Maximum Volume Clustering

Gang Niu[1,2]
gang@sg.cs.titech.ac.jp

Bo Dai[3]
bdai@nlpr.ia.ac.cn

Lin Shang[2]
shanglin@nju.edu.cn

Masashi Sugiyama[1]
sugi@cs.titech.ac.jp

[1]Department of Computer Science, Tokyo Institute of Technology
[2]State Key Laboratory for Novel Software Technology, Nanjing University
[3]NLPR/LIAMA, Institute of Automation, Chinese Academy of Science

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 clustering model based on the large volume principle called *maximum volume clustering* (MVC), and propose two algorithms to solve it approximately: a soft-label and a hard-label MVC algorithms based on sequential quadratic programming and semi-definite programming, respectively. Our MVC model includes spectral clustering and maximum margin clustering as special cases, and is substantially more general. We also establish the finite sample stability and an error bound for soft-label MVC method. Experiments show that the proposed MVC approach compares favorably with state-of-the-art clustering algorithms.

1 Introduction

Clustering has been an important topic in machine learning and data mining communities. Over past decades, a large number of clustering algorithms have been developed, including *k-means clustering* (Hartigan & Wong, 1979), *spectral clustering* (Shi & Malik, 2000; Meila & Shi, 2001; Ng et al., 2001), and *maximum margin clustering* (Xu et al., 2004; Xu & Schuurmans, 2005). They have been successfully applied to diverse real-world exploratory data-analysis tasks.

To the best of our knowledge, the *Maximum Margin Clustering* (MMC) (Xu et al., 2004)—which maximizes the margin between two opposite clusters—is the first clustering algorithm that is directly connected to statistical learning theory (Vapnik, 1998). For this reason, it has been extensively investigated recently, e.g., Generalized MMC (Valizadegan & Jin, 2006) and various approximation algorithms for speedup (Zhang et al., 2007; Zhao et al., 2008a; Zhao et al., 2008b; Li et al., 2009).

However, the *large margin principle* (LMP) is not the only way to go. There is a *large volume principle* (LVP) which was introduced by Vapnik (1982) for hyperplanes and extended by El-Yaniv et al. (2008) for soft response vectors. Roughly speaking, machine learning algorithms based on LVP should prefer hypotheses in some large-volume equivalence classes. See Figure 1 as an example. Here C_1, C_2 and C_3 represent data clouds, and we want to choose a better separating hyperplane from $\mathbf{h}_1$ and $\mathbf{h}_2$. Though LMP prefers $\mathbf{h}_1$ due to its large margin, we should choose $\mathbf{h}_2$ when considering LVP, since its equivalence class (a set of hyperplanes which equivalently separate samples) has a larger volume than $\mathbf{h}_1$'s.

In this paper, we propose a novel model for clustering called *Maximum Volume Clustering* (MVC), which serves as a prototype partitioning the data into two clusters based on LVP. Given the samples X_n, we construct an X_n-dependent hypothesis space $\mathcal{H}$. If there is a measure on $\mathcal{H}$, namely the *power* (Vapnik, 1998),

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

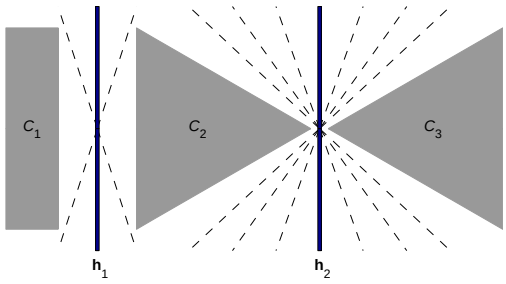

Figure 1: Large margin vs. large volume separation of samples into two clusters.

then we can talk about the *likelihood* or *confidence* of each equivalence class. Similarly to the *margin* 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. Thus we consider the partition lying in the equivalence class with the maximum volume as the best partition.

Similarly to other clustering approaches, the optimization problem involved in our MVC is NP-hard, so we introduce two approximation schemes: a soft-label MVC based on *sequential quadratic programming* (SQP) (Boggs & Tolle, 1995) that can be solved in $O(n^3)$ time (the same as a standard eigenvalue problem), and a hard-label MVC based on *semi-definite programming* (SDP) (De Bie & Cristianini, 2003) that can be solved by any standard SDP solver in $O(n^{6.5})$ time (the same as the original MMC). Moreover, we show that these two approximations can be reduced to spectral clustering and MMC in special cases. Hence the proposed MVC model may be regarded as a natural extension of existing spectral and large margin approaches. We also establish the finite sample stability and an error bound for soft-label MVC method. Experiments on benchmarks show that the proposed MVC approach is promising.

