^\top \nabla L(\boldsymbol{h},\eta) + L(\boldsymbol{h},\eta)$$

and form our objective at any fixed $(\boldsymbol{h},\eta)$ into

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{p}^\top \nabla^2 L(\boldsymbol{h},\eta)\boldsymbol{p} + \boldsymbol{p}^\top \nabla f(\boldsymbol{h}),$$

according to Boggs and Tolle (1995, p. 9). As a consequence, the subproblem of the t-th iteration is a simple QP at the current estimate $(\boldsymbol{h}_t, \eta_t)$:

$$\begin{aligned} \min_{\boldsymbol{p}_t\in\mathbb{R}^n} \ & \boldsymbol{p}_t^\top(\gamma Q - \eta_t I_n)\boldsymbol{p}_t + 2\boldsymbol{p}_t^\top\big(\gamma Q\boldsymbol{h}_t - \operatorname{sign}(\boldsymbol{h}_t)\big) \\ \text{s.t.} \ & 2\boldsymbol{p}_t^\top\boldsymbol{h}_t + \boldsymbol{h}_t^\top\boldsymbol{h}_t = 1 \\ & -b \leq \boldsymbol{p}_t^\top\mathbf{1}_n + \boldsymbol{h}_t^\top\mathbf{1}_n \leq b, \end{aligned} \tag{5}$$

where I_n is the identity matrix of size n. The new estimate $(\boldsymbol{h}_{t+1}, \eta_{t+1})$ is given by

$$\boldsymbol{h}_{t+1} = \boldsymbol{h}_t + \boldsymbol{p}_t^*, \tag{6}$$

$$\eta_{t+1} = \frac{\boldsymbol{h}_t^\top\big(\gamma Q\boldsymbol{h}_{t+1} - \eta_t\boldsymbol{p}_t^* - \operatorname{sign}(\boldsymbol{h}_t)\big)}{\boldsymbol{h}_t^\top\boldsymbol{h}_t}, \tag{7}$$

2. We will ignore variables μ and ν later, since first-order terms of $L(\boldsymbol{h},\eta,\mu,\nu)$ would disappear in the second-order derivative $\nabla^2 L(\boldsymbol{h},\eta,\mu,\nu)$. The Lagrange function $L(\boldsymbol{h},\eta)$ itself has no constraint on $\boldsymbol{h}$, so we impose $\eta < \gamma\lambda_1$ to make sure that $L(\boldsymbol{h},\eta)$ is bounded from below. Otherwise, the subproblem may be ill-defined.

3. Note that minimizing $-\boldsymbol{h}^\top\boldsymbol{y}$ in optimization (2) or $-\|\boldsymbol{h}\|_1$ in optimization (4) has an effect to push $\boldsymbol{h}$ away from the coordinate axes of $\mathbb{R}^n$. Thus, $[\boldsymbol{h}]_i = 0$ hardly happens in practice and we assume that $\|\boldsymbol{h}\|_1$ is always differentiable.

Algorithm 1 MVC-SL

Input: stopping criterion ε,
symmetric positive-definite matrix Q,
regularization parameter γ,
class balance parameter b
Output: soft response vector $\boldsymbol{h}^*$

Initialize $(\boldsymbol{h}_0, \eta_0)$, recommended but not required, from Equation (9)
$t \leftarrow 0$
repeat
  Obtain $\boldsymbol{p}_t^*$ through optimization (5)
  Update $\boldsymbol{h}_{t+1}$ through Equation (6)
  Update η_{t+1} through Equation (7)
  if $\eta_{t+1} \geq \gamma\lambda_1$ **then break**
  $t \leftarrow t+1$
until $\|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\|_2 + |\eta_t - \eta_{t-1}| \leq \varepsilon$
return $\boldsymbol{h}^* = \boldsymbol{h}_t$

where $\boldsymbol{p}_t^*$ is the optimal solution to (5). Notice that we cannot obtain η_{t+1} directly from (5) and in fact Equation (7) comes from the best fit in the least-square sense of the following equation

$$\nabla^2 L(\boldsymbol{h}_t, \eta_t)\boldsymbol{p}_t^* + \nabla f(\boldsymbol{h}_t) - \eta_{t+1}\nabla f_1(\boldsymbol{h}_t) = 0. \tag{8}$$

The MVC-SL algorithm based on SQP is summarized in Algorithm 1. In our experiments, we use an initial solution $(\boldsymbol{h}_0, \eta_0)$ defined as

$$\boldsymbol{h}_0 = \frac{1}{\sqrt{n}} \operatorname{sign}\left(\boldsymbol{v}_2 - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \boldsymbol{v}_2\right), \quad \eta_0 = 0, \tag{9}$$

where $\boldsymbol{v}_2$ is the eigenvector associated with the second smallest eigenvalue of Q. The construction of $\boldsymbol{h}_0$ is explained as follows. The term $(\boldsymbol{v}_2 - \mathbf{1}_n\mathbf{1}_n^\top \boldsymbol{v}_2/n)$ equals $C_n\boldsymbol{v}_2$, where $C_n = I_n - \mathbf{1}_n\mathbf{1}_n^\top/n$ is the centering matrix of size n. It means that we cut $\boldsymbol{v}_2$ at the mean value of its components to form two initial clusters, and normalize the corresponding soft response vector into the unit norm as $\boldsymbol{h}_0$. The asymptotic time complexity of each iteration is at most $O(n^3)$, and the convergence rate of SQP iterations is independent of n (Boggs and Tolle, 1995). Moreover, it takes $O(n^2)$ time to compute $\boldsymbol{h}_0$. Hence, the overall computational complexity of Algorithm 1 is no more than $O(n^3)$.

3.3 Hard-Label Approximation

As opposed to the soft-label approximation, we can also optimize $\boldsymbol{y}$ alone. Let $\boldsymbol{h} = \boldsymbol{\alpha} \circ \boldsymbol{y}$, where $\boldsymbol{\alpha} = |\boldsymbol{h}|$ is a vector of element-wise absolute values, $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{h})$ is a vector of the corresponding signs, and $\circ$ means the element-wise product. We would like to further introduce a hyperparameter C to bound each component of $\boldsymbol{\alpha}$, which might be helpful for dealing with outliers. Subsequently,

the primal problem of *hard-label MVC* (MVC-HL) is written as

$$\begin{aligned} \min_{\boldsymbol{y}\in\{-1,+1\}^n} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \; & -2\boldsymbol{\alpha}^\top \mathbf{1}_n + \gamma \boldsymbol{\alpha}^\top \left(Q \circ \boldsymbol{y}\boldsymbol{y}^\top\right) \boldsymbol{\alpha} \\ \text{s.t.} \; & \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1 \\ & \mathbf{0}_n \le \boldsymbol{\alpha} \le C\mathbf{1}_n, \end{aligned} \tag{10}$$

where $\mathbf{0}_n$ means the all-zero vector in $\mathbb{R}^n$.

By employing the technique described in Lanckriet et al. (2004), let $M = \boldsymbol{y}\boldsymbol{y}^\top$ and then optimization (10) can be relaxed to

$$\begin{aligned} \min_{M\in\mathbb{R}^{n\times n}} \min_{\eta\in\mathbb{R}} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \; & 2\boldsymbol{\alpha}^\top \mathbf{1}_n - \gamma \boldsymbol{\alpha}^\top (Q \circ M)\boldsymbol{\alpha} + \eta(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - 1) \\ \text{s.t.} \; & M \succeq \mathbf{0} \\ & \operatorname{diag}(M) = \mathbf{1}_n \\ & \mathbf{0}_n \le \boldsymbol{\alpha} \le C\mathbf{1}_n, \end{aligned} \tag{11}$$

where the function $\operatorname{diag}(\cdot)$ forms the diagonal entries of a square matrix into a column vector, and $\succeq \mathbf{0}$ indicates the positive semi-definiteness of a symmetric matrix.[4] The relaxation from (10) to (11) is mainly achieved by replacing $M \in \{-1,+1\}^{n\times n}$ and $\operatorname{rank}(M) = 1$ with $M \in \mathbb{R}^{n\times n}$, $M \succeq \mathbf{0}$ and $\operatorname{diag}(M) = \mathbf{1}_n$. As a result, optimization (11) is a *semi-definite programming* (SDP) provided $(\gamma Q \circ M - \eta I_n) \succeq \mathbf{0}$. Let $\boldsymbol{\nu} \ge \mathbf{0}_n$ and $\boldsymbol{\mu} \ge \mathbf{0}_n$ be the Lagrangian multipliers for $\mathbf{0}_n \le \boldsymbol{\alpha}$ and $\boldsymbol{\alpha} \le C\mathbf{1}_n$, then (11) is equivalent to

$$\begin{aligned} \min_{M,\boldsymbol{\mu},\boldsymbol{\nu},\eta} \max_{\boldsymbol{\alpha}} \; & 2\boldsymbol{\alpha}^\top (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu}) - \boldsymbol{\alpha}^\top (\gamma Q \circ M - \eta I_n)\boldsymbol{\alpha} + 2C\boldsymbol{\mu}^\top \mathbf{1}_n - \eta \\ \text{s.t.} \; & M \succeq \mathbf{0} \\ & \operatorname{diag}(M) = \mathbf{1}_n \\ & \boldsymbol{\mu} \ge \mathbf{0}_n, \boldsymbol{\nu} \ge \mathbf{0}_n. \end{aligned} \tag{12}$$

When considering the variable $\boldsymbol{\alpha}$ in (12), the optimal $\boldsymbol{\alpha}$ should be

$$\boldsymbol{\alpha} = (\gamma Q \circ M - \eta I_n)^\dagger (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu}),$$

where $\dagger$ is the operator of the pseudo inverse, and we can form (12) into

$$\begin{aligned} \min_{M,\boldsymbol{\mu},\boldsymbol{\nu},\eta} \; & (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu})^\top (\gamma Q \circ M - \eta I_n)^\dagger (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu}) + 2C\boldsymbol{\mu}^\top \mathbf{1}_n - \eta \\ \text{s.t.} \; & M \succeq \mathbf{0} \\ & \operatorname{diag}(M) = \mathbf{1}_n \\ & \boldsymbol{\mu} \ge \mathbf{0}_n, \boldsymbol{\nu} \ge \mathbf{0}_n \end{aligned}$$

under an additional condition that $(\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu})$ is orthogonal to the null space of $(\gamma Q \circ M - \eta I_n)$. Eventually, by the *extended Schur complement lemma* (De Bie and Cristianini, 2004), we arrive at

4. We imply by $M \succeq \mathbf{0}$ that M is symmetric and will not explicitly write $M^\top = M$ for convenience.

a standard SDP formulation:

$$\begin{aligned} \min_{M,\boldsymbol{\mu},\boldsymbol{\nu},\eta,t} \quad & t \\ \text{s.t.} \quad & M \succeq \mathbf{0} \\ & \operatorname{diag}(M) = \mathbf{1}_n \\ & \boldsymbol{\mu} \geq \mathbf{0}_n, \boldsymbol{\nu} \geq \mathbf{0}_n \\ & \begin{pmatrix} \gamma Q \circ M - \eta I_n & (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu}) \\ (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu})^\top & t + \eta - 2C\boldsymbol{\mu}^\top \mathbf{1}_n \end{pmatrix} \succeq \mathbf{0}. \end{aligned} \tag{13}$$

The asymptotic time complexity of optimization (13) is $O(n^{6.5})$ if directly solved by any standard SDP solver (De Bie and Cristianini, 2004). It could be reduced to $O(n^{4.5})$ with the *subspace tricks* (De Bie and Cristianini, 2006), which essentially make use of the spectral properties of Q to control the trade-off between the computational cost and the accuracy.

After we obtain M^*, $\boldsymbol{y}^*$ could be recovered from the rank one approximation of M^* by either *thresholding* (De Bie and Cristianini, 2004) or *randomized rounding* (Raghavan and Thompson, 1985; De Bie and Cristianini, 2006). In our experiments, we use the former technique: The eigenvector $\boldsymbol{v}^*$ associated with the largest eigenvalue of M^* is extracted, and then $\boldsymbol{y}^*$ is recovered as

$$\boldsymbol{y}^* = \operatorname{sign}\left(\boldsymbol{v}^* - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^\top \boldsymbol{v}^*\right),$$

where the threshold is the center of $\boldsymbol{v}^*$ (cf., the construction of $\boldsymbol{h}_0$ in MVC-SL).

4. Generality

MVC is a general framework and closely related to several existing clustering methods. The primal problem of MVC-SL can in fact be reduced to the optimization problems of *unnormalized spectral clustering* (USC) (von Luxburg, 2007, p. 6), relaxed plain and kernel *k-means clustering* (Ding and He, 2004), and *squared-loss mutual information based clustering* (SMIC) (Sugiyama et al., 2011) as special limit cases. We demonstrate these claims in this section.

First of all, consider USC. The relaxed *RatioCut* problem can formulate USC from a graph cut point of view as

$$\min_{f \in \mathbb{R}^n} f^\top L_{\text{un}} f \qquad \text{s.t. } f \perp \mathbf{1}_n, \|f\|_2 = \sqrt{n} \tag{14}$$

when the number of clusters is two, where L_{un} is the *unnormalized graph Laplacian* (von Luxburg, 2007, pp. 10–11). Note that we can rewrite the primal problem of MVC-SL defined in (4) as

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} -2\|\boldsymbol{h}\|_1/\gamma + \boldsymbol{h}^\top Q \boldsymbol{h} \qquad \text{s.t. } \|\boldsymbol{h}\|_2 = 1. \tag{15}$$

Optimizations (15) and (4) share exactly the same optimal solution with/without the class balance constraint $-b \leq \boldsymbol{h}^\top \mathbf{1}_n \leq b$, though (15) has an optimal objective value γ times smaller than (4)'s. Now, let $Q = L_{\text{un}} + \varepsilon I_n$ with arbitrarily chosen $\varepsilon > 0$ to make sure the positive definiteness of Q. Assume that f^* is the solution to (14), and $\boldsymbol{h}_m^*$ is the solution to (15) with Q specified as above, a class balance parameter $b = 0$, and a regularization parameter $\gamma_m = m$ given a natural number m. Subsequently, it is obvious that

$$\lim_{m \to \infty} \boldsymbol{h}_m^* = f^*/\sqrt{n},$$

and

$$\lim_{m\to\infty} -2\|\boldsymbol{h}_m^*\|_1/\gamma_m + \boldsymbol{h}_m^{*\top} Q \boldsymbol{h}_m^* = f^{*\top} L_{\text{un}} f^*/n + \varepsilon,$$

since $\|\boldsymbol{h}_m^*\|_1 \le \sqrt{n}\|\boldsymbol{h}_m^*\|_2 = \sqrt{n}$ and then $\lim_{m\to\infty} \|\boldsymbol{h}_m^*\|_1/\gamma_m = 0$. Therefore, USC may be viewed as a special limit case of MVC-SL, that is, a special case with the specification $Q = L_{\text{un}} + \varepsilon I_n$ of a limit case as $\gamma \to \infty$.

Remark 1 The motivation of $f \perp \mathbf{1}_n$ in USC is very different from $\boldsymbol{h}^\top \mathbf{1}_n = 0$ in MVC-SL for class balancing. When L_{un} is constructed from a fully connected similarity graph, the constraint $f \perp \mathbf{1}_n$ means that the feasible region of optimization (14) is in a space spanned by all eigenvectors of L_{un} except the trivial eigenvector $\mathbf{1}_n$. Note that $\boldsymbol{h}^\top \mathbf{1}_n = 0$ just asks for strictly balanced soft responses and is not equivalent to $\text{sign}(\boldsymbol{h})^\top \mathbf{1}_n = 0$ that demands strictly balanced cluster assignments.

On the other hand, continuous solutions to the relaxations of k-means clustering (MacQueen, 1967; Hartigan and Wong, 1979) and kernel k-means clustering (Girolami, 2002) can be obtained by principle component analysis (PCA) and kernel PCA, respectively (Zha et al., 2002; Ding and He, 2004). Now, let $Q = \varepsilon I_n - C_n K C_n$ with arbitrarily chosen $\varepsilon > \|C_n K C_n\|_2$, where K is the *kernel matrix*, $C_n = I_n - \mathbf{1}_n \mathbf{1}_n^\top/n$ is the centering matrix, and $\|\cdot\|_2$ here means the spectral norm (which is also known as the operator norm induced by the ℓ_2-norm) of a matrix. As a result,

$$\lim_{m\to\infty} \boldsymbol{h}_m^* = \boldsymbol{v}^*,$$

and

$$\lim_{m\to\infty} -2\|\boldsymbol{h}_m^*\|_1/\gamma_m + \boldsymbol{h}_m^{*\top} Q \boldsymbol{h}_m^* = \varepsilon - \boldsymbol{v}^* C_n K C_n \boldsymbol{v}^*,$$

where $\boldsymbol{h}_m^*$ is the solution to (15) with Q specified as above and $\gamma_m = m$, and $\boldsymbol{v}^*$ is the solution to the relaxed kernel k-means clustering (Ding and He, 2004, Theorem 3.5). In addition, if $\mathcal{X} \subset \mathbb{R}^d$ and $X \in \mathbb{R}^{n\times d}$ is the design matrix, we will have

$$\lim_{m\to\infty} \boldsymbol{h}_m'^* = \boldsymbol{v}'^*,$$

and

$$\lim_{m\to\infty} -2\|\boldsymbol{h}_m'^*\|_1/\gamma_m + \boldsymbol{h}_m'^{*\top} Q \boldsymbol{h}_m'^* = \varepsilon - \boldsymbol{v}'^* C_n X X^\top C_n \boldsymbol{v}'^*,$$

where $\boldsymbol{h}_m'^*$ is the solution to (15) with $Q = \varepsilon I_n - C_n X X^\top C_n$, $\varepsilon > \|C_n X X^\top C_n\|_2$ and $\gamma_m = m$, and $\boldsymbol{v}'^*$ is the solution to the relaxed plain k-means clustering (Ding and He, 2004, Theorem 2.2). In other words, plain k-means clustering and kernel k-means clustering after certain relaxations are special limit cases of MVC-SL.[5]

Similarly to USC and two k-means clustering, the optimization problem of the binary SMIC is another special limit case of MVC-SL. It involves the maximization of the following *squared-loss mutual information* approximator

$$\max_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\in\mathbb{R}^n} \frac{1}{n}\sum_{y=1}^{2} \boldsymbol{\alpha}_y^\top K^2 \boldsymbol{\alpha}_y - \frac{1}{2} \tag{16}$$

5. When considering k-means algorithms that are referred to as certain iterative clustering algorithms rather than clustering models, by no means they can be special limit cases of MVC-SL.

under an orthonormal constraint of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are model parameters of posterior probabilities and K is the kernel matrix. The optimal solutions to (16) can be obtained through

$$\boldsymbol{\alpha}_1^* = \mathop{\arg\max}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \boldsymbol{\alpha}^\top K^2 \boldsymbol{\alpha} \qquad \text{s.t. } \|\boldsymbol{\alpha}\|_2 = 1, \tag{17}$$

$$\boldsymbol{\alpha}_2^* = \mathop{\arg\max}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \boldsymbol{\alpha}^\top K^2 \boldsymbol{\alpha} \qquad \text{s.t. } \boldsymbol{\alpha}\perp\boldsymbol{\alpha}_1^*, \|\boldsymbol{\alpha}\|_2 = 1. \tag{18}$$

Now, let $Q = \varepsilon I_n - K^2$ with arbitrarily chosen $\varepsilon > \|K\|_2^2$. We could then know

$$\lim_{m\to\infty} \boldsymbol{h}_{1,m}^* = \boldsymbol{\alpha}_1^*,$$

and

$$\lim_{m\to\infty} -2\|\boldsymbol{h}_{1,m}^*\|_1/\gamma_m + \boldsymbol{h}_{1,m}^{*\top} Q \boldsymbol{h}_{1,m}^* = \varepsilon - \boldsymbol{\alpha}_1^{*} K^2 \boldsymbol{\alpha}_1^*,$$

where $\boldsymbol{h}_{1,m}^*$ is the solution to (15) with Q specified as above and $\gamma_m = m$. Likewise,

$$\lim_{m\to\infty} \boldsymbol{h}_{2,m}^* = \boldsymbol{\alpha}_2^*,$$

and

$$\lim_{m\to\infty} -2\|\boldsymbol{h}_{2,m}^*\|_1/\gamma_m + \boldsymbol{h}_{2,m}^{*\top} Q \boldsymbol{h}_{2,m}^* = \varepsilon - \boldsymbol{\alpha}_2^{*} K^2 \boldsymbol{\alpha}_2^*,$$

where $\boldsymbol{h}_{2,m}^*$ is the solution to (15) with Q specified as above, $\gamma_m = m$, and a constraint $\boldsymbol{h}^\top \boldsymbol{h}_{1,m}^* = 0$.