The rest of this paper is organized as follows. In Section 2 we briefly introduce a large volume approximation for clustering to be used in MVC. In Section 3, we present the MVC model, algorithms and theoretical analyses. A comparison with related works is made in Section 4. Experimental results are shown in Section 5, and finally we conclude in Section 6.

2 A Large Volume Approximation

Given a set of samples $X_n = \{x_1, \ldots, x_n\}$, where $x_i \in \mathcal{X}$ (most often but not necessarily, $\mathcal{X} \subset \mathbb{R}^d$ for some positive integer d), we will construct a hypothesis space $\mathcal{H}$ that depends on X_n, such that for any hypothesis $\mathbf{h} \in \mathcal{H} \subset \mathbb{R}^n$, $[\mathbf{h}]_i$ (the i-th component of $\mathbf{h}$) stands for a soft label or a confidence-rated label of x_i (El-Yaniv et al., 2008). We will then select an $\mathbf{h} \in \mathcal{H}$ following LVP, and the two resulting clusters will be $\{x_i \mid [\mathbf{h}]_i > 0\}$ and $\{x_i \mid [\mathbf{h}]_i < 0\}$.

As El-Yaniv et al. (2008), assume we have an $n \times n$ positive definite matrix Q that contains pairwise information about X_n. Consider the hypothesis space $\mathcal{H}_Q = \{\mathbf{h} \mid \mathbf{h}^\top Q \mathbf{h} \leq 1\}$, which is geometrically an origin-centered ellipsoid $\mathcal{E}(\mathcal{H}_Q)$ in $\mathbb{R}^n$. The set of sign vectors $\{\mathrm{sgn}(\mathbf{h}) \mid \mathbf{h} \in \mathcal{H}_Q\}$ contains all 2^n possible dichotomies of X_n. In other words, $\mathcal{H}_Q$ is now partitioned into a finite number of *equivalence classes* $H_1, \ldots, H_{2^n}$, such that for fixed $k \in \{1, \ldots, 2^n\}$, all hypotheses in H_k will generate the same dichotomy of X_n. The *power* of an equivalence class H_k (Vapnik, 1998) is defined as a probability mass

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

where $P(\mathbf{h})$ is the underlying distribution of $\mathbf{h}$ 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., $P(\mathbf{h})$ is unknown), it would be natural to assume the uniform distribution for $\mathbf{h}$ over $\mathcal{H}_Q$. Consequently, $\mathcal{P}(H_k)$ is proportional to H_k's *geometric volume*

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

Therefore, the larger $\mathcal{V}(H_k)$ is, the more confident we are of the partition $\mathrm{sgn}(\mathbf{h})$ where $\mathbf{h}$ is chosen from H_k. Note that each quadrant in $\mathbb{R}^n$ intersects with one equivalence class, and $\mathcal{V}(H_k)$ is also given by the geometric volume of the k-th quadrant of $\mathcal{E}(\mathcal{H}_Q)$.

However, it is too hard to compute $\mathcal{V}(H_k)$ exactly for all $k \in \{1, \ldots, 2^n\}$, so we employ an efficient approximation scheme introduced by El-Yaniv et al. (2008). Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of Q, and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the associated normalized eigenvectors. The direction and length of the i-th principal axis of $\mathcal{E}(\mathcal{H}_Q)$ are $\mathbf{v}_i$ and $1/\sqrt{\lambda_i}$ respectively. Then a small angle from $\mathbf{h} \in H_k$ to $\mathbf{v}_i$ 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 observation, we define

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

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

3 Maximum Volume Clustering

In this section, we will define our MVC model, derive practical algorithms and give theoretical analyses.

3.1 Basic Formulation

Motivated by MMC (Xu et al., 2004), we first formulate the clustering problems from the regularization viewpoint. If we have labels $Y_n = \{y_1, \ldots, y_n\}$ at hand where $y_i \in \{-1, +1\}$, we can find a base algorithm $\mathscr{B}$ to compute

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

where Δ is the overall loss, W is a regularizer, γ is a regularization parameter, and the hypothesis space $\mathcal{H}$ is dependent upon X_n and Y_n. The value $\vartheta(X_n, Y_n)$ is a measure of classification quality.

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

$$\mathbf{y}' = \arg\min_{\mathbf{y}\in\{-1,+1\}^n} \vartheta(X_n, \mathbf{y}).$$

Generally speaking, the value $\vartheta(X_n, \mathbf{y}')$ can be viewed as a measure of clustering quality. The smaller the value $\vartheta(X_n, \mathbf{y}')$ is, the more satisfied we are with the resulting data partition $\operatorname{sgn}(\mathbf{y}')$.