Remark 2 After optimizing (17) and (18), SMIC adopts the post-processing that encloses $\boldsymbol{\alpha}_1^*$ and $\boldsymbol{\alpha}_2^*$ into posterior probabilities and enables the out-of-sample clustering ability for any $x \in \mathcal{X}$ even if $x \notin X_n$ (Sugiyama et al., 2011), while MVC-SL can use

$$\boldsymbol{h}^* = \boldsymbol{\alpha}_1^* \operatorname{sign}(\mathbf{1}_n^\top \boldsymbol{\alpha}_1^*) - \boldsymbol{\alpha}_2^* \operatorname{sign}(\mathbf{1}_n^\top \boldsymbol{\alpha}_2^*)$$

as the optimal soft response vector since there are just two clusters.

5. Finite Sample Stability

The stability of the resulting clusters is important for clustering models whose non-convex primal problems are solved by randomized algorithms (e.g., MVC-SL and k-means clustering) rather than relaxed to convex dual problems (e.g., MVC-HL and MMC). To this end, we investigate the finite sample stability of the primal problem of MVC-SL in this section.

In the following, we presume that we are always able to find a locally optimal solution to optimization (4) accurately. Under this presumption, we prove that the instability is resulted from the symmetry of data samples: As long as the input matrix Q satisfies some asymmetry condition, we could obtain the same data partition based on different locally optimal solutions, and consequently the non-convex optimization of MVC-SL seems convex.

5.1 Definitions

Definition 3 *The Hamming clustering distance for two n-dimensional soft response vectors $\boldsymbol{h}$ and $\boldsymbol{h}'$ is defined as*

$$d_{\mathcal{H}}(\boldsymbol{h}, \boldsymbol{h}') := \frac{1}{2} \min(\|\operatorname{sign}(\boldsymbol{h}) + \operatorname{sign}(\boldsymbol{h}')\|_1, \|\operatorname{sign}(\boldsymbol{h}) - \operatorname{sign}(\boldsymbol{h}')\|_1).$$

When measuring the difference of two binary clusterings, $d_{\mathcal{H}}(\boldsymbol{h},\boldsymbol{h}')$ is always a natural number smaller than $n/2$, since $\|\operatorname{sign}(\boldsymbol{h})+\operatorname{sign}(\boldsymbol{h}')\|_1+\|\operatorname{sign}(\boldsymbol{h})-\operatorname{sign}(\boldsymbol{h}')\|_1=2n$.

Definition 4 (Irreducibility) *A sample x_i is isolated in X_n, if $Q_{i,i}>0$ and $\forall j\neq i, Q_{i,j}=0$. A set of samples X_n is irreducible, if no sample is isolated in X_n; otherwise X_n is reducible.*

The idea behind the irreducibility of X_n is simple: If x_i is isolated, we cannot decide its cluster based on the information contained in Q no matter what binary clustering algorithm is used, unless we assign x_i to one cluster and $X_n\setminus x_i$ to the other cluster. We would like to remove such isolated samples and reduce the clustering of X_n to a better-defined problem.

Next we define two symmetry concepts, the submatrix-information- (SI- for short) symmetry in Definition 5 and the axisymmetry in Definition 7. SI-asymmetry is a part of the sufficient condition for finite sample stability, and axisymmetry is a part of the sufficient condition for instability. The relationship of irreducibility, axisymmetry and SI-symmetry will be proved in Theorem 10.

Definition 5 (Submatrix-Information-Symmetry) *A set of samples X_n is submatrix-information-symmetric, if there exist $\{\delta_1,\ldots,\delta_n\}\in\{-1,+1\}^n$ and nonempty $\mathcal{K}\subsetneq\{1,\ldots,n\}$ such that*

$$\sum_{i\in\mathcal{K},j\notin\mathcal{K},\delta_i=\delta_j}Q_{i,j}=\sum_{i\in\mathcal{K},j\notin\mathcal{K},\delta_i\neq\delta_j}Q_{i,j}. \tag{19}$$

Otherwise, X_n is submatrix-information-asymmetric.[6]

Remark 6 It is clear that

$$\sum_{i\in\mathcal{K},j\notin\mathcal{K},\delta_i=\delta_j}Q_{i,j}=\sum_{i\in\mathcal{K},j\notin\mathcal{K}}\delta_i\delta_jQ_{i,j},$$
$$\sum_{i\in\mathcal{K},j\notin\mathcal{K},\delta_i\neq\delta_j}Q_{i,j}=-\sum_{i\in\mathcal{K},j\notin\mathcal{K}}\delta_i\delta_jQ_{i,j},$$

and thus Equation (19) is equivalent to

$$\left(\sum_{k\in\mathcal{K}}\delta_k\boldsymbol{e}_k\right)^{\top}Q\left(\sum_{k\notin\mathcal{K}}\delta_k\boldsymbol{e}_k\right)=0, \tag{20}$$

where $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}$ is a standard basis of $\mathbb{R}^n$. From now on, we may use Equation (20) as the condition to check SI-symmetry or SI-asymmetry for convenience.

Intuitively, the SI-symmetry of X_n says that Q has a submatrix containing the same amount of similarity and dissimilarity information. More specifically, both $\{\delta_1,\ldots,\delta_n\}$ and $\mathcal{K}$ are valid partitions of X_n, though they have different representations and functions. The partition $\{\delta_1,\ldots,\delta_n\}$ is a reference for similarity and dissimilarity, and based on this partition, we categorize the information $Q_{i,j}$ between x_i and x_j into similarity information if $\delta_i=\delta_j$ or dissimilarity information if $\delta_i\neq\delta_j$. On the other hand, we divide Q into four parts $Q[i\in\mathcal{K};j\in\mathcal{K}]$, $Q[i\in\mathcal{K};j\notin\mathcal{K}]$, $Q[i\notin\mathcal{K};j\in\mathcal{K}]$ and $Q[i\notin\mathcal{K};j\notin\mathcal{K}]$ according to the partition $\mathcal{K}$. The SI-symmetry of X_n shown in Equation (19)

6. Strictly speaking, saying that X_n is SI-symmetric is a bit abuse of terminology. In formal mathematical terminology, an object is symmetric with respect to some operation, if this operation, when applied to the object, preserves certain property. For example, in the axisymmetry, the object is X_n, the operation is ϕ and the property is Q. However, in the SI-symmetry, the object is a set of two vectors $\{\sum_{k\in\mathcal{K}}\delta_k\boldsymbol{e}_k,\sum_{k\notin\mathcal{K}}\delta_k\boldsymbol{e}_k\}$, the operation is replacing I_n with Q, and the property is the orthogonality (preserved from the standard orthogonality to the Q-orthogonality).

indicates that the submatrix $Q[i \in \mathcal{K}; j \notin \mathcal{K}]$ (and likewise $Q[i \notin \mathcal{K}; j \in \mathcal{K}]$) contains the same amount of similarity information (i.e., the left-hand side) and dissimilarity information (i.e., the right-hand side).

When X_n is SI-symmetric, we could easily find two feasible solutions to optimization (4), such that they would induce different partitions of X_n but share the same value of the objective function. To see this, let

$$\boldsymbol{h}_+ = \frac{\sum_{k\in\mathcal{K}} \delta_k \boldsymbol{e}_k + \sum_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k}{\sqrt{n}},$$
$$\boldsymbol{h}_- = \frac{\sum_{k\in\mathcal{K}} \delta_k \boldsymbol{e}_k - \sum_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k}{\sqrt{n}}.$$

It is easy to verify that $\|\boldsymbol{h}_\pm\|_2 = 1$, $\|\boldsymbol{h}_\pm\|_1 = \sqrt{n}$ and $d_{\mathcal{H}}(\boldsymbol{h}_+, \boldsymbol{h}_-) \geq 1$. Moreover,

$$\boldsymbol{h}_+^\top Q \boldsymbol{h}_+ = \boldsymbol{h}_+^\top Q \boldsymbol{h}_+ - (\boldsymbol{h}_+ + \boldsymbol{h}_-)^\top Q (\boldsymbol{h}_+ - \boldsymbol{h}_-) = \boldsymbol{h}_-^\top Q \boldsymbol{h}_-,$$

where we used $(\boldsymbol{h}_+ + \boldsymbol{h}_-)^\top Q (\boldsymbol{h}_+ - \boldsymbol{h}_-) = 0$ by the condition Equation (20) of SI-symmetry. However, $\boldsymbol{h}_+$ and $\boldsymbol{h}_-$ are not necessarily locally optimal solutions to optimization (4), and maybe no locally optimal solution could induce the same partition with $\boldsymbol{h}_+$ or $\boldsymbol{h}_-$. The real reason for finite sample instability is the axisymmetry of data samples defined below.

Definition 7 (Axisymmetry) *A set of samples X_n is axisymmetric, if there exists a permutation ϕ of $\{1,\ldots,n\}$, such that*

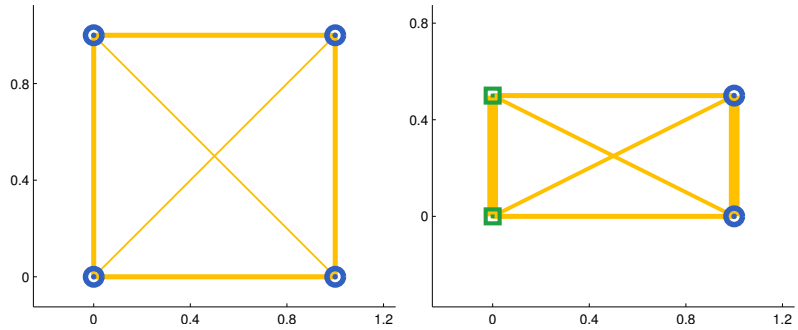
1. $\exists i \in \{1,\ldots,n\}, \phi(i) \neq i$;
2. $\forall i \in \{1,\ldots,n\}, \phi^{-1}(i) = \phi(i)$;
3. $\forall 1 \leq i,j \leq n, Q_{i,j} = Q_{\phi(i),\phi(j)}$.

The first property says that the permutation ϕ cannot be the identical mapping: It allows some, but not all, sample x_i mapped to itself. The second property requires that those samples mapped to others are all paired. In other words, X_n is separated into two types of disjoint subsets according to ϕ, and they are either cardinality one (i.e., $\{x_i \mid \phi(i) = i\}$) or two (i.e., $\{x_i, x_{\phi(i)} \mid \phi(i) \neq i\}$), but no greater cardinality. The third property guarantees that Q is ϕ-invariant, or equivalently the samples in the subset $\{x_i, x_{\phi(i)} \mid \phi(i) \neq i\}$ cannot be distinguished by all other subsets $\{x_j \mid \phi(j) = j, j \neq i\}$ or $\{x_j, x_{\phi(j)} \mid \phi(j) \neq j, j \neq i, j \neq \phi(i)\}$ based on the information contained in Q, so we can exchange x_i and $x_{\phi(i)}$ freely without modifying Q.

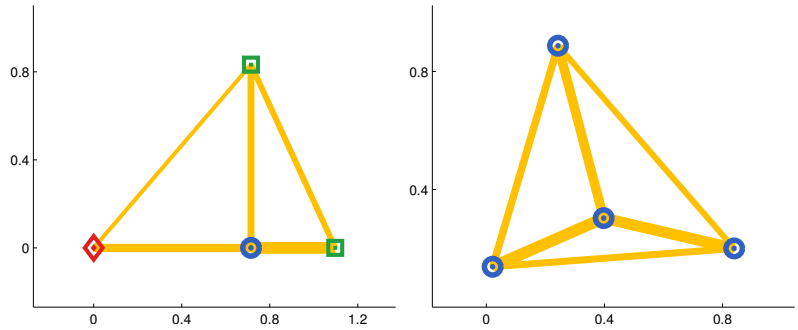
The axisymmetry of X_n in terms of Q is equivalent to the geometric axisymmetry of X_n in $\mathcal{X}$, if $\mathcal{X} \subset \mathbb{R}^d$ and Q is a matrix induced from the Euclidean distance such as a Gaussian kernel matrix. For example, as shown in Figure 2,

$$X_4 = \{(0,0),(1,0),(1,1),(0,1)\},$$
$$X_4' = \{(0,0),(1,0),(1,0.5),(0,0.5)\}$$

are axisymmetric both in $\mathbb{R}^2$ and in terms of Q if Q is a Gaussian kernel matrix, regardless of the kernel width. The permutation ϕ for X_4' could be $\{(1,2),(3,4)\}$, $\{(1,3),(2,4)\}$ or $\{(1,4),(2,3)\}$, and besides them, ϕ for X_4 could also be $\{(1),(3),(2,4)\}$ or $\{(1,3),(2),(4)\}$. We can identify an



(a) X_4 (b) X_4'



(c) X_4'' (d) Random

Figure 2: Four-point sets that are typical in the theory of finite sample stability. Gaussian similarities ($\sigma = 1/\sqrt{2}$) between nodes are visualized by the line thickness of edges. All sets in this figure are irreducible. X_4 in panel (a) is axisymmetric and SI-symmetric. X_4' in (b) is axisymmetric, SI-symmetric, anisotropic, and has a unique best partition. X_4'' in (c) is very special: It is anisotropic, *SI-symmetric but not axisymmetric*, since the similarity of the diamond and circle equals the sum of the similarities of the diamond and squares. A random set would be anisotropic and SI-asymmetric with high probability.

axis of symmetry geometrically in $\mathbb{R}^d$: It must pass through either x_i if $\phi(i) = i$ or $(x_i + x_{\phi(i)})/2$ if $\phi(i) \neq i$ for $i = 1, \ldots, n$. This is why we call such a property axisymmetry.

Generally speaking, the concepts of axisymmetry and SI-symmetry almost coincide, if Q is a Gaussian kernel matrix or the corresponding graph Laplacian matrix. While it is possible to deliberately construct counter-examples that are SI-symmetric but not axisymmetric, it is improbable to meet a counter-example in practice. For instance, as illustrated in panel (c) of Figure 2,

$$\begin{aligned} X_4'' &= \{(0,0), (\sqrt{\ln(5/3)}, 0), (\sqrt{\ln(10/3)}, 0), (\sqrt{\ln(5/3)}, \sqrt{\ln 2})\} \\ &\approx \{(0,0), (0.7147, 0), (1.0973, 0), (0.7147, 0.8326)\} \end{aligned}$$

is SI-symmetric but not axisymmetric in terms of Gaussian kernel matrix Q when $\sigma = 1/\sqrt{2}$, yet X_4'' is SI-asymmetric whenever $\sigma \neq 1/\sqrt{2}$.

Definition 8 (Anisotropy) *A set of samples X_n is anisotropic, if Q has n distinct eigenvalues.*

The anisotropy of X_n is the other part of the sufficient condition for finite sample stability. The name comes from a geometric interpretation of the ellipsoid $\mathcal{E}(\mathcal{H}_Q)$: All its principal axes achieve distinct lengths when Q has distinct eigenvalues, and thus $\mathcal{E}(\mathcal{H}_Q)$ is anisotropic and not rotatable. The concepts of anisotropy and axisymmetry are not complementary, since they concern different aspects of different objects, that is, the rotation of $\mathcal{E}(\mathcal{H}_Q)$ vs. the reflection of X_n. In Figure 2, X_4 is axisymmetric, X_4' is axisymmetric and anisotropic, and most random sets are just anisotropic. There might be X_n neither axisymmetric nor anisotropic. Nonetheless, when considering the more general SI-symmetry and certain families of Q such as Gaussian kernel matrices, X_n is anisotropic as long as it is SI-asymmetric.

All definitions have been discussed. The theoretical results will be presented next.

5.2 Theoretical Results

The following lemma will be used in Theorems 11 and 13. All proofs are provided in Appendix A.

Lemma 9 *Let X_n be an irreducible set of samples, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the normalized eigenvectors of Q, and $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ be a standard basis of $\mathbb{R}^n$. Then, $\forall i, j \in \{1, \ldots, n\}$, $\boldsymbol{v}_i \neq \pm \boldsymbol{e}_j$.*

The following two theorems describe the relationship between the properties defined above.

Theorem 10 *A set of samples X_n is SI-symmetric, if it is reducible or axisymmetric.*

Theorem 11 *If X_n is an SI-asymmetric set of samples, and there exists $\kappa > 0$ such that $Q_{1,1} = Q_{2,2} = \cdots = Q_{n,n} = \kappa$, then X_n is anisotropic.*

We are ready to deliver our main theorems. To begin with, given a constant η, we define (recall the assumption that $\|\boldsymbol{h}\|_1$ is differentiable thanks to the non-sparsity of $\boldsymbol{h}$)

$$G(\boldsymbol{h}) := \gamma \boldsymbol{h}^\top Q \boldsymbol{h} - \eta \|\boldsymbol{h}\|_2^2 - 2\|\boldsymbol{h}\|_1,$$
$$g(\boldsymbol{h}) := \frac{1}{2} \nabla G(\boldsymbol{h}) = \gamma Q \boldsymbol{h} - \eta \boldsymbol{h} - \operatorname{sign}(\boldsymbol{h}).$$

Theorem 12 (Twin Minimum Theorem) *Assume that $n > 2$, X_n is an axisymmetric set of samples, ϕ is the corresponding permutation, and $I = \{\{i, \phi(i)\} \mid \phi(i) \neq i\}$ is the index set of those paired samples given ϕ. For every minimum $\boldsymbol{h}^*$ of optimization* (4)*, if*

1. *$\forall i, [\boldsymbol{h}^*]_i \neq 0$, and*
2. *$\exists \{i, \phi(i)\} \in I, [\boldsymbol{h}^*]_{\phi(i)} [\boldsymbol{h}^*]_i < 0$,*

then $\boldsymbol{h}^*$ has a twin minimum $\boldsymbol{h}^\star$ satisfying $G(\boldsymbol{h}^\star) = G(\boldsymbol{h}^*)$ and $d_{\mathcal{H}}(\boldsymbol{h}^*, \boldsymbol{h}^\star) \geq 1$. The only exception is

$$\forall i \in \{1, \ldots, n\}, [\boldsymbol{h}^*]_{\phi(i)} [\boldsymbol{h}^*]_i < 0.$$

In order to explain the implication of Theorem 12, let us recall X_4 and X_4' shown in Figure 2. There are many twin minima when considering the perfectly symmetric X_4, but it is same even for those convex relaxations of MMC due to the post-processing. On the other hand, X_4' illustrates an exception: While X_4 allows $\phi(i) = i$, this is impossible for X_4'. More specifically, any minimum $\boldsymbol{h}^*$

corresponding to partition $(+1,-1,-1,+1)$ has no twin minimum, since ϕ can be $\{(1,2),(3,4)\}$, $\{(1,3),(2,4)\}$ or $\{(1,4),(2,3)\}$ for X_4', and

$$\forall \phi, (\exists i, [\boldsymbol{h}^*]_{\phi(i)}[\boldsymbol{h}^*]_i < 0) \rightarrow (\forall i, [\boldsymbol{h}^*]_{\phi(i)}[\boldsymbol{h}^*]_i < 0).$$

It suggests that if we permute $\boldsymbol{h}^*$ according to ϕ, then $\text{sign}(\boldsymbol{h}^\star) = \pm\text{sign}(\boldsymbol{h}^*)$ is the same partition and thus $d_{\mathcal{H}}(\boldsymbol{h}^*, \boldsymbol{h}^\star) = 0$. Another minimum $\boldsymbol{h}$ that corresponds to $(+1,+1,-1,-1)$ and satisfies $d_{\mathcal{H}}(\boldsymbol{h}^*, \boldsymbol{h}) \geq 1$ should have $G(\boldsymbol{h}) > G(\boldsymbol{h}^*)$. In a word, local minima that correspond to different partitions for X_4' are not equally good and the best partition is still unique, as illustrated in panel (b) of Figure 2. The genuine instability emerges only when the best partition is not unique, like X_4 in panel (a) of Figure 2.