In MVC, we hope to utilize (1) as our regularizer. Formally, given the matrix Q, by instantiating Δ to the linear loss function $-2\mathbf{h}^\top\mathbf{y}$, we define the *Maximum Volume Clustering* model as

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

where $\mathcal{H}_Q = \{\mathbf{h} \mid \mathbf{h}^\top Q\mathbf{h} \le 1\}$ is the hypothesis space, and $\gamma > 0$ is a regularization parameter. Optimization (2) is computationally intractable, because not only of the non-convexity of $V(\mathbf{h})$ but also of the integer feasible region of $\mathbf{y} \in \{-1, +1\}^n$. Next we will discuss two approximation schemes of (2) in detail.

3.2 Soft-Label Approximation

Now we try to optimize $\mathbf{h}$ alone by eliminating $\mathbf{y}$. After changing the order of $\min_{\mathbf{y}}$ and $\min_{\mathbf{h}}$ in (2), we see that $\mathbf{y}$ should be $\operatorname{sgn}(\mathbf{h})$, since the second term is independent of $\mathbf{y}$, and the first term is minimized when $\mathbf{y} = \operatorname{sgn}(\mathbf{h})$ for fixed $\mathbf{h}$. Therefore, (2) becomes

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

Clearly $\mathbf{h}^\top Q\mathbf{h}/\|\mathbf{h}\|_2^2$ equals $\mathbf{h}^\top Q\mathbf{h}$ under the condition $\|\mathbf{h}\|_2 = 1$, and then (3) can be expressed as

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

This is the primal problem of *Soft-Label Maximum Volume Clustering* (SL-MVC).

In order to solve (4), we resort to *sequential quadratic programming* (SQP) (Boggs & Tolle, 1995). Let[1]

$$\begin{aligned} f(\mathbf{h}) =& -2\mathbf{h}^\top\operatorname{sgn}(\mathbf{h}) + \gamma\mathbf{h}^\top Q\mathbf{h}, \\ f_1(\mathbf{h}) =& \mathbf{h}^\top\mathbf{h} - 1, \quad f_2(\mathbf{h}) = \mathbf{h}^\top\mathbf{1}, \\ L(\mathbf{h}, \eta) =& -2\mathbf{h}^\top\operatorname{sgn}(\mathbf{h}) + \gamma\mathbf{h}^\top Q\mathbf{h} - \eta\left(\mathbf{h}^\top\mathbf{h} - 1\right) \\ & - \mu\left(\mathbf{h}^\top\mathbf{1} - b\right) + \nu\left(\mathbf{h}^\top\mathbf{1} + b\right), \end{aligned}$$

where we include a class balance constraint $|\mathbf{h}^\top\mathbf{1}| \le b$ with a class balance parameter $b > 0$, $\eta \in \mathbb{R}$, $\mu \ge 0$, and $\nu \ge 0$ are Lagrangian multipliers for $\|\mathbf{h}\|_2 = 1$ and $|\mathbf{h}^\top\mathbf{1}| \le b$. Then, according to the problem

$$\begin{aligned} \min_{\mathbf{p}\in\mathbb{R}^n} \ & \frac{1}{2}\mathbf{p}^\top\nabla^2 L(\mathbf{h}, \eta)\mathbf{p} + \mathbf{p}^\top\nabla f(\mathbf{h}) \\ \text{s.t.} \ & \mathbf{p}^\top\nabla f_1(\mathbf{h}) + f_1(\mathbf{h}) = 0, \ \left|\mathbf{p}^\top\nabla f_2(\mathbf{h}) + f_2(\mathbf{h})\right| \le b, \end{aligned}$$

at the t-th iteration, the subproblem at the current solution $(\mathbf{h}_t, \eta_t)$ is a quadratic programming

$$\begin{aligned} \min_{\mathbf{p}_t\in\mathbb{R}^n} \ & \mathbf{p}_t^\top(\gamma Q - \eta_t I)\mathbf{p}_t + 2\mathbf{p}_t^\top\big(\gamma Q\mathbf{h}_t - \operatorname{sgn}(\mathbf{h}_t)\big) \\ \text{s.t.} \ & 2\mathbf{p}_t^\top\mathbf{h}_t + \mathbf{h}_t^\top\mathbf{h}_t = 1, \ \left|\mathbf{p}_t^\top\mathbf{1} + \mathbf{h}_t^\top\mathbf{1}\right| \le b, \end{aligned} \tag{5}$$

and the new η is given by

$$\eta_{t+1} = \big(\gamma Q\mathbf{h}_{t+1} - \eta_t\mathbf{p}_t - \operatorname{sgn}(\mathbf{h}_t)\big)^\top\mathbf{h}_t / \left(\mathbf{h}_t^\top\mathbf{h}_t\right). \tag{6}$$