Theorem 13 (Equivalent Minima Theorem) *All minima of optimization* (4) *are equivalent with respect to $d_{\mathcal{H}}$, provided that*

1. *X_n is SI-asymmetric;*
2. *X_n is anisotropic.*

By combining Theorem 11 and Theorem 13, we have a corollary immediately.

Corollary 14 *All minima of optimization* (4) *are equivalent with respect to $d_{\mathcal{H}}$, provided that*

1. *X_n is SI-asymmetric;*
2. *There exists $\kappa > 0$ such that $Q_{1,1} = Q_{2,2} = \cdots = Q_{n,n} = \kappa$.*

To sum up, if Q has the two properties listed above, different locally optimal solutions to optimization (4) would ideally induce the same data partition. Nevertheless, the output of the algorithm is not in the same form as the solution to optimization (4), since the variable η has been introduced, and we cannot foresee its optimal value when we analyze the original model. Spectral clustering is consistent (von Luxburg et al., 2008), but it has a similar problem in finite sample stability, that is, when the graph Laplacian has distinct eigenvalues and the unique spectral decomposition leads to a stable spectral embedding, the following k-means clustering can still introduce high instability due to the non-convex nature of the distortion function.

Remark 15 We rely on a Karush-Kuhn-Tucker stationarity condition $g(\boldsymbol{h}^*) = \mathbf{0}_n$ in the proofs of Theorems 12 and 13, where $\boldsymbol{h}^*$ is the optimal solution to (4). Actually, the objective function of (4) usually has a non-zero derivative and the objective function of (3) always has a non-zero derivative in their feasible regions. Therefore, we introduce the functions $G(\boldsymbol{h})$ and $g(\boldsymbol{h})$ to analyze MVC-SL from a theoretical point of view. In MVC-SL, Equation (7) comes from the least-square fitting of Equation (8), and if $t \rightarrow \infty$, we will have $\boldsymbol{p}_t^* \rightarrow \mathbf{0}_n$ and then Equation (8) will turn into $g(\boldsymbol{h}^*) = \lim_{t\rightarrow\infty} g(\boldsymbol{h}_t) = \mathbf{0}_n$.

6. Clustering Error Bound

In this section, we derive a clustering error bound for MVC-SL based on *transductive Rademacher complexity* (El-Yaniv and Pechyony, 2009).

It is extremely difficult, if possible, to evaluate clustering methods in an objective and domain-independent manner (von Luxburg et al., 2012). However, when the goals and interests are clear, it

makes sense to evaluate clustering results using classification benchmark data sets, where the class structure coincides with the desired cluster structure according to the goals and interests.

In real-world applications, we often find some experts to cluster a small portion $X_{n'}$ of X_n where $n' < n$ according to their professional knowledge, test a lot of clustering methods with a lot of similarity measures, see their agreement with given clustering of $X_{n'}$, and eliminate those low agreement methods. This procedure may be viewed as propagating the knowledge of experts from $X_{n'}$ to X_n.

Here, we derive a data-dependent clustering error bound to guarantee the quality of this propagation of knowledge. The key technique is transductive Rademacher complexity for deriving data-dependent transductive error bounds. To begin with, we follow El-Yaniv and Pechyony (2009) for the definition of transductive Rademacher complexity:

Definition 16 *Fix positive integers m and u. Let $\mathcal{H} \subseteq \mathbb{R}^{m+u}$ be a hypothesis space, $p \in [0, 1/2]$ be a parameter, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{m+u})^\top$ be a vector of independent and identically distributed random variables, such that*

$$\sigma_i := \begin{cases} +1 & \text{with probability} \quad p, \\ -1 & \text{with probability} \quad p, \\ 0 & \text{with probability} \quad 1-2p. \end{cases}$$

Then, the transductive Rademacher complexity of $\mathcal{H}$ with parameter p is defined as

$$\mathcal{R}_{m+u}(\mathcal{H}, p) := \left(\frac{1}{m} + \frac{1}{u}\right) \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\boldsymbol{h} \in \mathcal{H}} \boldsymbol{\sigma}^\top \boldsymbol{h} \right\}.$$

For the sake of comparison, we give a definition of inductive Rademacher complexity following El-Yaniv and Pechyony (2009).[7]

Definition 17 *Let $p(x)$ be a probability density on $\mathcal{X}$, and $X_n = \{x_1, \ldots, x_n\}$ be a set of independent observations drawn from $p(x)$. Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top$ be a vector of independent and identically distributed random variables, such that*

$$\sigma_i := \begin{cases} +1 & \text{with probability} \quad 1/2, \\ -1 & \text{with probability} \quad 1/2. \end{cases}$$

The empirical Rademacher complexity of $\mathcal{H}$ conditioned on X_n is

$$\widehat{\mathcal{R}}_u^{(ind)}(\mathcal{H}) := \frac{2}{n} \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{h \in \mathcal{H}} \boldsymbol{\sigma}^\top \boldsymbol{h} \mid X_n \right\},$$

where $\boldsymbol{h} = (h(x_1), \ldots, h(x_n))^\top$, and the inductive Rademacher complexity of $\mathcal{H}$ is

$$\mathcal{R}_u^{(ind)}(\mathcal{H}) := \mathbb{E}_{X_n} \left\{ \widehat{\mathcal{R}}_u^{(ind)}(\mathcal{H}) \right\}.$$

The transductive Rademacher complexity of $\mathcal{H}$ is an empirical quantity that depends only on p. Given fixed X_n, we have $\mathcal{R}_{m+u}(\mathcal{H}) = 2\widehat{\mathcal{R}}_{m+u}^{(ind)}(\mathcal{H})$ when $p = 1/2$ and $m = u$.[8] Whenever $p < 1/2$,

7. Albeit there are many definitions of Rademacher complexity, for example, Koltchinskii (2001), Bartlett and Mendelson (2002), Meir and Zhang (2003) and Bousquet et al. (2004), they are similar and conceptually equivalent.
8. A class of functions conditioned on fixed data is equivalent to a hypothesis space of soft response vectors.

some Rademacher variables will attain zero values and reduce the complexity. We simply consider $p_0 = mu/(m+u)^2$ and abbreviate $\mathcal{R}_{m+u}(\mathcal{H}, p_0)$ to $\mathcal{R}_{m+u}(\mathcal{H})$ as El-Yaniv and Pechyony (2009) in Lemma 18 and Theorem 19, though these theoretical results hold for all $p > p_0$ since $\mathcal{R}_{m+u}(\mathcal{H}, p)$ is monotonically increasing with p. Please refer to El-Yaniv and Pechyony (2009) for the detailed discussions about transductive Rademacher complexity.

Lemma 18 *Let $\mathcal{H}'_Q$ be the set of all possible $\boldsymbol{h}$ returned by Algorithm 1 for the given Q, η^* be the optimal η when Algorithm 1 stops,*

$$\mu = \sup_{\boldsymbol{h} \in \mathcal{H}'_Q} \operatorname{sign}(\boldsymbol{h})^\top (\gamma Q - \eta^* I_n)^{-1} \operatorname{sign}(\boldsymbol{h}),$$

and $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of Q. Then, for the transductive Rademacher complexity of $\mathcal{H}'_Q$, the following upper bound holds for any integer $n' = 1, 2, \ldots, n-1$,

$$\mathcal{R}_u(\mathcal{H}'_Q) \leq \sqrt{\frac{2}{n'(n-n')}} \min \left\{ \sqrt{n}, \left(\sum_{i=1}^{n} \frac{n}{(\gamma\lambda_i - \eta^*)^2} \right)^{1/2}, \left(\sum_{i=1}^{n} \frac{\mu}{\gamma\lambda_i - \eta^*} \right)^{1/2} \right\}.$$

The proof of Lemma 18 can be found in Appendix B. By Lemma 18 together with Theorem 2 of El-Yaniv and Pechyony (2009), we can immediately obtain the clustering error bound:

Theorem 19 *Assume that $\boldsymbol{y}^*$ is the ground truth partition of X_n, and $\mathcal{L}$ is a random set of size n' chosen uniformly from the set $\{\mathcal{L} \mid \mathcal{L} \subset \{1, \ldots, n\}, \#\mathcal{L} = n'\}$. Let $\ell(z) = \min(1, \max(0, 1-z))$ for $z \in \mathbb{R}$ be the surrogate loss, $\mathcal{H}'_Q$ be the set of all possible $\boldsymbol{h}$ returned by Algorithm 1 for the given Q, η^* be the optimal η when Algorithm 1 stops,*

$$\mu = \sup_{\boldsymbol{h} \in \mathcal{H}'_Q} \operatorname{sign}(\boldsymbol{h})^\top (\gamma Q - \eta^* I_n)^{-1} \operatorname{sign}(\boldsymbol{h}),$$

$\lambda_1, \ldots, \lambda_n$ be the eigenvalues of Q, and $c_0 = \sqrt{32(1+\ln 4)/3}$. For any $\boldsymbol{h} \in \mathcal{H}'_Q$, with probability at least $1-\delta$ over the choice of $\mathcal{L}$, we have

$$\begin{aligned} d_{\mathcal{H}}(\boldsymbol{h}, \boldsymbol{y}^*) \leq{}& \frac{n}{n'} \min \left\{ \sum_{i \in \mathcal{L}} \ell([\boldsymbol{h}]_i [\boldsymbol{y}^*]_i), \sum_{i \in \mathcal{L}} \ell(-[\boldsymbol{h}]_i [\boldsymbol{y}^*]_i) \right\} \\ &+ \frac{c_0 n}{\sqrt{n'}} + \sqrt{\frac{2n^2(n-n')^2}{n'(2n-1)(2n-2n'-1)} \ln(1/\delta)} \\ &+ \sqrt{\frac{2(n-n')}{n'}} \min \left\{ \sqrt{n}, \left(\sum_{i=1}^{n} \frac{n}{(\gamma\lambda_i - \eta^*)^2} \right)^{1/2}, \left(\sum_{i=1}^{n} \frac{\mu}{\gamma\lambda_i - \eta^*} \right)^{1/2} \right\}. \end{aligned} \tag{21}$$

There are four terms in the right-hand side of inequality (21). The first term is a measure of the clustering error on $X_{n'} = \{x_i \mid i \in \mathcal{L}\}$ by the surrogate loss times the ratio n/n'. More specifically, we would like to select a proper similarity measure via given clustering $\{[\boldsymbol{y}^*]_i \mid i \in \mathcal{L}\}$ to make the error on $X_{n'}$ as small as possible, under the assumption that the error rates on $X_{n'}$ and X_n should be close for a fixed similarity measure (the given $\{[\boldsymbol{y}^*]_i \mid i \in \mathcal{L}\}$ are not used for training). The second term depends only on n and n', i.e., the sizes of the whole set and the clustered subset. Besides n and n', the third term further depends on the significance level δ, as in common error bounds. The last term

is the upper bound of $(n-n')\mathcal{R}_u(\mathcal{H}'_Q)$, which carries out the complexity control of $\mathcal{H}'_Q$ implicitly: The smaller the value of $\mathcal{R}_u(\mathcal{H}'_Q)$ is, the more confident we are that $d_{\mathcal{H}}(\boldsymbol{h},\boldsymbol{y}^*)$ would be small, if the error on $X_{n'}$ is small. When considering the average clustering error measured by $d_{\mathcal{H}}(\boldsymbol{h},\boldsymbol{y}^*)/n$, the order of the error bound is $O(1/\sqrt{n'})$ since the second term dominates the third and fourth terms.

Remark 20 Our problem setting is equivalent to neither semi-supervised clustering nor transductive classification: We do not reveal any labels to the clustering algorithm in Theorem 19; instead, a set of randomly chosen labels are revealed to an evaluator who then returns the evaluation of the quality of any possible partition generated by the algorithm. We can use the theory of transductive Rademacher complexity to derive a clustering error bound for Algorithm 1, since it can be viewed as a transductive algorithm that ignores all revealed labels.

7. Related Works

In this section, we review related works and qualitatively compare the proposed MVC with them.

7.1 Maximum Margin Clustering

Among existing clustering methods, *maximum margin clustering* (MMC) is closest to MVC. Both of them originate from statistical learning theory, but their geneses and underlying criteria are still different: The primal problems of all MMC adopt a regularizer $\|\boldsymbol{w}\|_2^2$ from the margin, while MVC relies on the regularizer $V(\boldsymbol{h})$ in Equation (1) from the volume. The hypothesis shared by all MMC is the hyperplane for induction, while the hypothesis in MVC is the soft response vector for transduction. The latter is more natural, since clustering is more transductive than inductive.

The family of MMC algorithms was initiated by Xu et al. (2005). It follows the support vector machine (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) and its hard-margin version can be formulated as

$$\min_{\boldsymbol{y}\in\{-1,+1\}^n} \min_{\boldsymbol{w},\boldsymbol{\xi}} \ \|\boldsymbol{w}\|_2^2 \\ \text{s.t. } y_i\boldsymbol{w}^\top x_i \geq 1, \quad i=1,\ldots,n.$$

The value of $y_i\boldsymbol{w}^\top x_i$ is called the functional margin of (x_i,y_i), whereas the value of $y_i\boldsymbol{w}^\top x_i/\|\boldsymbol{w}\|_2$ is called the geometric margin of (x_i,y_i). MMC can maximize the geometric margin of all $x_i \in X_n$ over $\boldsymbol{y}\in\{-1,+1\}^n$ by minimizing $\|\boldsymbol{w}\|_2$ and requiring the minimal functional margin to be one simultaneously. Likewise, the primal problem of the soft-margin MMC is

$$\min_{\boldsymbol{y}\in\{-1,+1\}^n} \min_{\boldsymbol{w},\boldsymbol{\xi}} \ \|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{s.t. } y_i\boldsymbol{w}^\top x_i \geq 1-\xi_i, \xi_i \geq 0, \quad i=1,\ldots,n,$$

where $C > 0$ is a regularization parameter, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is a vector of slack variables. Then, it can be relaxed into a standard SDP dual

$$\begin{aligned} \min_{M,\boldsymbol{\mu},\boldsymbol{\nu},t} \quad & t \\ \text{s.t.} \quad & M \succeq \mathbf{0} \\ & \operatorname{diag}(M) = \mathbf{1}_n \\ & \boldsymbol{\mu} \geq \mathbf{0}_n, \boldsymbol{\nu} \geq \mathbf{0}_n \\ & \begin{pmatrix} M \circ K & (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu}) \\ (\mathbf{1}_n - \boldsymbol{\mu} + \boldsymbol{\nu})^\top & t - 2C\boldsymbol{\mu}^\top \mathbf{1}_n \end{pmatrix} \succeq \mathbf{0}, \end{aligned} \tag{22}$$

and solved by any standard SDP solver in $O(n^{6.5})$ time.

Remark 21 Xu et al. (2005) initially imposed three groups of linear constraints on the entries of M in MMC:

1. $\forall ijk$, $M_{i,k} \geq M_{i,j} + M_{j,k} - 1$;
2. $\forall ijk$, $M_{i,k} \geq -M_{i,j} - M_{j,k} - 1$;
3. $\forall i$, $-b \leq \sum_j M_{i,j} \leq b$.

However, Xu and Schuurmans (2005) and Valizadegan and Jin (2007) considered (22) as the dual problem of MMC, sometimes equipped with an additional class balance constraint $-b\mathbf{1}_n \leq M\mathbf{1}_n \leq b\mathbf{1}_n$. In other words, the first and second groups of constraints were ignored.

Subsequently, *generalized maximum margin clustering* (GMMC) (Valizadegan and Jin, 2007) relaxes the restriction that the original MMC only considers homogeneous hyperplanes and hence demands every possible clustering boundary to pass through the origin. Furthermore, GMMC is a convex relaxation of MMC, and its computational complexity is $O(n^{4.5})$ that is remarkably faster than MMC. In fact, GMMC optimizes an n-dimensional vector rather than an $n \times n$ matrix. More specifically, the hard-margin GMMC converts the original MMC following Lanckriet et al. (2004) into a dual problem as

$$\begin{aligned} \min_{\boldsymbol{y} \in \{-1,+1\}^n} \min_{\boldsymbol{\nu},\lambda} \quad & \frac{1}{2}(\mathbf{1}_n + \boldsymbol{\nu} + \lambda \boldsymbol{y})^\top \operatorname{diag}(\boldsymbol{y}) K^{-1} \operatorname{diag}(\boldsymbol{y})(\mathbf{1}_n + \boldsymbol{\nu} + \lambda \boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{\nu} \geq \mathbf{0}_n, \end{aligned}$$

where the function $\operatorname{diag}(\cdot)$ here converts a column vector into a diagonal matrix. The trick here is

$$\left(K \circ \boldsymbol{y}\boldsymbol{y}^\top\right)^{-1} = (\operatorname{diag}(\boldsymbol{y}) K \operatorname{diag}(\boldsymbol{y}))^{-1} = \operatorname{diag}(\boldsymbol{y}) K^{-1} \operatorname{diag}(\boldsymbol{y}),$$

since $\boldsymbol{y} \in \{-1,+1\}^n$. By a tricky substitution $\boldsymbol{w} = (\operatorname{diag}(\boldsymbol{y})(\mathbf{1}_n + \boldsymbol{\nu}); \lambda) \in \mathbb{R}^{n+1}$ where we use the semicolon to separate the rows of a vector or matrix (i.e., $(A;B) = (A^\top, B^\top)^\top$), it becomes

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{n+1}} \quad & \boldsymbol{w}^\top (I_n; \mathbf{1}_n^\top) K^{-1} (I_n, \mathbf{1}_n) \boldsymbol{w} + C_e \left((\mathbf{1}_n^\top, 0) \boldsymbol{w} \right)^2 \\ \text{s.t.} \quad & [\boldsymbol{w}]_i^2 \geq 1, i = 1, \ldots, n, \end{aligned} \tag{23}$$

where $((\mathbf{1}_n^\top, 0)\boldsymbol{w})^2$ is another regularization to remove the translation invariance from the objective function and C_e is the corresponding regularization parameter. Let

$$W = (I_n; \mathbf{1}_n^\top) K^{-1} (I_n, \mathbf{1}_n) + C_e (\mathbf{1}_n; 0)(\mathbf{1}_n^\top, 0) - \mathrm{diag}((\boldsymbol{\gamma}; 0)).$$

The SDP dual of optimization (23) is then

$$\max_{\boldsymbol{\gamma} \in \mathbb{R}^n} \boldsymbol{\gamma}^\top \mathbf{1}_n \qquad \text{s.t. } W \succeq \mathbf{0}, \boldsymbol{\gamma} \geq \mathbf{0}_n.$$

This is the dual problem of the hard-margin GMMC. The dual problem of the soft-margin GMMC is slightly different such that $\boldsymbol{\gamma}$ is upper bounded:

$$\max_{\boldsymbol{\gamma} \in \mathbb{R}^n} \boldsymbol{\gamma}^\top \mathbf{1}_n \qquad \text{s.t. } W \succeq \mathbf{0}, \mathbf{0}_n \leq \boldsymbol{\gamma} \leq C_\delta \mathbf{1}_n, \tag{24}$$

where C_δ is a regularization parameter to control the trade-off between the clustering error and the margin. After obtaining the optimal $\boldsymbol{\gamma}$, the partition can be inferred from the sign of the eigenvector of W associated with the zero eigenvalue, since the Karush-Kuhn-Tucker complementary condition is $W\boldsymbol{w} = \mathbf{0}_{n+1}$, and $\mathrm{sign}([\boldsymbol{w}]_i) = \mathrm{sign}([\boldsymbol{y}]_i)$ for $i = 1, \ldots, n$.