The SL-MVC algorithm based on SQP is summarized in *Algorithm 1*. We use an initial solution $\mathbf{h}_0 = \operatorname{sgn}(\mathbf{v}_2 - \frac{1}{n}\mathbf{1}^\top\mathbf{v}_2\mathbf{1})/\|\operatorname{sgn}(\mathbf{v}_2 - \frac{1}{n}\mathbf{1}^\top\mathbf{v}_2\mathbf{1})\|_2$ and $\eta_0 = -0.001$ in our experiments, where $\mathbf{v}_2$ is the eigenvector corresponding to the second smallest eigenvalue of Q. The asymptotic time-complexity of each subproblem is $O(n^3)$, and SQP converges in $O(1)$ iterations. Likewise it takes $O(n^3)$ time to compute the initial solution $\mathbf{h}_0$. Thus the overall computational complexity of *Algorithm 1* is $O(n^3)$.

Additionally we have,

Theorem 1. *Spectral clustering could be derived from SL-MVC when the number of clusters is 2.*

Proof. Recall the relaxed RatioCut that formulates spectral clustering from the graph cut point of view when there are two clusters (von Luxburg, 2007),

$$\min_{f\in\mathbb{R}^n} f^\top Lf \quad \text{s.t. } f\perp\mathbf{1}, \ \|f\|_2 = \sqrt{n}, \tag{7}$$

where L is the *unnormalized graph Laplacian*. Obviously (7) can be derived from (4) by setting $Q = L, \gamma = \infty$ and a strict class balance constraint $\mathbf{h}^\top\mathbf{1} = 0$. □

[1] Note that the term $-\mathbf{h}^\top\mathbf{y}$ or $-\|\mathbf{h}\|_1$ combined with $\min_{\mathbf{h}}$ has an effect to push $\mathbf{h}$ away from the coordinate axes of $\mathbb{R}^n$. Thus $h_i = 0$ hardly happens in practice and we assume that $\|\mathbf{h}\|_1$ is always differentiable.

Algorithm 1 SL-MVC (SQP ver.)

Input: stop criterion ϵ,
positive definite matrix Q,
regularization parameter γ,
class balance parameter b
Output: soft response vector $\mathbf{h}_{t+1}$
Initialize $\mathbf{h}_0$ and η_0
Let $t = -1$
repeat
$t = t + 1$
Optimize (5) to obtain $\mathbf{p}_t$
Update $\mathbf{h}_{t+1} = \mathbf{h}_t + \mathbf{p}_t$
Update η_{t+1} through (6)
until $\|\mathbf{h}_{t+1} - \mathbf{h}_t\|_2^2 + \|\eta_{t+1} - \eta_t\|_2^2 \leq \epsilon$

3.3 Hard-Label Approximation

Similarly to the soft-label case, we can optimize $\mathbf{y}$ alone. Let $\mathbf{h} = \boldsymbol{\alpha} \circ \mathbf{y}$, where $y_i = \operatorname{sgn}(h_i), \alpha_i = |h_i|$ and $\circ$ denotes the element-wise product. After the introduction of a parameter C that may work against outliers, the primal problem of *Hard-Label Maximum Volume Clustering* (HL-MVC) is rewritten as

$$\begin{aligned} \min_{\mathbf{y} \in \{-1,+1\}^n} \min_{\boldsymbol{\alpha}} \; & -2\boldsymbol{\alpha}^\top \mathbf{1} + \gamma \boldsymbol{\alpha}^\top (Q \circ \mathbf{y}\mathbf{y}^\top)\boldsymbol{\alpha} \\ \text{s.t.} \; & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \; \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1. \end{aligned} \tag{8}$$

By employing the techniques in De Bie and Cristianini (2003), let $M = \mathbf{y}\mathbf{y}^\top$ and then (8) is relaxed to

$$\begin{aligned} \min_{M \in \mathbb{R}^{n \times n}} \max_{\boldsymbol{\alpha}, \eta} \; & 2\boldsymbol{\alpha}^\top \mathbf{1} - \gamma \boldsymbol{\alpha}^\top (Q \circ M)\boldsymbol{\alpha} + \eta \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \eta \\ \text{s.t.} \; & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \; M \succeq 0, \; \operatorname{diag}(M) = \mathbf{1}, \end{aligned} \tag{9}$$

where $\succeq$ indicates the positive semi-definiteness of a matrix. The relaxation mainly comes from replacing the non-convex constraint $\operatorname{rank}(M) = 1$ with a convex alternative $M \succeq 0$. Actually, (9) is a *semi-definite programming* (SDP) provided $(\gamma Q \circ M - \eta I) \succeq 0$. Let $\boldsymbol{\nu} \geq \mathbf{0}$ and $\boldsymbol{\mu} \geq \mathbf{0}$ be Lagrangian multipliers for $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\alpha} \leq C\mathbf{1}$. Due to the convexity of (9) when $(\gamma Q \circ M - \eta I) \succeq 0$ holds, (9) is equivalent to

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

Consider the variable $\boldsymbol{\alpha}$ alone. The optimal $\boldsymbol{\alpha}$ should be $\boldsymbol{\alpha} = (Q \circ M - \eta I)^{-1}(\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu})$, and we can form the above problem as

$$\begin{aligned} \min_{M, \boldsymbol{\mu}, \boldsymbol{\nu}, \eta} \; & (\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu})^\top (Q \circ M - \eta I)^{-1} (\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu}) \\ & + 2C\boldsymbol{\mu}^\top \mathbf{1} - \eta \\ \text{s.t.} \; & \boldsymbol{\mu} \geq \mathbf{0}, \; \boldsymbol{\nu} \geq \mathbf{0} \\ & M \in \mathbb{R}^{n \times n}, \; M \succeq 0, \; \operatorname{diag}(M) = \mathbf{1}, \end{aligned}$$

under an additional condition that $\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu}$ is orthogonal to the null space of $Q \circ M - \eta I$. Eventually, we arrive at a formulation in standard SDP form:

$$\begin{aligned} \min_{M, \boldsymbol{\mu}, \boldsymbol{\nu}, \eta, t} \; & t \\ \text{s.t.} \; & \boldsymbol{\mu} \geq \mathbf{0}, \; \boldsymbol{\nu} \geq \mathbf{0} \\ & M \in \mathbb{R}^{n \times n}, \; M \succeq 0, \; \operatorname{diag}(M) = \mathbf{1} \\ & \begin{pmatrix} \gamma Q \circ M - \eta I & (\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu}) \\ (\mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu})^\top & t + \eta - 2C\boldsymbol{\mu}^\top \mathbf{1} \end{pmatrix} \succeq 0. \end{aligned} \tag{10}$$

The computational complexity of (10) is $O(n^{6.5})$ by a standard SDP solver. It could be reduced to $O(n^{4.5})$ with the *subspace tricks* (De Bie & Cristianini, 2003; De Bie & Cristianini, 2006) which use essentially the spectral properties of Q to control the trade off between the computational cost and the accuracy.

When we get the optimal M, the optimal $\mathbf{y}$ can be recovered from M by techniques such as *randomized rounding* (Raghavan & Thompson, 1985). In our implementation, we first extract the eigenvector $\bar{\mathbf{v}}$ associated with the largest eigenvalue of M, and the optimal $\mathbf{y}$ is then recovered as $\mathbf{y} = \operatorname{sgn}(\bar{\mathbf{v}} - \frac{1}{n}\mathbf{1}^\top \bar{\mathbf{v}} \mathbf{1})$.

Additionally we have,

Theorem 2. *Maximum margin clustering could be derived from HL-MVC.*

Proof. Let $Q = K, \gamma = 1$ and fix $\eta = 0$, then the dual problem of HL-MVC coincides with the SDP dual of MMC exactly (without class balance constraints). □

3.4 Finite Sample Stability

In this subsection, we investigate the finite sample stability of SL-MVC method. Stability of the resulting clusters is especially important for those solved by randomized algorithms (e.g., SL-MVC and k-means clustering) rather than by transforming themselves to convex dual problems (e.g., HL-MVC, and MMC in Xu et al. (2004) and Valizadegan and Jin (2006)). In the following, we presume that we are able to find local minima of (4) accurately. Under this presumption, we prove that the instability of SL-MVC is solely resulted from the symmetry of samples, that is, we cannot deduce unstable clusters from any asymmetric samples. The proofs are omitted here due to limited space.

To begin with, given a constant η we define (remember the assumption that $\|\mathbf{h}\|_1$ is differentiable)

$$G(\mathbf{h}) = \gamma \mathbf{h}^\top Q \mathbf{h} - \eta \|\mathbf{h}\|_2^2 - 2\|\mathbf{h}\|_1,$$

$$g(\mathbf{h}) = \frac{1}{2}\nabla G(\mathbf{h}) = \gamma Q \mathbf{h} - \eta \mathbf{h} - \operatorname{sgn}(\mathbf{h}).$$

Definition 1. The Hamming clustering distance for two soft response vectors in $\mathbb{R}^n$ is defined as

$$d_{\mathcal{H}}(\mathbf{h}_1, \mathbf{h}_2) = \tfrac{1}{2} \min(\|\mathbf{y}_1 + \mathbf{y}_2\|_1, \|\mathbf{y}_1 - \mathbf{y}_2\|_1),$$

where $\mathbf{y}_1 = \text{sgn}(\mathbf{h}_1)$ and $\mathbf{y}_2 = \text{sgn}(\mathbf{h}_2)$.