There exist a few faster MMC algorithms. *Iterative support vector regression* (IterSVR) (Zhang et al., 2007) replaces SVM with the hinge loss in the inner optimization subproblem with SVR with the Laplacian loss, while for each inner SVR the time complexity is at most $O(n^3)$ and the empirical time complexity is usually between $O(n)$ and $O(n^{2.3})$. *Cutting-plane maximum margin clustering* (CPMMC) (Zhao et al., 2008b) can be solved by a series of constrained concave-convex procedures within a linear time complexity $O(sn)$ where s is the average number of non-zero features. Unlike MMC and GMMC that rely on SDP or IterSVR and CPMMC that are non-convex, *label-generation maximum margin clustering* (LGMMC) (Li et al., 2009) is scalable yet convex so that it can achieve its globally optimal solution. Roughly speaking, LGMMC replaces the hinge loss in SVM with the squared hinge loss to get an alternative MMC:

$$\begin{aligned} \min_{\boldsymbol{y} \in \{-1,+1\}^n} \min_{\boldsymbol{w}, \boldsymbol{\xi}} \; & \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \rho + \frac{C}{2}\sum_{i=1}^n \xi_i^2 \\ \text{s.t. } & y_i \boldsymbol{w}^\top x_i \geq \rho - \xi_i, \quad i = 1, \ldots, n \\ & -b \leq \boldsymbol{y}^\top \mathbf{1}_n \leq b. \end{aligned}$$

After a long derivation, LGMMC can be expressed as a *multiple kernel learning* problem:

$$\begin{aligned} \min_{\boldsymbol{\mu} \in \mathbb{R}^{2^n}} \max_{\boldsymbol{\alpha}} \; & -\frac{1}{2}\boldsymbol{\alpha}^\top \left(\sum\nolimits_{t: -b \leq \boldsymbol{y}_t^\top \mathbf{1}_n \leq b} \mu_t K \circ \boldsymbol{y}_t \boldsymbol{y}_t^\top + \frac{1}{C} I_n \right) \boldsymbol{\alpha} \\ \text{s.t. } & \boldsymbol{\mu}^\top \mathbf{1}_{2^n} = 1, \boldsymbol{\mu} \geq \mathbf{0}_{2^n} \\ & \boldsymbol{\alpha}^\top \mathbf{1}_n = 1, \boldsymbol{\alpha} \geq \mathbf{0}_n. \end{aligned}$$

This optimization is again solved by the cutting plane method, that is, finding the most violated $\boldsymbol{y}_t$ iteratively, and the empirical time complexity of multiple kernel learning has the same order as the complexity of SVM which usually scales between $O(n)$ and $O(n^{2.3})$.

On the other hand, the stability of MVC is by no means inferior to those non-convex MMC in terms of the resulting clusters. The optimization involved in MVC-HL is a convex SDP problem;

the optimization involved in MVC-SL is a non-convex SQP problem, while under mild conditions, it seems convex if one only cares the resulting clusters. Moreover, MVC-SL has a clustering error bound, and to the best of our knowledge no MMC has such a result. Although the asymptotic time complexity of MVC-SL is $O(n^3)$, its computation time has exhibited less potential of growth in our experiments than the computationally-efficient LGMMC (see Figure 5 in page 2669).

7.2 Spectral Clustering

Spectral clustering (SC) (Shi and Malik, 2000; Meila and Shi, 2001; Ng et al., 2002) is also closely related to MVC. SC algorithms include two steps, a spectral embedding step to unfold the manifold structure and embed the input data into a low-dimensional space in a geodesic manner, and then a k-means clustering step to carry out clustering using the embedded data.

Given a similarity matrix $W \in \mathbb{R}^{n\times n}$ and the corresponding degree matrix $D = \mathrm{diag}(W\mathbf{1}_n)$, we have three popular graph Laplacian matrices: The unnormalized graph Laplacian is defined as

$$L_{\mathrm{un}} := D - W,$$

and two normalized graph Laplacian are

$$L_{\mathrm{sym}} := D^{-1/2} L_{\mathrm{un}} D^{-1/2} = I_n - D^{-1/2} W D^{-1/2}$$
$$L_{\mathrm{rw}} := D^{-1} L_{\mathrm{un}} = I_n - D^{-1} W.$$

The first matrix is denoted by L_{sym} since it is a symmetric matrix and the second one by L_{rw} since it is closely related to a random walk. Each popular graph Laplacian corresponds to a popular SC algorithm according to von Luxburg (2007). Unnormalized SC computes the first k eigenvectors of L_{un} where the eigenvalues are all positive and listed in an increasing order. Shi and Malik (2000) computes the first k generalized eigenvectors of the generalized eigenvalue problem $L_{\mathrm{un}}\boldsymbol{u} = \lambda D\boldsymbol{u}$ that are also the eigenvectors of L_{rw}, and hence it is called normalized SC.[9] The other normalized SC algorithm, namely Ng et al. (2002), computes the first k eigenvectors of L_{sym}, puts them into an $n \times k$ matrix, and normalizes all rows of that matrix to the unit norm, that is, projects the embedded data further to the k-dimensional unit sphere. Anyway, the main idea is to change the representation from $\mathbb{R}^d$ to $\mathbb{R}^k$ and then run k-means clustering.

MVC-SL is able to integrate the two steps of unnormalized SC into a single optimization when the number of clusters is two and the highly non-convex k-means step is unnecessary. Furthermore, a vital difference between MVC and SC is that the basic model of MVC has a loss function which pushes hypotheses away from the coordinate axes and always leads to non-sparse optimal solutions. When considering the finite sample stability, the spectral embedding step of SC is stable if MVC-SL is stable but not vice versa, since SC only requires that the graph Laplacian has distinct eigenvalues; the k-means step is always unstable for fixed data due to the non-convex distortion function which is essentially an integer programming, but it is stable for different random samplings from the same underlying distribution, if the globally optimal solution is unique (Rakhlin and Caponnetto, 2007). In addition, there are a few theoretical results about the infinite sample stability or the consistency of SC. Globally optimal solutions to k-means clustering converge to a limit partition of the whole data space $\mathcal{X}$, if the underlying distribution has a finite support, and the globally optimal solution

9. Actually, two algorithms were proposed in Shi and Malik (2000): The two-way cut algorithm only makes use of the second eigenvector and the k-way cut algorithm uses all first k eigenvectors.

to the expectation of the distortion function with respect to the underlying distribution is unique (Ben-David et al., 2007). Eigenvectors of graph Laplacian also converge to eigenvectors of certain limit operators, while the conditions for convergence are very general for L_{sym}, but are very special for L_{un} so that they are not easily satisfied (von Luxburg et al., 2005, 2008). In contrast, the infinite sample stability of MVC is currently an open problem.

Remark 22 Certain SC algorithms such as Belkin and Niyogi (2002) ignore the first eigenvector by extracting the second to k-th eigenvectors of some graph Laplacian, and thus change the representation to $\mathbb{R}^{k-1}$ rather than $\mathbb{R}^k$. Nevertheless, the multiplicity of the eigenvalue zero of the graph Laplacian equals the number of connected components of the similarity graph, and the eigenspace of eigenvalue zero is spanned by the indicator vectors of the connected components (von Luxburg, 2007, Propositions 2 and 4). As a consequence, all three aforementioned SC algorithms keep the first eigenvector in order to deal with disconnected similarity graphs.

7.3 Approximate Volume Regularization

The connection of *approximate volume regularization* (AVR) (El-Yaniv et al., 2008) and MVC is analogous with the connection of SVM and MMC.

Compared with MVC, AVR is a transductive method for classification so that the label vector $\boldsymbol{y}$ is constant and only the soft response vector $\boldsymbol{h}$ needs to be optimized. More specifically, given m labeled data $\{(x_1,y_1),\ldots,(x_m,y_m)\}$ and u unlabeled data $\{x_{m+1},\ldots,x_{m+u}\}$, the label vector is denoted by $\boldsymbol{y}=(y_1,\ldots,y_m,0,\ldots,0)^\top\in\mathbb{R}^{m+u}$, and the primal problem of AVR is defined as

$$\min_{\boldsymbol{h}\in\mathbb{R}^{m+u}} -\frac{1}{m}\boldsymbol{h}^\top\boldsymbol{y}+\gamma\boldsymbol{h}^\top Q\boldsymbol{h} \qquad \text{s.t. } \|\boldsymbol{h}\|_2=t, \tag{25}$$

where t is a hyperparameter to control the scale of $\boldsymbol{h}$. Since $\boldsymbol{y}$ is constant, optimization (25) can be directly solved using Lagrangian multipliers and the Karush-Kuhn-Tucker conditions

$$\begin{aligned} -\boldsymbol{y}/m+2\gamma Q\boldsymbol{h}-2\eta\boldsymbol{h}&=0,\\ \boldsymbol{h}^\top\boldsymbol{h}-t^2&=0. \end{aligned}$$

Let the eigen-decomposition of Q be $Q=V\Lambda V^\top$ and $d_i=[V^\top\boldsymbol{y}]_i$, then we get an equation about the optimal η:

$$\frac{1}{4m^2}\sum_{i=1}^{m+u}\frac{d_i^2}{(\gamma\lambda_i-\eta)^2}-t^2=0. \tag{26}$$

Thanks to the special structure of (26), a binary search procedure is enough for finding its smallest root η^*, and the optimal $\boldsymbol{h}$ is recovered by

$$\boldsymbol{h}^*=\frac{1}{2m}(\gamma Q-\eta^* I_{m+u})^{-1}\boldsymbol{y}.$$

On the other hand, MVC involves a combinatorial optimization similarly to the most clustering models and several semi-supervised learning models such as MMC. This difficulty caused by the integer feasible region is intrinsically owing to the clustering problem and has no business with the large volume approximation $V(\boldsymbol{h})$. In order to solve the basic model, we proposed two approximation schemes based on sequential quadratic programming and semi-definite programming that are much more complicated than finding the smallest root of Equation (26) as in AVR.

8. Experiments

In this section, we numerically evaluate the performance of the proposed MVC algorithms.

8.1 Setup

Seven clustering algorithms were included in our experiments:

- Kernel k-means clustering (KM; Zha et al., 2002),
- Normalized spectral clustering (NSC; Ng et al., 2002),
- Maximum margin clustering (MMC; Xu et al., 2005),
- Generalized MMC (GMMC; Valizadegan and Jin, 2007),
- Label-generation MMC (LGMMC; Li et al., 2009),
- Soft-label maximum volume clustering (MVC-SL),
- Hard-label maximum volume clustering (MVC-HL).

The *CVX* package (Grant and Boyd, 2011), which is a Matlab-based modeling system for disciplined convex programming, was used to solve the QP problem (5) for MVC-SL and the SDP problems (13), (22) and (24) for MVC-HL, MMC and GMMC.

Table 1 summarizes the specification of data sets in our experiments. We first evaluated all seven algorithms on three artificial data sets. MVC-HL and MMC were excluded from the middle-scale experiments since they were very time-consuming when $n > 100$. The *IDA benchmark repository*[10] contains thirteen benchmark data sets for binary classification, and ten of them that have no intrinsic within-class multi-modality were included. Additionally, we made intensive comparisons based on four well-known benchmark data sets for classification: *USPS* and *MNIST*[11] contain 8-bit gray-scale images of handwritten digits '0' through '9' with the resolution 16×16 and 28×28, *20Newsgroups sorted by date*[12] contains term-frequency vectors of documents that come from twenty newsgroups, and *Isolet*[13] contains acoustic features of isolated spoken letters from 'A' to 'Z'.

In our experiments, the performance was measured by the clustering error rate

$$\frac{1}{n} d_{\mathcal{H}}(\boldsymbol{y}, \boldsymbol{y}^*) = \frac{1}{2n} \min(\|\boldsymbol{y} + \boldsymbol{y}^*\|_1, \|\boldsymbol{y} - \boldsymbol{y}^*\|_1),$$

where $\boldsymbol{y}$ is the label vector returned by clustering algorithms and $\boldsymbol{y}^*$ is the ground truth label vector. The similarity measure was either the Gaussian similarity

$$W_{i,j} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$

with a hyperparameter σ, the cosine similarity

$$W_{i,j} = \begin{cases} \dfrac{\langle x_i, x_j \rangle}{\|x_i\|_2 \|x_j\|_2} & \text{if } x_i \sim_k x_j, \\ 0 & \text{otherwise,} \end{cases}$$

10. The data sets were downloaded from `http://ida.first.gmd.de/~raetsch/data/benchmarks.htm`.
11. The data sets are available at `http://cs.nyu.edu/~roweis/data.html`.
12. The data set is available at `http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html`.
13. The data set is available at `http://archive.ics.uci.edu/ml/datasets/isolet`.

	# Classes	# Features	# Data	# Samplings
Artificial Data				
2gaussians	2	3	-	12×10
2moons	2	2	400	12×10
2circles	2	2	315	12×10
IDA Benchmarks				
Breast-cancer	2	9	200	100
Diabetes	2	8	468	100
Flare-solar	2	9	666	100
German	2	20	700	100
Heart	2	13	170	100
Image	2	18	1300	20
Ringnorm	2	20	400	100
Splice	2	60	1000	20
Titanic	2	3	150	100
Twonorm	2	20	400	100
Other Benchmarks				
USPS	10	256	11000	8×10
MNIST	10	784	70000	8×10
20Newsgroups	7	26214	18846	8×10
Isolet	26	617	7797	8×10

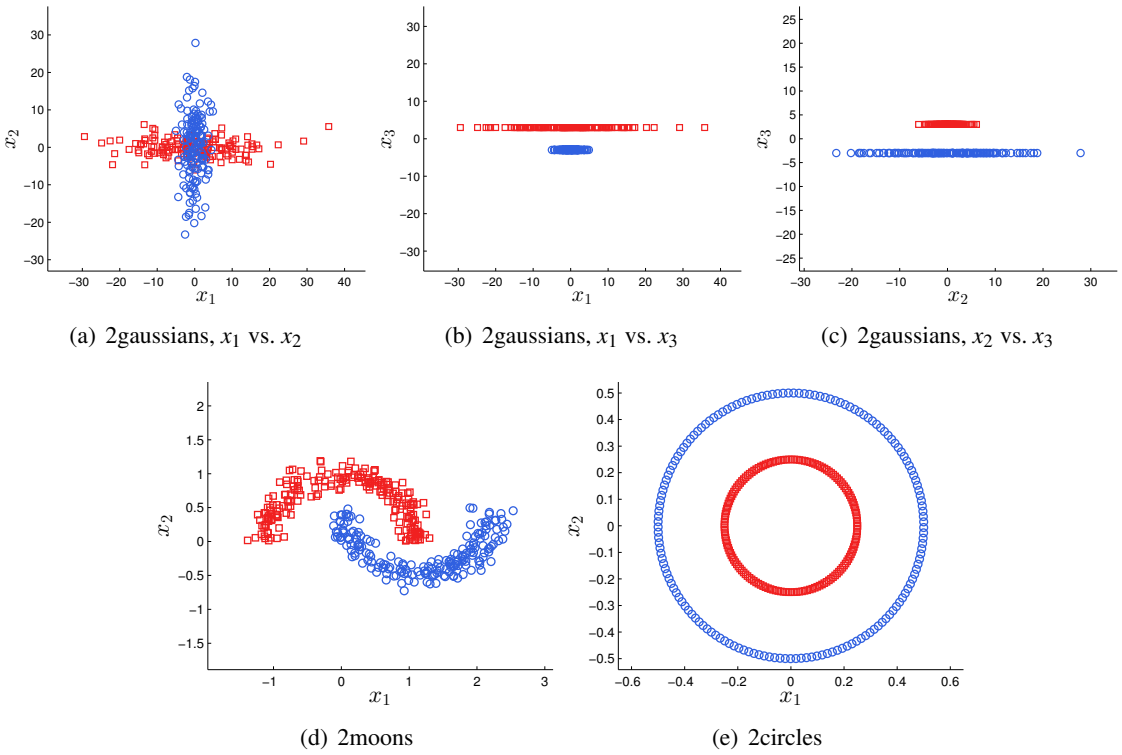
Table 1: Specification of artificial and benchmark data sets.

with a hyperparameter k, where $x_i \sim_k x_j$ means that x_i and x_j are among the k-nearest neighbors of each other, or the locally-scaled Gaussian-like similarity (Zelnik-Manor and Perona, 2005)

$$W_{i,j} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma_i\sigma_j}\right)$$

with a hyperparameter k, where $\sigma_i = \|x_i - x_i^{(k)}\|_2$ is called the local scaling factor of x_i and $x_i^{(k)}$ is the k-th nearest neighbor of x_i in X_n. The kernel matrix was $K = W$ for KM, MMC and LGMMC, and $K = W + I_n/n$ for GMMC since it would be very unstable without this small eigenvalue shift. NSC relied on the graph Laplacian L_{sym} constructed from W. Due to the requirement of positive definiteness of Q for MVC, we also slightly shifted the eigenvalues of certain positive semi-definite matrices and adopted $Q = L_{\mathrm{sym}} + I_n/n$ for MVC-SL and $Q = W + I_n/n$ for MVC-HL.

Numerical issues always exist and there may be more than one candidate $\boldsymbol{h}_0$ for MVC-SL. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of Q, and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the associated normalized eigenvectors. In our implementation, we initialize MVC-SL by a few eigenvectors whose eigenvalues are close to λ_2. Specifically, we construct a set of candidate eigenvectors $\mathcal{V} = \{\boldsymbol{v}_i \mid |\lambda_i - \lambda_2| < 10^{-4}\}$, and if $\#\mathcal{V} > 10$, we say that Q is ill-defined and only keep ten such $\boldsymbol{v}_i$ in $\mathcal{V}$. Next we obtain one $\boldsymbol{h}_0$ from each $\boldsymbol{v}_i \in \mathcal{V}$ and solve the SQP problem based on each $\boldsymbol{h}_0$. At last, the solution $\boldsymbol{h}^*$ resulting in the smallest objective value $-2\|\boldsymbol{h}^*\|_1 + \gamma \boldsymbol{h}^{*\top} Q \boldsymbol{h}^*$ would be selected as the final solution to MVC-SL. This trick can sometimes improve the performance significantly, while the cost is the increase of the computation time by no more than ten times.