Definition 2. We say that X_n is an axisymmetric set of samples w.r.t. Q, if there exists a bijection $\phi : \{1, \ldots, n\} \mapsto \{1, \ldots, n\}$ such that (a) $\exists i, \phi(i) \neq i$, (b) $\forall i, \phi^{-1}(i) = \phi(i)$ and (c) $Q_{i,\phi(k)} = Q_{\phi(i),k}$ for all $1 \leq i, k \leq n$.

In other words, ϕ divides X_n into subsets of single element (i.e., $\{x_i\}$ if $\phi(i) = i$) or paired elements (i.e., $\{x_i, x_j\}$ if $\phi(i) = j, \phi(j) = i$), and for each x_i and x_j in the latter case, they cannot be distinguished by all other samples as a whole based on the information of Q, so they can exchange with each other without changing Q. There might be more than one eligible ϕ. In fact, the axisymmetry of X_n w.r.t. Q is equivalent to the geometric axisymmetry of X_n in $\mathcal{X} \subset \mathbb{R}^d$ if Q is constructed from the *Euclidean distance* defined on $\mathbb{R}^d$. For example, $X_4 = \{(0,0), (1,0), (1,1), (0,1)\}$, which is axisymmetric in $\mathbb{R}^2$, is an axisymmetric set of samples in the sense of Gaussian similarity.

It is, however, not true that a symmetric X_n must result in unstable best partition. It occurs only when the best partition is not unique. For example, $X_4' = \{(0,0), (2,0), (2,1), (0,1)\}$ has the unique best partition $(1, -1, -1, 1)$, but intuitively X_4 has two best partitions $(1, -1, -1, 1)$ and $(1, 1, -1, -1)$. Formally,

Theorem 3 (**Twin Minimum Theorem**). *Assume that $n > 2$, X_n is an axisymmetric set of samples w.r.t. Q, and $\mathcal{I} = \{\{i, j\} \mid x_i, x_j$ cannot be distinguished by $X_n \setminus \{x_i, x_j\}\}$. For every minimum $\mathbf{h}^*$ of (4), if $\forall i, [\mathbf{h}^*]_i \neq 0$ and $\exists \{i, j\} \in \mathcal{I}$, $[\text{sgn}(\mathbf{h}^*)]_i \neq [\text{sgn}(\mathbf{h}^*)]_j$, then $\mathbf{h}^*$ has a twin minimum $\mathbf{h}^\star$ satisfying $G(\mathbf{h}^\star) = G(\mathbf{h}^*)$ and $d_{\mathcal{H}}(\mathbf{h}^*, \mathbf{h}^\star) \geq 1$. The only exception is that there exists $\mathcal{I}' \subseteq \mathcal{I}$, such that $\mathcal{I}'$ consists of disjoint index pairs, all indices in $\mathcal{I}'$ cover $\{1, \ldots, n\}$, and $\forall \{i, j\} \in \mathcal{I}', [\text{sgn}(\mathbf{h}^*)]_i \neq [\text{sgn}(\mathbf{h}^*)]_j$.*

In order to explain the implication of above theorem, consider X_4' again. Minima corresponding to $\mathbf{y} = (1, -1, -1, 1)$ have no twin minimum, since $\forall \{i, j\} \in \mathcal{I} = \{\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$, we have $[\mathbf{y}]_i = [\mathbf{y}]_j$ or $[\mathbf{y}]_i \neq [\mathbf{y}]_j, [\mathbf{y}]_{i'} \neq [\mathbf{y}]_{j'}$ where $\{i', j'\} = \{1, \ldots, 4\} \setminus \{i, j\}$, which says that if we switch entries of $\mathbf{h}^*$ according to ϕ, it is still $d_{\mathcal{H}}(\mathbf{h}^*, \mathbf{h}^\star) = 0$. Otherwise, we get $d_{\mathcal{H}}(\mathbf{h}^*, \mathbf{h}^\star) \geq 1$.

However, *Theorem 3* gives only a sufficient condition. The minima corresponding to $\mathbf{y}$ and $(1, 1, -1, -1)$ are twins when considering X_4, which is also an unstable input for spectral clustering and MMC. The difference between X_4 and X_4' is explained below.

Definition 3. We say that X_n is an anisotropic set of samples w.r.t. Q, if Q has n distinct eigenvalues.

The name "anisotropic" comes from a geometric interpretation of the ellipsoid $\mathcal{E}(\mathcal{H}_Q)$. When Q has n distinct eigenvalues, principal axes of $\mathcal{E}(\mathcal{H}_Q)$ have different length and $\mathcal{E}(\mathcal{H}_Q)$ is irrotational about any principal axis. The concepts of anisotropy and axisymmetry are not complementary. Indeed some X_n is only axisymmetric or anisotropic, but some X_n is both axisymmetric and anisotropic, e.g., X_4' in the sense of Gaussian similarity, and some X_n can be neither axisymmetric nor anisotropic w.r.t. Q. Nevertheless, when considering certain families of Q like Gaussian kernel matrices, we have this relationship between axisymmetry and anisotropy, if a set X_n is not axisymmetric then X_n must be anisotropic w.r.t. this Q.

Theorem 4. *If X_n is not an axisymmetric set of samples w.r.t. Q, and there exists $\kappa > 0$ such that $\forall i, Q_{i,i} = \kappa$, then X_n is an anisotropic set of samples w.r.t. Q.*

Theorem 5 (**Equivalent Minima Theorem**). *All minima of (4) are equivalent w.r.t. $d_{\mathcal{H}}$ provided X_n is an anisotropic set of samples w.r.t. Q.*

Though *Algorithm 1* is stable for fixed $\mathbf{h}_0$ and η_0 since that algorithm is derandomized by the initial solution, we have a stronger result immediately from *Theorem 5*.

Corollary 6. *If X_n is an anisotropic set of samples w.r.t. Q, then the optimization problem (4) will always lead to the same partition of X_n.*

To sum up, the finite sample stability of SL-MVC is theoretically as strong as MMC with the convex SDP formula. Things will become complicated if we consider numerical issues such as $\|\mathbf{h}\|_2 \approx 1$ and $g(\mathbf{h}) \approx \mathbf{0}$ when *Algorithm 1* stops. As a consequence, there are algorithmically very special anisotropic sets with more than one best partition. For instance, $X_5' = X_4' \cup \{(1, 0.5)\}$, which is an unstable input for all aforementioned clustering algorithms without an assumption that $\text{sgn}(0) = 1$ or $\text{sgn}(0) = -1$.

3.5 Data-Dependent Error Bound

Moreover, we give a data-dependent error bound of *Algorithm 1* using *transductive Rademacher complexity* (El-Yaniv & Pechyony, 2007). It is argued that, no matter how hard it is to evaluate clustering in an objective and domain-independent way, when our goal is clear and a proper similarity measure is chosen, it makes sense to evaluate clustering on certain classification datasets, if the underlying assumption that points with the same class labels form clusters is true (Guyon et al., 2009).

In practical clustering tasks, we often find some experts to label a small portion of samples $X_{n'}$ according to their knowledge, run a pool of candidate clustering algorithms, see their agreement with the labels and eliminate results of those low agreement algorithms.

This may be viewed as propagating the knowledge of experts from $X_{n'}$ to X_n. Here, we present a data-dependent error bound to guarantee the quality of this propagation.

Lemma 7. *Let $\widetilde{\mathcal{H}}_Q$ be the set of all possible* $\mathbf{h}$ *returned by* Algorithm 1 *for a given Q, η^* be the maximal η when* Algorithm 1 *stops, $\{\hat{\lambda}_i\}_{i=1}^n$ be the eigenvalues of $\hat{Q} = (\gamma Q - \eta^* I)^{-1}$, and $\mu = \sup_{\mathbf{h}\in\widetilde{\mathcal{H}}_Q} \operatorname{sgn}(\mathbf{h})^\top \hat{Q} \operatorname{sgn}(\mathbf{h})$. Then, for the transductive Rademacher complexity $\mathcal{R}(\widetilde{\mathcal{H}}_Q)$, the following upper bound holds for any integer n' between 0 and n,*

$$\mathcal{R}(\widetilde{\mathcal{H}}_Q) \leq \sqrt{\frac{2}{n'(n-n')}} \min\left\{\sqrt{n}, \left(n\sum_{i=1}^n \hat{\lambda}_i^2\right)^{\frac{1}{2}}, \left(\mu\sum_{i=1}^n \hat{\lambda}_i\right)^{\frac{1}{2}}\right\}.$$

Use *Lemma 7* in conjunction with *Theorem 2* of (El-Yaniv & Pechyony, 2007) and obtain

Theorem 8. *Assume that we know the ground truth partition on X_n (denoted by $\mathbf{y}^*$), and $\mathcal{L}$ is selected uniformly over $\{\mathcal{L} \mid \mathcal{L} \subset \{1,\ldots,n\}, |\mathcal{L}| = n'\}$. Let ℓ be the 0/1 loss function, $\widetilde{\mathcal{H}}_Q$ be the set of all possible* $\mathbf{h}$ *returned by* Algorithm 1 *for a given Q, $\mathcal{Y} = \{\operatorname{sgn}(\mathbf{h}) \mid \mathbf{h} \in \widetilde{\mathcal{H}}_Q\}$, η^* be the maximal η when* Algorithm 1 *stops, $\{\hat{\lambda}_i\}_{i=1}^n$ be the eigenvalues of $\hat{Q} = (\gamma Q - \eta^* I)^{-1}$, $\mu = \sup_{\mathbf{y}\in\mathcal{Y}} \mathbf{y}^\top \hat{Q} \mathbf{y}$, and $c_0 = \sqrt{\frac{32}{3}\ln(4\mathrm{e})}$. For any $\mathbf{y} \in \mathcal{Y}$, with probability of at least $1-\delta$ over the choice of $\mathcal{L}$, we have*