(a) 2gaussians, x_1 vs. x_2 (b) 2gaussians, x_1 vs. x_3 (c) 2gaussians, x_2 vs. x_3

(d) 2moons (e) 2circles

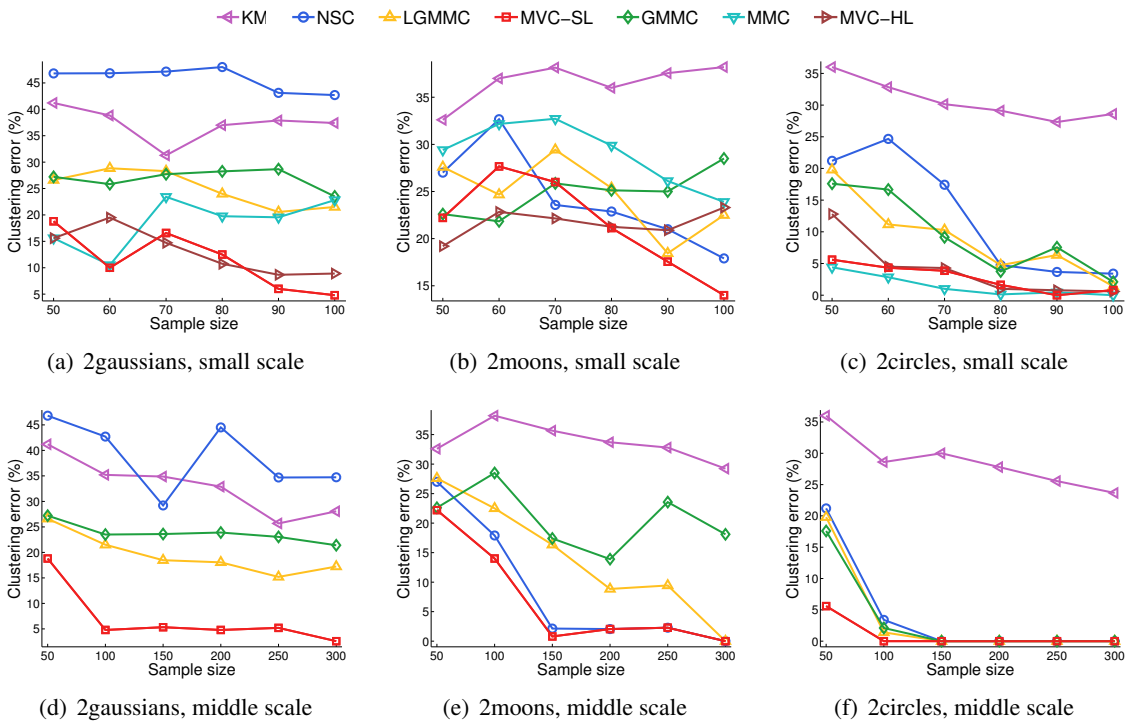
Figure 3: Visualization of artificial data sets.

8.2 Artificial Data Sets

To begin with, we compare the clustering error and the computation time of all seven algorithms based on three artificial data sets. As visualized in Figure 3, *2gaussians* is a three-dimensional data set generated as follows. We first randomly sampled $X_{n/2}^+$ from a Gaussian distribution with zero mean and covariance matrix $\mathrm{diag}(100,4)$ and $X_{n/2}^-$ from the other Gaussian distribution with zero mean and covariance matrix $\mathrm{diag}(4,100)$, set the third dimension as $+3$ for $X_{n/2}^+$ and -3 for $X_{n/2}^-$ and combined $X_{n/2}^+$ and $X_{n/2}^-$ into X_n. Subsequently, *2moons* is a two-dimensional data set with two non-Gaussian crescent-like clusters, and *2circles* is another two-dimensional data set with two non-Gaussian ring-like clusters. The Gaussian similarity was applied to all algorithms, and σ was fixed to $m_\sigma/10$, where m_σ is the mean pairwise distance given by

$$m_\sigma = \frac{\sum_{1\le i<j\le n}\|x_i-x_j\|_2}{n(n-1)/2} = \frac{\sum_{i,j=1}^n\|x_i-x_j\|_2}{n(n-1)}, \tag{27}$$

since $\|x_i-x_j\|_2=0$ when $i=j$. Then, the regularization parameter C of MMC was the best value among $\{10^{-3},1,10^3\}$, that is, we ran MMC three times using $C=10^{-3},1,10^3$ and recorded the best performance, since there lacks a uniformly effective model selection framework for clustering algorithms. The regularization parameter C of LGMMC was also selected from $\{10^{-3},1,10^3\}$ in the same way. For GMMC, the regularization parameter C_e was set to 10^4 following Valizadegan and Jin (2007) and the other regularization parameter C_δ was the best candidate in $\{10^{-3},1,10^3\}$. We fixed the stopping threshold ε to 10^{-6}, the regularization parameter γ to 10^{-2} and let the class



(a) 2gaussians, small scale (b) 2moons, small scale (c) 2circles, small scale

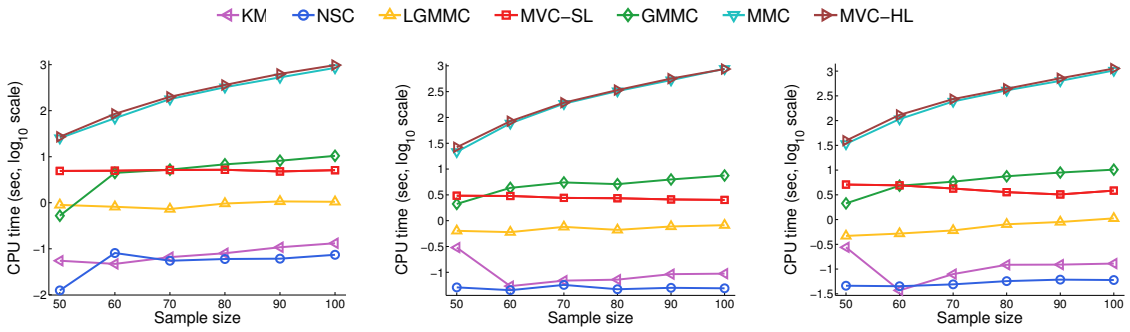
(d) 2gaussians, middle scale (e) 2moons, middle scale (f) 2circles, middle scale

Figure 4: Means of the clustering error (in %) on 2gaussians, 2moons and 2circles.

balance parameter b adaptively be $1/n$ for MVC-SL, while for MVC-HL, we fixed C to 1 and tried $\gamma \in \{10^{-3}, 1, 10^3\}$.

The experimental results in terms of the means of the clustering error are reported in Figure 4. All of the results were obtained by repeatedly running an algorithm on 10 random samplings with given sample size n, and the sample sizes were $\{50, 60, 70, 80, 90, 100\}$ for the small-scale experiments and $\{50, 100, 150, 200, 250, 300\}$ for the middle-scale experiments. We can see that among the three data sets, 2gaussians is most difficult such that LGMMC still had a mean clustering error around twenty percents even when $n = 300$, and 2circles is easiest because MMC, MVC-SL and MVC-HL already got near zero errors when $n = 80$ and LGMMC, GMMC and NSC also achieved perfect partitions after $n = 150$. In contrast, KM cannot deal with these artificial data well due to the non-convex distortion function and the random initialization of cluster centers, even though it was equipped with the Gaussian similarity. Surprisingly, NSC was worse than KM on 2guassians, whereas MVC-SL based on the almost same input $Q = L_{\mathrm{sym}} + I_n/n$ had much lower clustering errors, which implies that the highly non-convex k-means step may be a bottleneck of NSC.

Next we report the corresponding computation time of these algorithms in Figure 5. All of the results were measured in average seconds per run on Xeon X5670 processors. Note that the worst case running time (i.e., the asymptotic time complexity) of KM is super-polynomial in the sample size n (Arthur and Vassilvitskii, 2006), and so is the worst case running time of NSC. On the other hand, the asymptotic time complexities of LGMMC, MVC-SL, GMMC, MMC and MVC-HL are $O(n^3)$, $O(n^3)$, $O(n^{4.5})$, $O(n^{6.5})$ and $O(n^{6.5})$, respectively. In our experiments, NSC was the most computationally-efficient algorithm and almost always faster than KM, since the k-means invoked by NSC after the spectral embedding converged in fewer iterations than KM. While LGMMC was



(a) 2gaussians, small scale (b) 2moons, small scale (c) 2circles, small scale

(d) 2gaussians, middle scale (e) 2moons, middle scale (f) 2circles, middle scale

Figure 5: Means of the CPU time (in sec, per run) on 2gaussians, 2moons and 2circles.

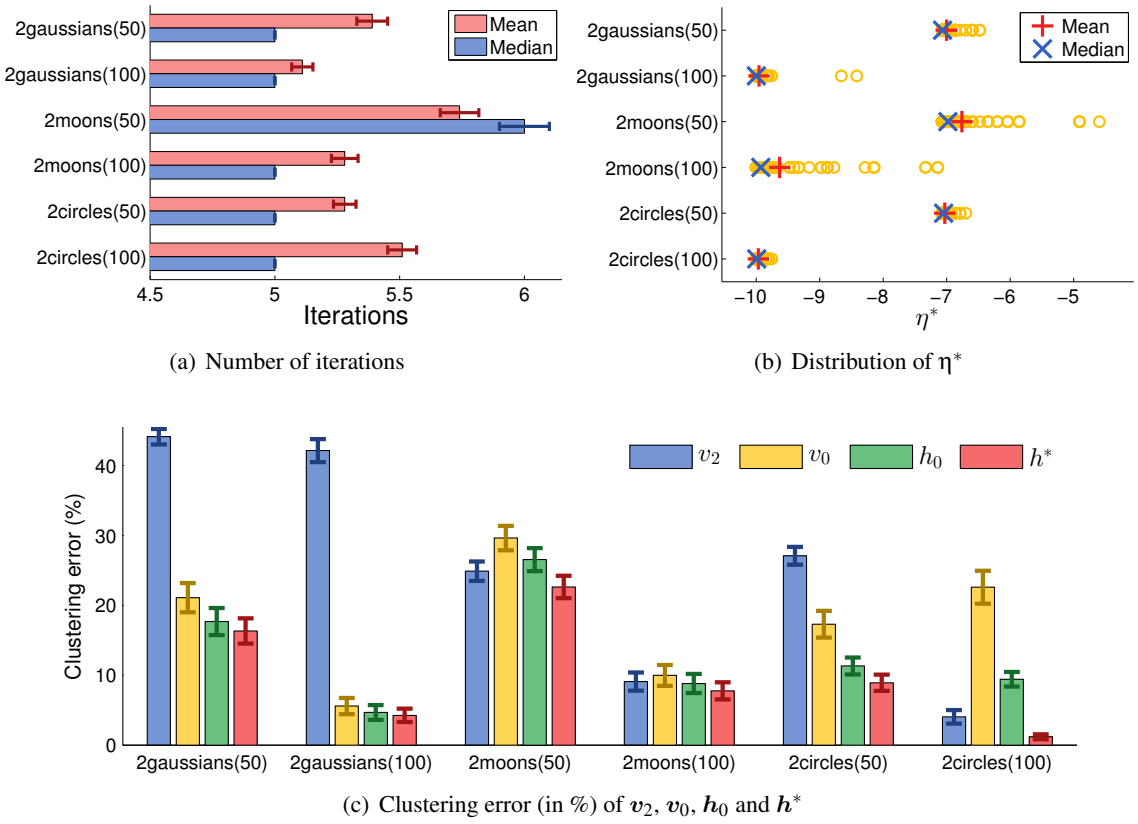
consistently faster than GMMC, MVC-SL lay between them and was comparable with GMMC in the small-scale experiments and comparable with LGMMC in the middle-scale experiments. As a result, the computation time or empirical time complexity of MVC-SL exhibited less potential of growth than LGMMC and GMMC. The worst-case computational complexities of MVC-HL and MMC made them extremely time-consuming, poorly scalable to middle or large sample sizes, and hence impractical despite their low mean clustering errors on 2guassians and 2circles.

Furthermore, we investigate three important properties of MVC-SL, and report the results over 100 random samplings in Figure 6.

Firstly, panel (a) shows the mean and median values about the number of iterations required by MVC-SL, where each mean is shown with the *standard error*, and each median is shown with the *median absolute deviation* divided by the square root of the number of random samplings (i.e., 10). As mentioned before, the convergence rate of SQP iterations is independent of the sample size n, and we can see that MVC-SL usually stopped within just a few iterations in our experiments. This phenomenon implies that the empirical time complexity of MVC-SL is directly proportional to the internal QP solver.

Secondly, we examine the distribution of η^* which may influence the stability of the resulting clusters. Fortunately, panel (b) shows that η^* for fixed data set and fixed sample size were highly concentrated, and the mean and median values exhibited a strong correlation with the sample size as well as a weak correlation with the data set.

Thirdly, recall that there may be more than one candidate $\boldsymbol{h}_0$ and we initialize MVC-SL using $\mathcal{V} = \{\boldsymbol{v}_i \mid |\lambda_i - \lambda_2| < 10^{-4}\}$. Although all $\boldsymbol{v} \in \mathcal{V}$ appear nearly equally good to NSC, they could induce initial solutions of very different qualities for MVC-SL, as shown in panel (c). The vectors



(a) Number of iterations

(b) Distribution of η^*

(c) Clustering error (in %) of $\boldsymbol{v}_2$, $\boldsymbol{v}_0$, $\boldsymbol{h}_0$ and $\boldsymbol{h}^*$

Figure 6: Experimental results concerning three important properties of MVC-SL.

$\boldsymbol{v}_2$, $\boldsymbol{v}_0$, $\boldsymbol{h}_0$, and $\boldsymbol{h}^*$ are all treated as soft response vectors, and the means with standard errors of the clustering error are plotted in panel (c), where $\boldsymbol{v}_2$ is the eigenvector of Q and L_{sym} associated with λ_2, $\boldsymbol{v}_0$ is the eigenvector selected by MVC-SL, and $\boldsymbol{h}_0$ and $\boldsymbol{h}^*$ are the corresponding initial and final solutions. We can see that $\boldsymbol{h}^*$ was better than $\boldsymbol{h}_0$ and $\boldsymbol{h}_0$ was better than $\boldsymbol{v}_0$. Moreover, $\boldsymbol{v}_0$ was significantly superior to $\boldsymbol{v}_2$ on 2guassians. It is interesting and surprising that both $\boldsymbol{h}_0$ and $\boldsymbol{v}_0$ were significantly inferior to $\boldsymbol{v}_2$ on 2circles when $n = 100$, but they still resulted in $\boldsymbol{h}^*$ with the lowest mean clustering error. In a word, not only good initial solutions but also the SQP method contribute to the success of MVC-SL, which in turn implies that the underlying large volume principle should be reasonable for clustering.

8.3 Benchmark Data Sets

In the following, we discuss the experiments on the benchmarks listed in Table 1: The experiments involving ten IDA benchmarks are discussed in the first part, then USPS and MNIST in the second part, 20Newsgroups in the third part, and Isolet in the fourth part.

8.3.1 IDA BENCHMARKS

We compare KM, NSC, LGMMC, GMMC, and MVC-SL on ten data sets in the IDA benchmark repository that are designed for binary classification tasks and have one hundred fixed realizations

	KM	NSC	LGMMC	MVC-SL	GMMC	SVM
Breast-cancer	38.9 ± 0.65	26.4 ± 0.18	27.2 ± 0.19	$\mathbf{25.6 \pm 0.17}$	30.5 ± 0.23	26.0
Diabetes	30.3 ± 0.17	30.6 ± 0.18	$\mathbf{27.6 \pm 0.13}$	30.4 ± 0.15	28.6 ± 0.15	23.5
Flare-solar	$\mathbf{35.5 \pm 0.20}$	44.9 ± 0.11	37.6 ± 0.16	44.5 ± 0.12	N/A	32.4
German	39.4 ± 0.20	$\mathbf{30.2 \pm 0.09}$	$\mathbf{30.1 \pm 0.09}$	$\mathbf{30.2 \pm 0.09}$	N/A	23.6
Heart	$\mathbf{18.5 \pm 0.38}$	$\mathbf{18.0 \pm 0.23}$	18.7 ± 0.28	18.8 ± 0.22	18.9 ± 0.21	16.0
Image	41.0 ± 0.36	40.5 ± 0.15	$\mathbf{39.7 \pm 0.20}$	40.9 ± 0.11	N/A	2.96
Ringnorm	4.68 ± 0.11	$\mathbf{2.20 \pm 0.06}$	6.61 ± 0.11	$\mathbf{2.17 \pm 0.06}$	$\mathbf{2.07 \pm 0.06}$	1.66
Splice	29.1 ± 1.41	35.5 ± 0.44	$\mathbf{25.5 \pm 0.72}$	36.1 ± 0.44	N/A	10.9
Titanic	27.2 ± 0.59	26.8 ± 0.42	23.1 ± 0.36	$\mathbf{21.9 \pm 0.37}$	26.1 ± 0.43	22.4
Twonorm	3.61 ± 0.78	2.28 ± 0.07	$\mathbf{2.18 \pm 0.07}$	$\mathbf{2.20 \pm 0.07}$	$\mathbf{2.08 \pm 0.06}$	2.96

Table 2: Means with standard errors of the clustering error (in %) on IDA benchmark data sets. For each data set, the best algorithm and comparable ones based on the unpaired t-test at the significance level 5% are highlighted in boldface. Additionally, means of the classification error of highly-tuned SVM provided by IDA are also listed for comparison.

for each data set except that the data sets Image and Splice only have twenty realizations. For each realization of each data set, we ignored the test data and tested five clustering algorithms using the training data, yet GMMC was not tested on the data sets Flare-solar, German, Image and Splice as it required a very long execution time when $n \geq 600$. The Gaussian similarity was applied and σ was the best value among $\{4m_\sigma, 2m_\sigma, m_\sigma, m_\sigma/2, m_\sigma/4\}$ for each realization and each algorithm, where the variable m_σ was the mean pairwise distance defined in Equation (27). An exception is the data set Ringnorm where the locally-scaled similarity with $k = 7$ was applied, since it consists of data from two highly overlapped Gaussian distributions and can be treated as a multi-scale data set.[14] The settings for other hyperparameters of LGMMC, GMMC, and MVC-SL were exactly same as the experiments on the artificial data sets, specifically, $C \in \{10^{-3}, 1, 10^3\}$ for LGMMC, $C_e = 10^4$ and $C_\delta \in \{10^{-3}, 1, 10^3\}$ for GMMC, and $\varepsilon = 10^{-6}$, $\gamma = 10^{-2}$ and $b = 1/n$ for MVC-SL.

Table 2 describes the means with standard errors of the clustering error rate by each algorithm on each data set. For the sake of comparison, Table 2 also lists the means of the classification error rate of highly-tuned SVM provided by the official web site of the IDA benchmark repository.

We could see from Table 2 that LGMMC and MVC-SL were either the best algorithm or comparable to the best algorithm based on the unpaired t-test at the significance level 5% on five data sets. The clustering errors of five algorithms exhibited large differences on five data sets, namely, Breast-cancer, Flare-solar, German, Ringnorm and Splice, among which MVC-SL was one of the best algorithms on three data sets, and LGMMC was one of the best algorithms on two data sets. The clustering errors exhibited merely small differences on the other five data sets. Moreover, the fully supervised SVM has a mean classification error obviously smaller than the lowest mean clustering error on the data sets German, Image and Splice, and larger than the lowest mean clustering error on the data sets Breast-cancer, Titanic and Twonorm. It should not be surprising or confusing since the classification error is the out-of-sample test error on the test data whereas the clustering error is the in-sample test error on the same data to be clustered.

14. In fact, Ringnorm violates the underlying assumption when evaluating clustering results using classification data sets, that is, the class structure and the cluster structure must coincide with each other. However, 2circles does not violate this assumption, since those ring-like clusters are neither Gaussian distributions nor overlapped clusters.

8.3.2 IMAGES OF HANDWRITTEN DIGITS

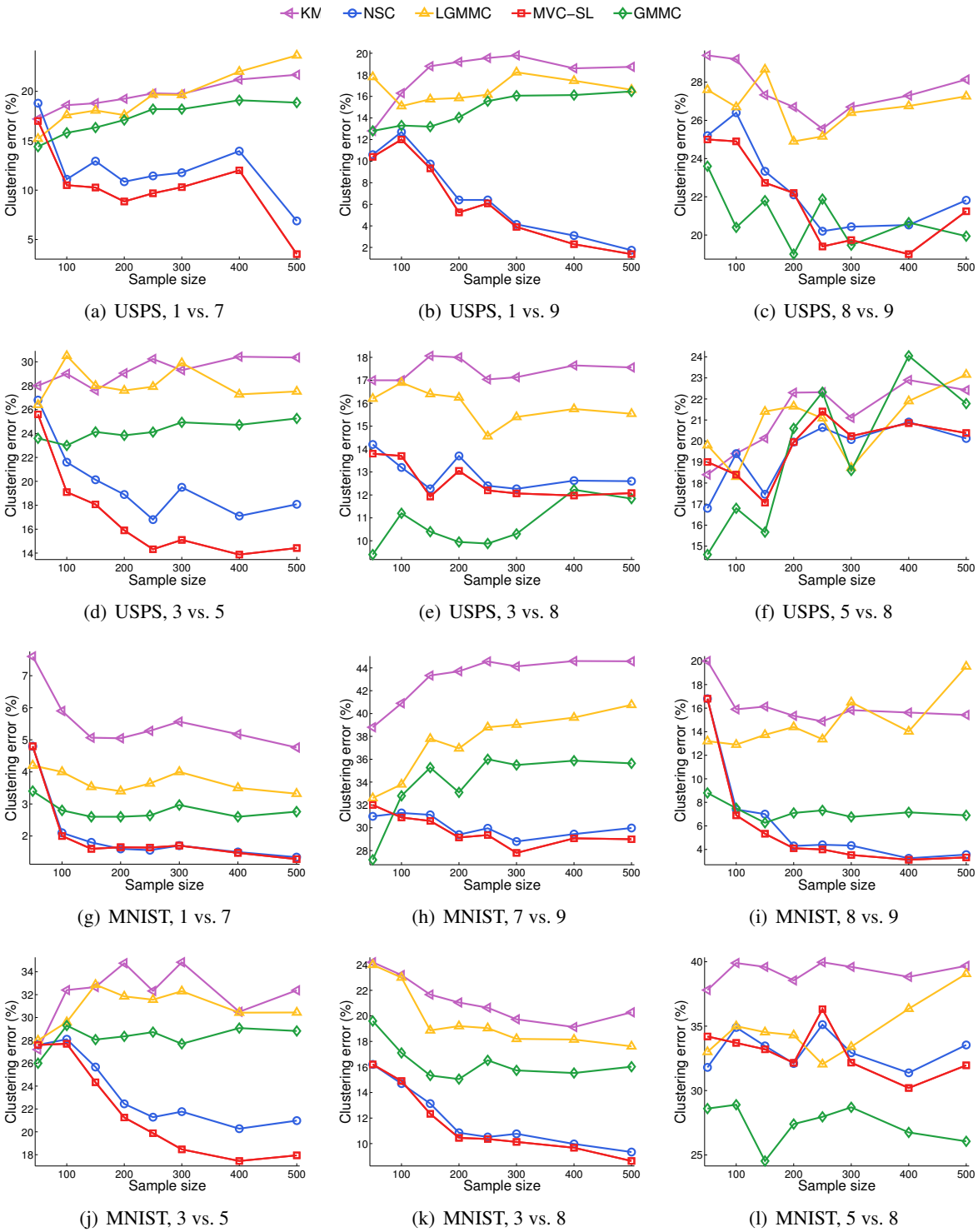
Secondly, we take the images of handwritten digits in USPS and MNIST. Instead of testing KM, NSC, LGMMC, GMMC and MVC-SL on all forty-five pairwise clustering tasks, a few challenging tasks were selected, namely, the pairs $\{1,7\},\{1,9\},\{8,9\},\{3,5\},\{3,8\},\{5,8\}$ of USPS and $\{1,7\},\{7,9\},\{8,9\},\{3,5\},\{3,8\},\{5,8\}$ of MNIST. The task digits 7 vs. 9 of USPS is too hard for all algorithms and then we selected an easier task digits 1 vs. 9. Unlike the training data in the IDA benchmark repository that are already standardized (i.e., normalized to mean zero and standard deviation one) by the provider, the 8-bit gray-scale images in USPS/MNIST are raw data represented by 256-/784-dimensional vectors of integers between 0 and 255. The popular pre-processing is to divide each integer by 255 and thus change the representation to vectors of floating-point numbers between 0 and 1. As a consequence, $\langle x_i, x_j \rangle$ is always nonnegative for any $1 \leq i, j \leq n$ and we can use the cosine similarity for NSC, where in our experiments the hyperparameter k of the k-nearest neighbors was the best value among $\{3,4,5,6,7,8\}$ for each random sampling. The same cosine similarity was also applied to MVC-SL. However, this cosine similarity did not work for the other three algorithms here, and then we still used the Gaussian similarity with σ as the best value among $\{4m_\sigma, 2m_\sigma, m_\sigma, m_\sigma/2, m_\sigma/4\}$ for each random sampling, where m_σ was the mean pairwise distance defined in Equation (27). The settings for other hyperparameters of LGMMC, GMMC, and MVC-SL were exactly same as the experiments on the artificial data sets.

Figure 7 reports the means of the clustering error by each algorithm on each task. The sample sizes were $\{50,100,150,200,250,300,400,500\}$ for all tasks, and each mean value was obtained by repeatedly running an algorithm on 10 random samplings. Given a certain task with sample size n, we first merged all data of the two classes and then randomly sampled a subset of size n, so the classes in the resulting subset were not necessarily balanced when n was small. Moreover, Table 3 summarizes the means with standard errors of the clustering error, in which each algorithm has 80 random samplings on each task. Since the sample sizes here varied in a large range, we performed the paired t-test of the null hypothesis that the difference of the clustering error is from a Gaussian distribution with mean zero and unknown variance, against the alternative hypothesis that the mean is not zero.

We can see from Figure 7 that the easiest task is MNIST 1 vs. 7, such that the mean clustering errors of MVC-SL and NSC were less than two percents when $n \geq 100$, and the hardest tasks are MNIST 7 vs. 9 and 5 vs. 8, where no algorithm was better than twenty-five percents. Both Figure 7 and Table 3 show that the relatively easy tasks include the pairs $\{1,7\},\{1,9\},\{3,8\}$ of USPS and $\{1,7\},\{8,9\},\{3,8\}$ of MNIST, while the relatively hard tasks are the pairs $\{8,9\},\{3,5\},\{5,8\}$ of USPS and $\{7,9\},\{3,5\},\{5,8\}$ of MNIST. In addition, according to Figure 7, the mean clustering errors of MVC-SL were basically non-increasing except in panel (f) USPS 5 vs. 8, and MVC-SL, NSC and GMMC usually outperformed KM and LGMMC, as in Table 3. Similarly, MVC-SL was either the best algorithm or comparable to the best algorithm on ten out of twelve tasks according to Table 3, among which it was best on eight tasks and outperformed all others on seven tasks. The second best algorithm GMMC was best on four tasks, and then NSC was comparable on two tasks. In a word, MVC-SL was fairly promising on USPS and MNIST.

8.3.3 NEWSGROUP DOCUMENTS

The benchmark 20Newsgroups has three versions containing 19997, 18846, and 18828 newsgroup documents, partitioned nearly evenly across twenty different newsgroups. The second version with



(a) USPS, 1 vs. 7 (b) USPS, 1 vs. 9 (c) USPS, 8 vs. 9

(d) USPS, 3 vs. 5 (e) USPS, 3 vs. 8 (f) USPS, 5 vs. 8

(g) MNIST, 1 vs. 7 (h) MNIST, 7 vs. 9 (i) MNIST, 8 vs. 9

(j) MNIST, 3 vs. 5 (k) MNIST, 3 vs. 8 (l) MNIST, 5 vs. 8

Figure 7: Means of the clustering error (in %) on USPS and MNIST.

	KM	NSC	LGMMC	MVC-SL	GMMC
USPS, 1 vs. 7	19.5 ± 0.36	12.2 ± 0.91	19.2 ± 0.68	$\mathbf{10.3 \pm 0.89}$	17.3 ± 0.37
USPS, 1 vs. 9	18.0 ± 0.55	6.9 ± 0.85	16.6 ± 0.54	$\mathbf{6.3 \pm 0.77}$	14.7 ± 0.38
USPS, 8 vs. 9	27.5 ± 0.72	22.5 ± 0.89	26.7 ± 0.87	$\mathbf{21.8 \pm 0.93}$	$\mathbf{20.8 \pm 0.80}$
USPS, 3 vs. 5	29.2 ± 0.61	19.9 ± 0.97	28.1 ± 0.72	$\mathbf{17.0 \pm 0.96}$	24.2 ± 0.62
USPS, 3 vs. 8	17.4 ± 0.52	12.9 ± 0.46	15.9 ± 0.47	12.6 ± 0.47	$\mathbf{10.6 \pm 0.48}$
USPS, 5 vs. 8	21.1 ± 0.65	$\mathbf{19.4 \pm 0.59}$	20.8 ± 0.74	$\mathbf{19.7 \pm 0.67}$	$\mathbf{19.3 \pm 0.84}$
MNIST, 1 vs. 7	5.5 ± 0.36	$\mathbf{2.1 \pm 0.35}$	3.7 ± 0.21	$\mathbf{2.0 \pm 0.35}$	2.8 ± 0.19
MNIST, 7 vs. 9	43.1 ± 0.44	30.1 ± 0.59	37.4 ± 0.58	$\mathbf{29.7 \pm 0.62}$	33.9 ± 0.56
MNIST, 8 vs. 9	16.1 ± 0.96	6.4 ± 0.77	14.7 ± 0.80	$\mathbf{5.9 \pm 0.75}$	7.2 ± 0.36
MNIST, 3 vs. 5	32.1 ± 0.65	23.5 ± 0.66	30.9 ± 0.52	$\mathbf{21.8 \pm 0.72}$	28.3 ± 0.47
MNIST, 3 vs. 8	21.2 ± 0.49	11.9 ± 0.54	19.8 ± 0.58	$\mathbf{11.6 \pm 0.59}$	16.4 ± 0.54
MNIST, 5 vs. 8	39.2 ± 0.47	33.2 ± 1.17	34.7 ± 0.79	33.0 ± 1.22	$\mathbf{27.4 \pm 0.80}$

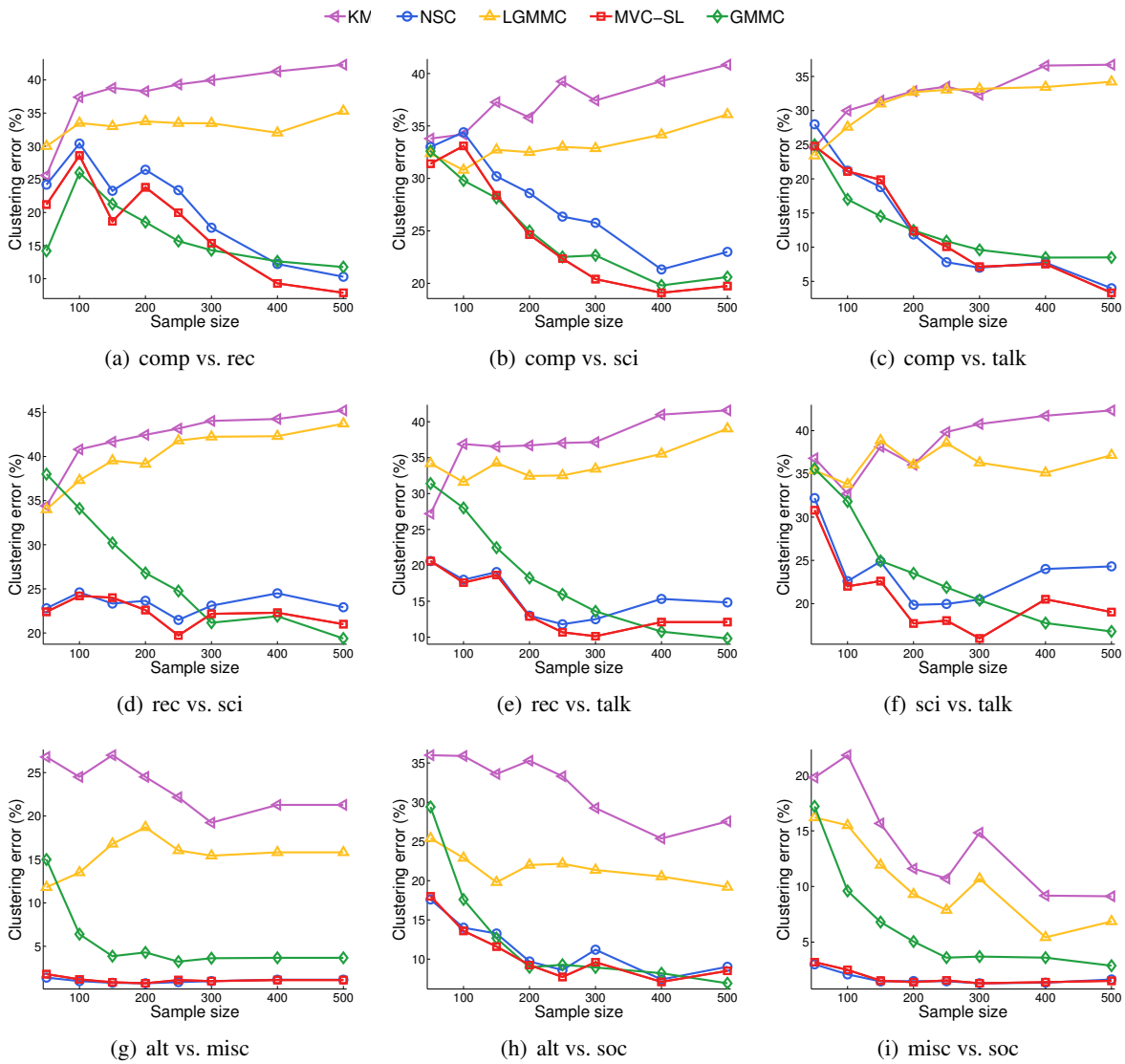
Table 3: Means with standard errors of the clustering error (in %) on USPS and MNIST. For each task, the best algorithm and comparable ones based on the paired t-test at the significance level 5% are highlighted in boldface.

18846 documents is recommended by the original provider[15] and hence is used in our experiments. The documents in 20Newsgroups can be further grouped into seven topics: They are 'alt', 'comp', 'misc', 'rec', 'sci', 'soc' and 'talk', with 799, 4891, 975, 3979, 3952, 997 and 3253 documents respectively, where comp consists of five classes, each of rec, sci and talk consists of four classes, and each of alt, misc and soc consists of a single class. We prepared nine pairwise clustering tasks which included all tasks between the four multi-modal topics and all tasks between the three uni-modal topics. The term-frequency vectors were processed into term-frequency-inverse-document-frequency vectors using the script written by the provider[16] for the whole data set. We tried all of the three similarity measures, and found that for any algorithm no one was consistently better than the other two. However, the locally-scaled similarity generally fitted all five algorithms, where the hyperparameter k was the best value in $\{3,4,5,6,7,8\}$ for each random sampling. The settings for other hyperparameters of LGMMC, GMMC and MVC-SL were exactly same as the experiments on the artificial data sets.

Figure 8 reports the means of the clustering error by each algorithm on each task. The sample sizes were $\{50,100,150,200,250,300,400,500\}$ for all tasks, and each mean value was averaged over 10 random samplings. Similarly to the random samplings of USPS and MNIST, the classes in each random sampling here were not necessarily balanced when n was small. In addition, Table 4 summarizes the means with standard errors of the clustering error, in which each algorithm has 80 random samplings on each task. The paired t-test was performed due to the varied sample sizes.

We can see from Figure 8 and Table 4 that the tasks between the four multi-modal topics are more difficult than the tasks between the three uni-modal topics. Two tasks involving misc (i.e., alt vs. misc and misc vs. soc) are easiest, and three tasks involving sci (i.e., comp vs. sci, rec vs. sci, and sci vs. talk) are hardest. Moreover, MVC-SL, NSC and GMMC usually outperformed KM and LGMMC, and Figure 8 also illustrates that the mean clustering errors of MVC-SL were basically non-increasing. As shown in Table 4, MVC-SL was either the best algorithm or comparable to the

15. See `http://qwone.com/~jason/20Newsgroups/`.
16. See `http://www.cad.zju.edu.cn/home/dengcai/Data/code/tfidf.m`.



(a) comp vs. rec (b) comp vs. sci (c) comp vs. talk

(d) rec vs. sci (e) rec vs. talk (f) sci vs. talk

(g) alt vs. misc (h) alt vs. soc (i) misc vs. soc

Figure 8: Means of the clustering error (in %) on 20Newsgroups.

best algorithm on eight out of nine tasks, among which it was best on six tasks and outperformed all others on four tasks. The second best algorithm NSC was best on three tasks, and then GMMC was best on two tasks and comparable on one task. In a word, MVC-SL was also fairly promising on 20Newsgroups.

8.3.4 ISOLATED SPOKEN LETTERS

The final benchmark is Isolet from the *UCI machine learning repository*. The data were collected by letting 150 subjects speak the name of each letter of the alphabet twice, while two 'F' and one 'M' were dropped due to difficulties in recording. Unlike the features of the previous benchmarks USPS, MNIST and 20Newsgroups, the acoustic features of Isolet are extracted by different ways and possess different physical meanings, including spectral coefficients, contour features, sonorant

	KM	NSC	LGMMC	MVC-SL	GMMC
comp vs. rec	37.9 ± 0.77	21.0 ± 1.46	33.1 ± 0.57	$\mathbf{18.1 \pm 1.41}$	$\mathbf{16.8 \pm 0.74}$
comp vs. sci	37.2 ± 0.65	27.8 ± 1.20	33.1 ± 0.61	$\mathbf{24.9 \pm 1.17}$	$\mathbf{25.1 \pm 0.69}$
comp vs. talk	32.3 ± 0.93	$\mathbf{13.3 \pm 1.69}$	31.1 ± 0.73	$\mathbf{13.3 \pm 1.65}$	$\mathbf{13.3 \pm 0.80}$
rec vs. sci	42.0 ± 0.55	23.3 ± 0.84	40.0 ± 0.73	$\mathbf{22.3 \pm 0.85}$	27.0 ± 1.01
rec vs. talk	36.8 ± 0.76	15.6 ± 1.11	34.1 ± 1.02	$\mathbf{14.3 \pm 1.08}$	18.8 ± 1.08
sci vs. talk	38.5 ± 0.71	23.5 ± 1.01	36.4 ± 0.67	$\mathbf{20.8 \pm 0.97}$	24.1 ± 0.86
alt vs. misc	23.3 ± 1.85	$\mathbf{1.0 \pm 0.12}$	15.5 ± 1.07	1.1 ± 0.13	5.5 ± 0.60
alt vs. soc	32.0 ± 1.05	11.3 ± 1.01	21.7 ± 0.95	$\mathbf{10.7 \pm 0.85}$	12.7 ± 0.91
misc vs. soc	14.1 ± 1.32	$\mathbf{1.7 \pm 0.16}$	10.5 ± 0.67	$\mathbf{1.8 \pm 0.16}$	6.6 ± 0.60

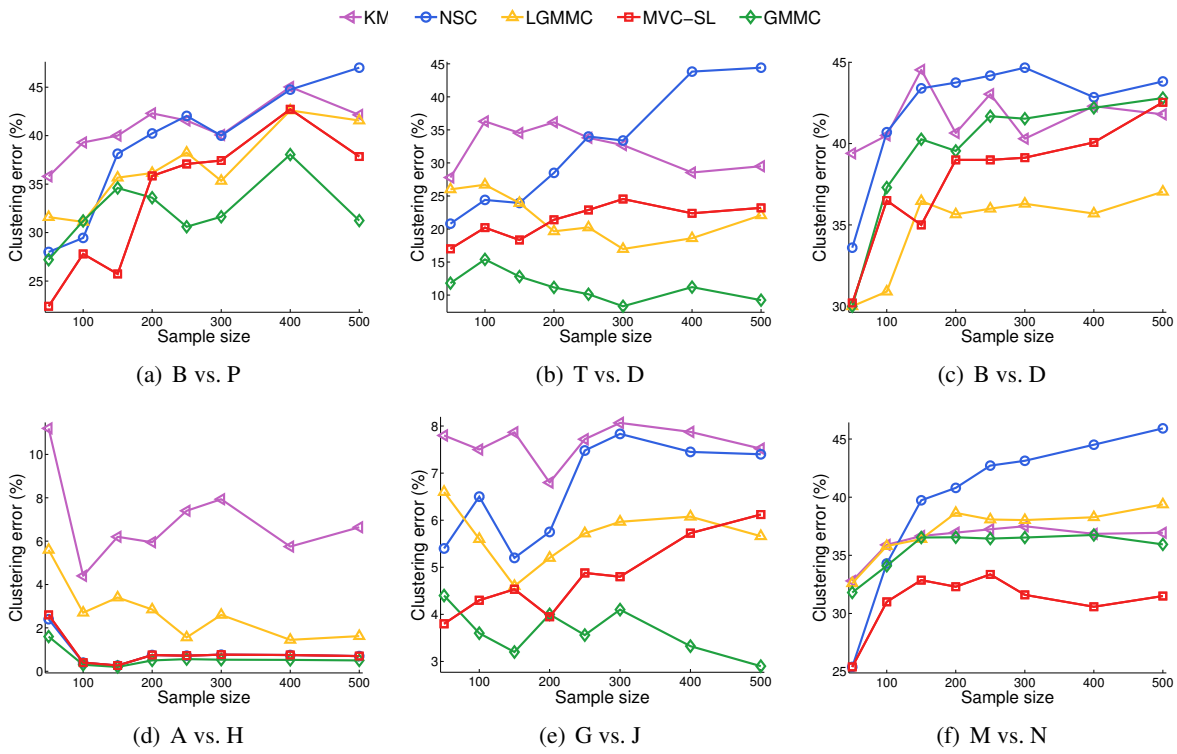
Table 4: Means with standard errors of the clustering error (in %) on 20Newsgroups. For each task, the best algorithm and comparable ones based on the paired t-test at the significance level 5% are highlighted in boldface.

	KM	NSC	LGMMC	MVC-SL	GMMC
B vs. P	40.8 ± 0.64	38.7 ± 1.04	36.5 ± 0.86	$\mathbf{33.4 \pm 1.25}$	$\mathbf{32.3 \pm 1.30}$
T vs. D	32.4 ± 0.93	31.7 ± 1.48	21.8 ± 1.24	21.2 ± 1.02	$\mathbf{11.2 \pm 1.09}$
B vs. D	41.6 ± 0.55	42.1 ± 0.63	$\mathbf{34.8 \pm 0.73}$	37.7 ± 0.82	39.4 ± 0.73
A vs. H	6.9 ± 0.68	0.8 ± 0.19	2.7 ± 0.41	0.9 ± 0.21	$\mathbf{0.6 \pm 0.15}$
G vs. J	7.6 ± 0.32	6.6 ± 0.72	5.7 ± 0.28	4.8 ± 0.28	$\mathbf{3.6 \pm 0.22}$
M vs. N	36.4 ± 0.49	39.6 ± 0.87	37.2 ± 0.47	$\mathbf{31.1 \pm 0.64}$	35.6 ± 0.47

Table 5: Means with standard errors of the clustering error (in %) on Isolet. For each task, the best algorithm and comparable ones based on the paired t-test at the significance level 5% are highlighted in boldface.

features, pre-sonorant features and post-sonorant features. All features are real-valued and scaled into the range -1 to $+1$. Generally speaking, all five algorithms can easily deal with the majority of pairwise clustering tasks, if we randomly choose two letters. Therefore, similarly to USPS and MNIST, a few challenging tasks that might sometimes be difficult for the mankind were selected: The letters B vs. P, T vs. D, B vs. D, A vs. H, G vs. J, and M vs. N. The hyperparameters here were slightly different from the previous experiments for better performance. The cosine similarity was applied to NSC, and the hyperparameter k was the best value in $\{1,2,3,4,5,6\}$ for each random sampling. The Gaussian similarity was still used for KM, LGMMC and GMMC, and the hyperparameter σ was the best value in $\{2m_\sigma, m_\sigma, m_\sigma/2, m_\sigma/4, m_\sigma/8\}$ for each random sampling, where m_σ was defined in Equation (27). For MVC-SL, we adopted either $Q = L_{\mathrm{sym}} + I_n/n$ where L_{sym} was constructed from the cosine similarity or $Q = nI_n - W$ with the Gaussian similarity depending on the task and the sample size n, and the hyperparameter k or σ was chosen in the same way. A key observation here was that for certain tasks such as M vs. N, the former specification was preferable for small n, whereas the latter specification was more advisable for relatively large n. The settings for other hyperparameters were exactly same as the experiments on the artificial data sets.

Figure 9 reports the means of the clustering error by each algorithm on each task. The sample sizes were $\{50,100,150,200,250,300,400,500\}$ for all tasks, and each mean value was averaged



(a) B vs. P (b) T vs. D (c) B vs. D

(d) A vs. H (e) G vs. J (f) M vs. N

Figure 9: Means of the clustering error (in %) on Isolet.

over 10 random samplings. Similarly to the random samplings of USPS and MNIST, the classes in each random sampling here were not necessarily balanced when n was small. In addition, Table 5 summarizes the means with standard errors of the clustering error, in which each algorithm has 80 random samplings on each task. The paired t-test was performed due to the varied sample sizes.

We can see from Figure 9 and Table 5 that the tasks A vs. H and G vs. J are very easy, and the tasks B vs. P, B vs. D and M vs. N are very hard. Interestingly, T vs. D is much easier than B vs. P and B vs. D, such that the lowest mean clustering errors on B vs. P and B vs. D were almost three times larger than the lowest mean clustering error on T vs. D. Unlike the curves shown in Figures 7 and 8, the mean clustering errors of MVC-SL in Figure 9 were basically non-increasing only in panel (d) A vs. H. Furthermore, LGMMC instead of NSC became a competitive algorithm besides GMMC and MVC-SL in Table 5, unlike the performance in Tables 3 and 4. According to Table 5, GMMC was the best algorithm on four tasks, MVC-SL was best on one task and also comparable to the best algorithm on one task, and LGMMC was best on one task. Nevertheless, MVC-SL was still satisfying on Isolet, if considering that MVC-SL consumed less than five percents of the total computation time while GMMC consumed over ninety percents, and thus GMMC was remarkably less computationally-efficient than MVC-SL.

9. Conclusions

We proposed a new discriminative clustering model called maximum volume clustering (MVC) to partition the data samples into two clusters based on the large volume principle. Two algorithms to

approximate the basic model of MVC were developed: MVC-HL relaxes MVC to a semi-definite programming problem that is convex but time-consuming; MVC-SL employs sequential quadratic programming that is non-convex but computationally-efficient. Then, we demonstrated that MVC includes the optimization problems of some well-known clustering methods as special limit cases, and discussed the finite sample stability and the clustering error bound of MVC-SL in great detail. Based on the encouraging experimental results on three artificial and fourteen benchmark data sets, we conclude that the proposed MVC approach is promising, especially for images and text.

The future work includes but is not limited to the following three directions: Multi-way extension, improved optimization, and model selection and specification of Q. We briefly discuss these future directions below.

First of all, the basic model of MVC is currently binary, and it needs a multi-way extension to partition the data samples into more than two clusters. To this end, we should extend the definition of the volume before extending the basic model of MVC. Unlike the margin, there exists no multi-class definition of the volume hitherto. We may borrow the idea of the multi-class definition of the margin in Crammer and Singer (2001) based on which the first multi-way extension of MMC was proposed (Xu and Schuurmans, 2005).

Secondly, the proposed approximation schemes and optimization algorithms for MVC may be improved. However, we believe that the improvement cannot be straightforward. We have considered several options and found that none of them befits MVC well. Recall that the primal problem of MVC-SL defined in (4) is non-convex, and the *concave-convex procedure* and *constrained concave-convex procedure* (CCCP) (Yuille and Rangarajan, 2003; Smola et al., 2005) seem able to solve it. In fact, the former technique can only be applied to the Lagrange function $L(\boldsymbol{h},\eta)$, and η as an optimization variable may diverge even though $\boldsymbol{h}$ is guaranteed to converge given constant η. On the other hand, the latter technique accepts any first-order equality constraint and any inequality constraint involving the difference of two convex functions, but the second-order equality constraint like $\boldsymbol{h}^\top\boldsymbol{h}=1$ is unacceptable. If we relax the equality constraint $\boldsymbol{h}^\top\boldsymbol{h}=1$ into an inequality constraint $\boldsymbol{h}^\top\boldsymbol{h}\leq 1$, we will get

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; -2\|\boldsymbol{h}\|_1+\gamma\boldsymbol{h}^\top Q\boldsymbol{h} \qquad \text{s.t. } \boldsymbol{h}^\top\boldsymbol{h}\leq 1. \tag{28}$$

Unfortunately, CCCP fails to solve optimization (28) again, since now we cannot assume that $\|\boldsymbol{h}\|_1$ is differentiable, and then we cannot easily linearize the concave part of the energy function. Note that the popular trick to cope with ℓ_1-regularization is futile here, since (28) is never equivalent to

$$\begin{aligned}\min_{\boldsymbol{h}\in\mathbb{R}^n} \; & -2\boldsymbol{\alpha}^\top\mathbf{1}_n+\gamma\boldsymbol{h}^\top Q\boldsymbol{h}\\ \text{s.t. } & \boldsymbol{h}^\top\boldsymbol{h}\leq 1,-\boldsymbol{\alpha}\leq\boldsymbol{h}\leq\boldsymbol{\alpha},\boldsymbol{\alpha}\geq\mathbf{0}_n.\end{aligned}$$

Similarly, (28) itself is not *quadratically-constrained quadratic programming* (QCQP) (Boyd and Vandenberghe, 2004) due to the minimization of negative ℓ_1-norm, but it can be reformulated as a QCQP with an optimization variable essentially in $\mathbb{R}^{2n}$:

$$\min_{\boldsymbol{y}\in[-1,+1]^n}\min_{\boldsymbol{h}\in\mathbb{R}^n} \; -2\boldsymbol{h}^\top\boldsymbol{y}+\gamma\boldsymbol{h}^\top Q\boldsymbol{h} \qquad \text{s.t. } \boldsymbol{h}^\top\boldsymbol{h}\leq 1. \tag{29}$$

Although optimization (29) is convex in $\boldsymbol{y}$ and convex in $\boldsymbol{h}$, it is not jointly convex in $\boldsymbol{y}$ and $\boldsymbol{h}$, so no off-the-shelf QCQP solver is applicable and we need relax it via semi-definite programming or *reformulation-linearization technique* (Sherali and Adams, 1998) once more. Actually, the feasible

region $[-1,+1]^n$ of $\boldsymbol{y}$ is as difficult as the combinatorial $\{-1,+1\}^n$, and all of optimizations (2), (4), (28) and (29) are NP-hard, regardless of the different feasible regions of $\boldsymbol{h}$. That being said, the current implementation using sequential quadratic programming is imperfect as the final $\boldsymbol{h}^*$ is a bit sensitive to the initial $\boldsymbol{h}_0$ (see the experimental results reported in Figure 6 for details).

In contrast to MVC-SL, there is much more room for MVC-HL to be improved. GMMC uses a tricky substitution to get (23), and that substitution is so specific that it does not work for MVC-HL. Following the idea of LGMMC, we can obtain an alternative relaxation as

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathbb{R}^{2^n}} \min_{\boldsymbol{\alpha}} \; & -2\boldsymbol{\alpha}^\top \mathbf{1}_n + \gamma\boldsymbol{\alpha}^\top \left(\sum\nolimits_{t:-b\le \boldsymbol{y}_t^\top \mathbf{1}_n \le b} \mu_t Q \circ \boldsymbol{y}_t \boldsymbol{y}_t^\top\right) \boldsymbol{\alpha} \\ \text{s.t.} \; & \boldsymbol{\mu}^\top \mathbf{1}_{2^n} = 1, \boldsymbol{\mu} \ge \mathbf{0}_{2^n} \\ & \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1, \boldsymbol{\alpha} \ge \mathbf{0}_n.\end{aligned}$$

Similarly, this optimization can also be regarded as a multiple kernel learning problem and solved by the cutting plane method, as LGMMC. However, the inner optimization subproblem is difficult due to $\boldsymbol{\alpha}^\top\boldsymbol{\alpha} = 1$ instead of $\boldsymbol{\alpha}^\top\mathbf{1}_n = 1$ in LGMMC, and we decide to investigate how to solve it in our future study since MVC-HL is not the main focus of the current paper.

Thirdly, in our experiments we always use the best candidate hyperparameters in the hindsight, since there lacks a systematic way to tune the hyperparameters for clustering. Such choices may be acceptable from the theoretical standpoint but not enough from the practical standpoint. Notice that any (cross-) validation technique using the clustering error, which is the in-sample test error on the same data to be clustered, simply does not work for model selection. In order to do model selection, a criterion other than the clustering error is necessary. Fortunately, a few information criteria exist though they are not uniformly effective for all clustering algorithms. In Sugiyama et al. (2011), the *mutual information* (MI) (Shannon, 1948) was used for *MI based clustering* (Gomes et al., 2010) via *maximum likelihood MI* (Suzuki et al., 2008) for model selection, and *squared-loss MI* (Suzuki et al., 2009) was used for *squared-loss MI based clustering* (Sugiyama et al., 2011) via *least-squares MI* (Suzuki et al., 2009) for model selection.

What is more, it is unclear how to specify the input matrix Q appropriately for a given data set, including a proper similarity measure and the construction of Q from it. According to von Luxburg et al. (2012), the former issue is actually open for all existing clustering algorithms and it probably has no uniformly effective solution. For the latter issue, we suggest MVC-SL with $Q = L_{\text{sym}} + I_n/n$, where L_{sym} is the normalized graph Laplacian, and the underlying similarity measure can be any similarity suitable for spectral clustering. Then, it is still unsolved when we should use MVC-SL, and when we should use the family of MMC or other clustering algorithms. Unfortunately, there is no answer from a theoretical point of view since clustering has no supervision at all. Nevertheless, MVC-SL may work with high probability in practice when spectral clustering works. We argue that it may be the minimization of negative ℓ_1-norm in MVC-SL that has improved spectral clustering as shown in panel (c) of Figure 6. Its preference of non-sparse optimal solutions may lead to a better approximation to the normalized cut criterion (Shi and Malik, 2000) than spectral clustering.

Acknowledgments

The authors would like to thank anonymous reviewers for the helpful comments. GN is supported by the MEXT scholarship No. 103250, LS is supported by the NSFC programs No. 61170180 and

No. 61035003, and MS is supported by the FIRST program. The preliminary work has been done when GN was studying at Department of Computer Science and Technology, Nanjing University, and BD was studying at Institute of Automation, Chinese Academy of Sciences.

Appendix A. Proofs of Theoretical Results in Section 5.2

In this appendix, we prove the lemmas and theorems appeared in Section 5.2.

A.1 Proof of Lemma 9

If $\exists j \in \{1,\dots,n\}$, $\boldsymbol{e}_j$ or $-\boldsymbol{e}_j$ is an eigenvector of Q, there should exist an eigenvalue $\lambda > 0$ such that $Q\boldsymbol{e}_j = \lambda \boldsymbol{e}_j$. This equation means that $Q_{j,j} = \lambda$ and $\forall i \neq j, Q_{i,j} = 0$. In other words, x_j is isolated and X_n is reducible. ■

A.2 Proof of Theorem 10

If x_i is isolated in X_n, let $\delta_1 = \cdots = \delta_n = 1$, $\mathcal{K} = \{i\}$ and by definition X_n is SI-symmetric.

If X_n is axisymmetric under a permutation ϕ, without loss of generality, we assume $\phi(1) = 2$ and let $\delta_1 = -1, \delta_2 = \cdots = \delta_n = 1$ and $\mathcal{K} = \{1,2\}$. Then X_n is SI-symmetric by Equation (20),

$$\left(\sum\nolimits_{k\in\mathcal{K}} \delta_k \boldsymbol{e}_k\right)^\top Q \left(\sum\nolimits_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k\right) = \sum_{i=3}^{n} (Q_{2,i} - Q_{1,i}) = 0,$$

since $\forall i \in \{3,\dots,n\}$, $\phi(i) \notin \{1,2\}$ and $Q_{1,i} = Q_{2,\phi(i)}$. ■

A.3 Proof of Theorem 11

When $n = 2$, X_2 must be axisymmetric if $Q_{1,1} = Q_{2,2}$, and we can know that X_2 is SI-symmetric by Theorem 10.

When $n > 2$, assume that X_n is irreducible due to Theorem 10, and then $\mathbb{R}^n$ has two disjoint bases according to Lemma 9: The standard basis and the set of the principle axes of $\mathcal{E}(\mathcal{H}_Q)$. We present an indirect proof of the theorem as follows.

Step 1. Let $\lambda_1,\dots,\lambda_n$ be the eigenvalues of Q and $\boldsymbol{v}_1,\dots,\boldsymbol{v}_n$ be the associated normalized eigenvectors. Suppose that $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the directions of two principal axes of $\mathcal{E}(\mathcal{H}_Q)$ with the same length $1/\sqrt{\lambda_i} = 1/\sqrt{\lambda_j}$. There should be at least one principal axis $\boldsymbol{v}_l$ such that $l \notin \{i,j\}$, $\lambda_l \neq \lambda_i$, and $\mathcal{E}(\mathcal{H}_Q)$ is rotational about $\boldsymbol{v}_l$ along the circle

$$C(\boldsymbol{v}_i, \boldsymbol{v}_j) := \{\cos(\theta)\boldsymbol{v}_i + \sin(\theta)\boldsymbol{v}_j \mid \theta \in [0, 2\pi)\}.$$

Otherwise, all principal axes have the same length and thus $\mathcal{E}(\mathcal{H}_Q)$ is a perfect ball, which contradicts the fact that $\boldsymbol{e}_1,\dots,\boldsymbol{e}_n$ are not eigenvectors of Q.

Further suppose that $\lambda_k \neq \lambda_l$ for any $k \neq l$, that is, the principal axis with the direction $\boldsymbol{v}_l$ has a unique length. As a consequence, $\boldsymbol{v}_l$ has a fixed position and cannot rotate within $C(\boldsymbol{v}_k, \boldsymbol{v}_l)$ for any $k \notin \{i,j,l\}$. Otherwise, all vectors in $C(\boldsymbol{v}_k, \boldsymbol{v}_l)$ are legal principal axes and can be considered as $\boldsymbol{v}_l$ with a fixed position.

We can know that $\mathcal{E}(\mathcal{H}_Q)$ intersects the k-th coordinate axis at $\pm\boldsymbol{e}_k/\sqrt{\kappa}$ from $Q_{k,k} = \kappa$, and the intersections compose an $(n-1)$-dimensional hyperplane. Principal axes of $\mathcal{E}(\mathcal{H}_Q)$ are orthogonal

and have at most $(n-1)$ distinct lengths, and $\mathcal{E}(\mathcal{H}_Q)$ also has a set of n orthogonal axes with the same length $1/\sqrt{\kappa}$, that is, the set $\{\boldsymbol{e}_1/\sqrt{\kappa},\ldots,\boldsymbol{e}_n/\sqrt{\kappa}\}$. Hence, any principal axis in a fixed position, especially $\boldsymbol{v}_l$, should lie on the central direction of a certain quadrant with dimensionality at least two. In other words, $\boldsymbol{v}_l$ can be written in the form of

$$\boldsymbol{v}_l = \frac{1}{\sqrt{\sum_{k=1}^n \delta_k^2}} \sum_{k=1}^n \delta_k \boldsymbol{e}_k, \quad \delta_k \in \{-1,0,1\},$$

where $\delta_1,\ldots,\delta_n$ cannot be all zeros.

Step 2. Let $\mathcal{K} = \{k \mid \delta_k = 0\}$ and one has $0 \le \#\mathcal{K} < n$ where $\#$ measures the cardinality. We discuss the cases $\#\mathcal{K} > 0$ and $\#\mathcal{K} = 0$ separately.

If $\#\mathcal{K} > 0$, we reset $\delta_k = 1$ for $k \in \mathcal{K}$. Subsequently,

$$\begin{aligned}
\left(\sum\nolimits_{k\in\mathcal{K}} \delta_k \boldsymbol{e}_k\right)^\top Q \left(\sum\nolimits_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k\right) &= \left(\sum\nolimits_{k\in\mathcal{K}} \boldsymbol{e}_k\right)^\top Q \left(\sqrt{n-\#\mathcal{K}}\boldsymbol{v}_l\right) \\
&= \left(\sum\nolimits_{k\in\mathcal{K}} \boldsymbol{e}_k\right)^\top \sqrt{n-\#\mathcal{K}}(Q\boldsymbol{v}_l) \\
&= \left(\sum\nolimits_{k\in\mathcal{K}} \boldsymbol{e}_k\right)^\top \sqrt{n-\#\mathcal{K}}(\lambda_l \boldsymbol{v}_l) \\
&= \lambda_l \left(\sum\nolimits_{k\in\mathcal{K}} \boldsymbol{e}_k\right)^\top \left(\sqrt{n-\#\mathcal{K}}\boldsymbol{v}_l\right) \\
&= \lambda_l \left(\sum\nolimits_{k\in\mathcal{K}} \boldsymbol{e}_k\right)^\top \left(\sum\nolimits_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k\right) \\
&= \lambda_l \sum_{k\in\mathcal{K},k'\notin\mathcal{K}} \delta_{k'} \boldsymbol{e}_k^\top \boldsymbol{e}_{k'} \\
&= 0,
\end{aligned}$$

due to $Q\boldsymbol{v}_l = \lambda_l \boldsymbol{v}_l$ and the orthonormal condition of the basis $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}$. If $\#\mathcal{K} = 0$, without loss of generality, assume that $\delta_1 = -\delta_2 = 1$ since $n > 2$ and the sign of $\boldsymbol{v}_l$ is arbitrary. The first two rows of the eigenvalue equation $Q\boldsymbol{v}_l = \lambda_l \boldsymbol{v}_l$ tell us

$$\begin{cases}
\kappa - Q_{1,2} + \sum_{k=3}^n \delta_k Q_{1,k} = \lambda_l \\
Q_{2,1} - \kappa - \sum_{k=3}^n \delta_k Q_{2,k} = -\lambda_l
\end{cases} \quad \Rightarrow \quad \sum_{k=3}^n \delta_k (Q_{1,k} - Q_{2,k}) = 0.$$

Hence by resetting $\mathcal{K} = \{1,2\}$, we obtain

$$\left(\sum\nolimits_{k\in\mathcal{K}} \delta_k \boldsymbol{e}_k\right)^\top Q \left(\sum\nolimits_{k\notin\mathcal{K}} \delta_k \boldsymbol{e}_k\right) = \sum_{k=3}^n \delta_k (Q_{1,k} - Q_{2,k}) = 0.$$

Both cases lead to a contradiction since X_n is SI-asymmetric.

Therefore, all principal axes of $\mathcal{E}(\mathcal{H}_Q)$ have distinct lengths, which is exactly what we were to prove. ■

A.4 Proof of Theorem 12

Let us denote $\boldsymbol{h}^* = (h_1, \ldots, h_n)^\top$ and consider $\boldsymbol{h}^\star = (h_{\phi(1)}, \ldots, h_{\phi(n)})^\top$.

Obviously, $\|\boldsymbol{h}^\star\|_1 = \|\boldsymbol{h}^*\|_1$ and $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{h}^*\|_2$. Moreover,

$$\sum_{i,j=1}^{n} Q_{i,j} h_{\phi(i)} h_{\phi(j)} = \sum_{i,j=1}^{n} Q_{\phi(i),\phi(j)} h_{\phi(i)} h_{\phi(j)} = \sum_{k,l=1}^{n} Q_{k,l} h_k h_l,$$

because of the third property of Definition 7. Hence, $\boldsymbol{h}^{\star\top} Q \boldsymbol{h}^\star = \boldsymbol{h}^{*\top} Q \boldsymbol{h}^*$ and then $G(\boldsymbol{h}^\star) = G(\boldsymbol{h}^*)$.

Similarly, $\forall i \in \{1, \ldots, n\}$,

$$\sum_{j=1}^{n} Q_{i,j} h_{\phi(j)} = \sum_{j=1}^{n} Q_{\phi(i),\phi(j)} h_{\phi(j)} = \sum_{k=1}^{n} Q_{\phi(i),k} h_k,$$

where we use the third property of Definition 7 again. As a result,

$$\begin{aligned}
[g(\boldsymbol{h}^\star)]_i &= \gamma \sum_{j=1}^{n} Q_{i,j} h_{\phi(j)} - \eta h_{\phi(i)} - \operatorname{sign}(h_{\phi(i)}) \\
&= \gamma \sum_{k=1}^{n} Q_{\phi(i),k} h_k - \eta h_{\phi(i)} - \operatorname{sign}(h_{\phi(i)}) \\
&= [g(\boldsymbol{h}^*)]_{\phi(i)}.
\end{aligned}$$

Hence, $g(\boldsymbol{h}^\star) = \mathbf{0}_n$ according to the Karush-Kuhn-Tucker condition $g(\boldsymbol{h}^*) = \mathbf{0}_n$, which means that $\boldsymbol{h}^\star$ is also a minimum of optimization (4), since the Hessian matrix $\nabla^2 G(\boldsymbol{h}) = 2(\gamma Q - \eta I_n)$ must be symmetric and positive-definite.

Notice that $d_{\mathcal{H}}(\boldsymbol{h}^*, \boldsymbol{h}^\star) \geq 1$ since $\exists i, h_{\phi(i)} h_i < 0$, with the only exception $d_{\mathcal{H}}(\boldsymbol{h}^*, \boldsymbol{h}^\star) = 0$ when $\operatorname{sign}(\boldsymbol{h}^\star) = -\operatorname{sign}(\boldsymbol{h}^*)$, that is, $\forall i, h_{\phi(i)} h_i < 0$. This completes the proof. ■

A.5 Proof of Theorem 13

We prove the theorem in three steps.

Step 1. Let $0 < \lambda_1 < \cdots < \lambda_n$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the eigenvalues and eigenvectors of Q. Given a minimum $\boldsymbol{h}$, the Karush-Kuhn-Tucker condition $g(\boldsymbol{h}) = \mathbf{0}$ implies that

$$\boldsymbol{h} = \hat{Q}\boldsymbol{y}, \tag{30}$$

where $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{h})$, $\hat{Q} = (\gamma Q - \eta I_n)^{-1}$, and the unknown η satisfies $\eta < \gamma\lambda_1$. Plug Equation (30) into the constraint $\|\boldsymbol{h}\|_2 = 1$, note that $\hat{Q}$ is a symmetric matrix, and then we will have

$$\boldsymbol{y}^\top \hat{Q}^2 \boldsymbol{y} = (\hat{Q}\boldsymbol{y})^\top (\hat{Q}\boldsymbol{y}) = \boldsymbol{h}^\top \boldsymbol{h} = 1. \tag{31}$$

All eigenvalues of Q are different and positive since X_n is anisotropic, so are all eigenvalues of $\hat{Q}$. Consequently, $\hat{Q}^2$ has a unique spectral decomposition. It is easy to see that

$$\boldsymbol{y}^\top \hat{Q}^2 \boldsymbol{y} = \boldsymbol{y}^\top \left(\sum_{i=1}^{n} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \right) \boldsymbol{y} = \sum_{i=1}^{n} \mu_i (\boldsymbol{v}_i^\top \boldsymbol{y})^2, \tag{32}$$

where $\mu_i = 1/(\gamma\lambda_i - \eta)^2$ is the i-th largest eigenvalue of $\hat{Q}^2$.

Step 2. Define a linear mapping

$$\begin{aligned}\psi : \mathbb{R}^n_\beta &\mapsto \mathbb{R}^n \\ \boldsymbol{\beta} &\mapsto \beta_1 \boldsymbol{v}_1 + \cdots + \beta_n \boldsymbol{v}_n,\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$, $\mathbb{R}^n_\beta = \mathbb{R}^n$ and we just use the symbol $\mathbb{R}^n_\beta$ to distinguish the domain and the range of ψ. It is obvious that ψ is a vector space automorphism, and the set of vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ and the set of images $\{\psi(\boldsymbol{e}_1), \ldots, \psi(\boldsymbol{e}_n)\} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ are completely different bases due to Theorem 10 and Lemma 9.

Let $\boldsymbol{\beta} = \psi^{-1}(\boldsymbol{y})$. Then,

$$\|\boldsymbol{y}\|_2 = \sqrt{n} \quad \Rightarrow \quad \beta_1^2 + \cdots + \beta_n^2 = n \tag{33}$$

$$(31) + (32) \quad \Rightarrow \quad \mu_1 \beta_1^2 + \cdots + \mu_n \beta_n^2 = 1. \tag{34}$$

Equation (33) represents a hyper-ball in $\mathbb{R}^n_\beta$, and Equation (34) represents an irrotational ellipsoid in $\mathbb{R}^n_\beta$ since $\mu_1, \ldots, \mu_n$ are distinct eigenvalues. As a result, given any other $\boldsymbol{\beta}' = (\beta'_1, \ldots, \beta'_n)^\top$ satisfying (33) and (34), there exist three disjoint index sets $\mathcal{J}_+, \mathcal{J}_-, \mathcal{J}_0$ such that $\mathcal{J}_+ \cup \mathcal{J}_- \cup \mathcal{J}_0 = \{1, \ldots, n\}$ and

$$\begin{aligned}&\forall j \in \mathcal{J}_+, \beta_j \neq 0, \beta_j + \beta'_j = 0 \\ &\forall j \in \mathcal{J}_-, \beta_j \neq 0, \beta_j - \beta'_j = 0 \\ &\forall j \in \mathcal{J}_0, \beta_j = \beta'_j = 0.\end{aligned}$$

Step 3. For another arbitrarily chosen minimum $\boldsymbol{h}'$ of (4), let $\boldsymbol{y}' = \operatorname{sign}(\boldsymbol{h}')$ and $\boldsymbol{\beta}' = \psi^{-1}(\boldsymbol{y}')$, then $\boldsymbol{\beta}'$ is also a solution to the system of Equations (33) and (34), and it is guaranteed the existence of aforementioned $\mathcal{J}_+, \mathcal{J}_-, \mathcal{J}_0$.

Notice that $\forall j \in \mathcal{J}_+$,

$$\boldsymbol{v}_j^\top (\boldsymbol{y} + \boldsymbol{y}') = \beta_j + \beta'_j = 0 \quad \Rightarrow \quad \boldsymbol{v}_j^\top \boldsymbol{y}' = -\boldsymbol{v}_j^\top \boldsymbol{y}.$$

Similarly, $\forall j \in \mathcal{J}_-, \boldsymbol{v}_j^\top \boldsymbol{y}' = \boldsymbol{v}_j^\top \boldsymbol{y}$ and $\forall j \in \mathcal{J}_0, \boldsymbol{v}_j^\top \boldsymbol{y}' = \boldsymbol{v}_j^\top \boldsymbol{y} = 0$. In a word, we have $(\boldsymbol{v}_j^\top \boldsymbol{y}')^2 = (\boldsymbol{v}_j^\top \boldsymbol{y})^2$ for all $j = 1, \ldots, n$. Hence,

$$\boldsymbol{y}^\top Q \boldsymbol{y} = \sum_{j=1}^n \lambda_j (\boldsymbol{v}_j^\top \boldsymbol{y})^2 = \sum_{j=1}^n \lambda_j (\boldsymbol{v}_j^\top \boldsymbol{y}')^2 = \boldsymbol{y}'^\top Q \boldsymbol{y}',$$

which indicates that $(\boldsymbol{y} + \boldsymbol{y}')^\top Q (\boldsymbol{y} - \boldsymbol{y}') = 0$.

Let $\delta_1 = [\boldsymbol{y}]_1, \cdots, \delta_n = [\boldsymbol{y}]_n$ and $\mathcal{K} = \{k \mid [\boldsymbol{y}]_k = [\boldsymbol{y}']_k, 1 \leq k \leq n\}$. Subsequently, by checking the condition Equation (20) we would find that

$$\left(\sum\nolimits_{k \in \mathcal{K}} \delta_k \boldsymbol{e}_k\right)^\top Q \left(\sum\nolimits_{k \notin \mathcal{K}} \delta_k \boldsymbol{e}_k\right) = \frac{1}{4} (\boldsymbol{y} + \boldsymbol{y}')^\top Q (\boldsymbol{y} - \boldsymbol{y}') = 0.$$

However, X_n is SI-asymmetric and thus there must be $\#\mathcal{K} = 0$ or $\#\mathcal{K} = n$, that is, $\boldsymbol{y}' = -\boldsymbol{y}$ or $\boldsymbol{y}' = \boldsymbol{y}$. Therefore, $d_{\mathcal{H}}(\boldsymbol{h}, \boldsymbol{h}') = 0$ and $\boldsymbol{h}'$ is equivalent to $\boldsymbol{h}$. ■

Appendix B. Proof of Lemma 18

For any $\boldsymbol{h} \in \mathcal{H}'_Q$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\boldsymbol{h} = U\boldsymbol{\alpha}$, where U consists of n orthonormal eigenvectors of Q, and $\|\boldsymbol{\alpha}\|_2 = 1$ since $\|\boldsymbol{h}\|_2 = 1$ and $U^\top U = I$. The expression $\boldsymbol{h} = U\boldsymbol{\alpha}$ is an unlabeled-labeled representation (ULR) since U only has the information about unlabeled samples extracted from Q. Each column of U has a unit length, and thus $\|U\|_F^2 = n$ where $\|\cdot\|_F$ is the Frobenius norm. The first part of the upper bound, namely,

$$\mathcal{R}_u(\mathcal{H}'_Q) \leq \sqrt{2n/n'(n-n')},$$

comes from Equations (20)–(22) of El-Yaniv and Pechyony (2009).

Let $\hat{Q} = (\gamma Q - \eta^* I_n)^{-1}$. Another ULR is shown in Equation (30), in the proof of Theorem 13:

$$\boldsymbol{h} = \hat{Q}\,\mathrm{sign}(\boldsymbol{h}).$$

It is clear that $1/(\gamma\lambda_1 - \eta^*), \ldots, 1/(\gamma\lambda_n - \eta^*)$ are the eigenvalues of $\hat{Q}$ given that $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of Q. Subsequently, the second part of the upper bound, that is,

$$\mathcal{R}_u(\mathcal{H}'_Q) \leq \sqrt{\frac{2}{n'(n-n')}} \left(\sum_{i=1}^{n} \frac{n}{(\gamma\lambda_i - \eta^*)^2} \right)^{1/2},$$

can be derived from Equations (20)–(22) of El-Yaniv and Pechyony (2009) with $\mu_1 = \sqrt{n}$. Furthermore, Equation (30) is also a kernel ULR, since $\hat{Q}$ is symmetric positive definite and can be viewed as a kernel matrix. Thereby we can obtain the third part of the upper bound

$$\mathcal{R}_u(\mathcal{H}'_Q) \leq \sqrt{\frac{2}{n'(n-n')}} \left(\sum_{i=1}^{n} \frac{\mu}{\gamma\lambda_i - \eta^*} \right)^{1/2}$$

based on Equations (23)–(25) of El-Yaniv and Pechyony (2009) with $\mu_2 = \sqrt{\mu}$. ■

References

F. Agakov and D. Barber. Kernelized infomax clustering. In *Advances in Neural Information Processing Systems 18 (NIPS)*, 2006.

D. Arthur and S. Vassilvitskii. How slow is the k-means method? In *Proceedings of 22nd ACM Symposium on Computational Geometry (SoCG)*, 2006.

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 (NIPS)*, 2002.

S. Ben-David, D. Pál, and H. U. Simon. Stability of *k*-means clustering. In *Proceedings of 20th Annual Conference on Learning Theory (COLT)*, 2007.

P. T. Boggs and J. W. Tolle. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1995.

B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of 5th Annual Workshop on Computational Learning Theory (COLT)*, 1992.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems 16 (NIPS)*, 2004.

T. De Bie and N. Cristianini. Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems. *Journal of Machine Learning Research*, 7:1409–1436, 2006.

C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of 21st International Conference on Machine Learning (ICML)*, 2004.

R. El-Yaniv and D. Pechyony. Transductive Rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

R. El-Yaniv, D. Pechyony, and V. Vapnik. Large margin vs. large volume in transductive learning. *Machine Learning*, 72(3):173–188, 2008.

L. Faivishevsky and J. Goldberger. A nonparametric information theoretic clustering algorithm. In *Proceedings of 27th International Conference on Machine Learning (ICML)*, 2010.

M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3):780–784, 2002.

R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems 23 (NIPS)*, 2010.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21 (web page and software). http://cvxr.com/cvx, 2011.

J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

G. R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

Y. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou. Tighter and convex maximum margin clustering. In *Proceedings of 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

M. Meila and J. Shi. A random walks view of spectral segmentation. In *Proceedings of 8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2001.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

A. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 (NIPS)*, 2002.

G. Niu, B. Dai, L. Shang, and M. Sugiyama. Maximum volume clustering. In *Proceedings of 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

P. Raghavan and C. Thompson. Randomized rounding: A technique for provably good algorithms and algorithmic proofs. Technical Report UCB/CSD-85-242, UC Berkeley, 1985.

A. Rakhlin and A. Caponnetto. Stability of k-means clustering. In *Advances in Neural Information Processing Systems 19 (NIPS)*, 2007.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423 & 623–656, 1948.

H. Sherali and W. Adams. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Kluwer Academic Publishers, 1998.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

A. Smola, S.V.N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Proceedings of 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

L. Song, A. Smola, A. Gretton, and K. Borgwardt. A dependence maximization view of clustering. In *Proceedings of 24th International Conference on Machine Learning (ICML)*, 2007.

M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In *Proceedings of 28th International Conference on Machine Learning (ICML)*, 2011.

T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *JMLR Workshop and Conference Proceedings*, volume 4, pages 5–20, 2008.

T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

H. Valizadegan and R. Jin. Generalized maximum margin clustering and unsupervised kernel learning. In *Advances in Neural Information Processing Systems 19 (NIPS)*, 2007.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In *Advances in Neural Information Processing Systems 17 (NIPS)*, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

U. von Luxburg, R. Williamson, and I. Guyon. Clustering: Science or art? In *JMLR Workshop and Conference Proceedings*, volume 27, pages 65–80, 2012.

F. Wang, B. Zhao, and C. Zhang. Linear time maximum margin clustering. *IEEE Transactions on Neural Networks*, 21(2):319–332, 2010.

L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. In *Proceedings of 20th National Conference on Artificial Intelligence (AAAI)*, 2005.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems 17 (NIPS)*, 2005.

A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17 (NIPS)*, 2005.

H. Zha, X. He, C. Ding, M. Gu, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems 14 (NIPS)*, 2002.

K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. In *Proceedings of 24th International Conference on Machine Learning (ICML)*, 2007.

B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. In *Proceedings of 25th International Conference on Machine Learning (ICML)*, 2008a.

B. Zhao, F. Wang, and C. Zhang. Efficient maximum margin clustering via cutting plane algorithm. In *Proceedings of 8th SIAM International Conference on Data Mining (SDM)*, 2008b.