$$d_{\mathcal{H}}(\mathbf{y}, \mathbf{y}^*) \leq \frac{n}{n'} \min\left\{\sum_{i\in\mathcal{L}} \ell([\mathbf{y}]_i, [\mathbf{y}^*]_i), \sum_{i\in\mathcal{L}} \ell([-\mathbf{y}]_i, [\mathbf{y}^*]_i)\right\} + \frac{c_0 n}{\sqrt{n'}} + \sqrt{\frac{2n(n-n')}{n'}\ln\frac{1}{\delta}} + \sqrt{\frac{2(n-n')}{n'}} \min\left\{\sqrt{n}, \left(n\sum_{i=1}^n \hat{\lambda}_i^2\right)^{\frac{1}{2}}, \left(\mu\sum_{i=1}^n \hat{\lambda}_i\right)^{\frac{1}{2}}\right\}. \tag{11}$$

There are four terms in (11). The first term is the amplified empirical error on the subset of samples $X_{n'} = \{x_i \mid i \in \mathcal{L}\}$. More specifically, we want to choose a proper similarity measure via the given labels to make the empirical error small on $X_{n'}$. The second term depends only on n and n', that is, the size of the whole set and the labeled subset. The next is a term which depends further on the significance level δ as in common error bounds. The last term is the upper bound of $(n-n')\mathcal{R}(\widetilde{\mathcal{H}}_Q)$, which carries out implicitly the complexity control of the hypothesis space, that is, when $\mathcal{R}(\widetilde{\mathcal{H}}_Q)$ is small, we are confident that $d_{\mathcal{H}}(\mathbf{y}, \mathbf{y}^*)$ will be small if the error on $X_{n'}$ is small.

Remark 1. Our setting is equivalent to neither semi-supervised clustering nor transductive classification. We do not reveal any labels to the algorithm, but we do reveal a portion of randomly selected labels to an evaluator which then returns an evaluation of the quality of any possible partition generated by the algorithm. We can use transductive Rademacher complexity for our algorithm since it can be viewed as a transductive algorithm that ignores all revealed labels.

4 Related Works

Among the existing methods, maximum margin clustering algorithms (MMC) are closest to maximum volume clustering (MVC). Although MMC and HL-MVC share similar dual problems, their geneses and underlying criteria are quite different. The primal problems of various MMC use a regularizer $\|\mathbf{w}\|_2^2$ originated from the margin, while MVC uses $V(\mathbf{h})$ originated from the volume. Moreover, the basis of MMC is hyperplanes for induction, whereas the basis of MVC is soft response vectors for transduction.

After the proposition of MMC (Xu et al., 2004), Generalized MMC (GMMC) (Valizadegan & Jin, 2006) has relaxed the restriction that MMC requires the center of samples passing through the origin. There are two fast iterative MMC algorithms based on support vector regression (Zhang et al., 2007) and cutting plane techniques (Zhao et al., 2008a), but both have troubles of local optima. Unlike Zhang et al. (2007) and Zhao et al. (2008a), HL-MVC involves convex optimization, and SL-MVC is proved to perform as a convex optimization under mild conditions, if we only concern the final partition $\mathbf{y}$ rather than the hypothesis $\mathbf{h}$. In a word, the stability of MVC is by no means inferior to non-convex MMC variations. In our experiments, two MVC algorithms run even more stably than GMMC given different candidate parameters, since the latter tries to invert a kernel matrix. What is more, we have a data-dependent error bound, and to the best of our knowledge MMC algorithms have no such result.

Another related work is spectral clustering (SC) (Shi & Malik, 2000; Meila & Shi, 2001; Ng et al., 2001). Many SC algorithms have two steps, first a spectral embedding step and then a k-means step. Note that SL-MVC is able to integrate unnormalized SC into a single optimization and the highly non-convex k-means step is unnecessary. Next, the motivation of $f \perp \mathbf{1}$ in SC is different from $\mathbf{h}^\top \mathbf{1} = 0$ in SL-MVC. When L is constructed from a fully connected similarity graph, $f \perp \mathbf{1}$ means that the feasible region of (7) is included in the space spanned by all eigenvectors of L except the trivial eigenvector $\mathbf{1}$. The last and most vital difference between SC and MVC is that MVC has a linear loss

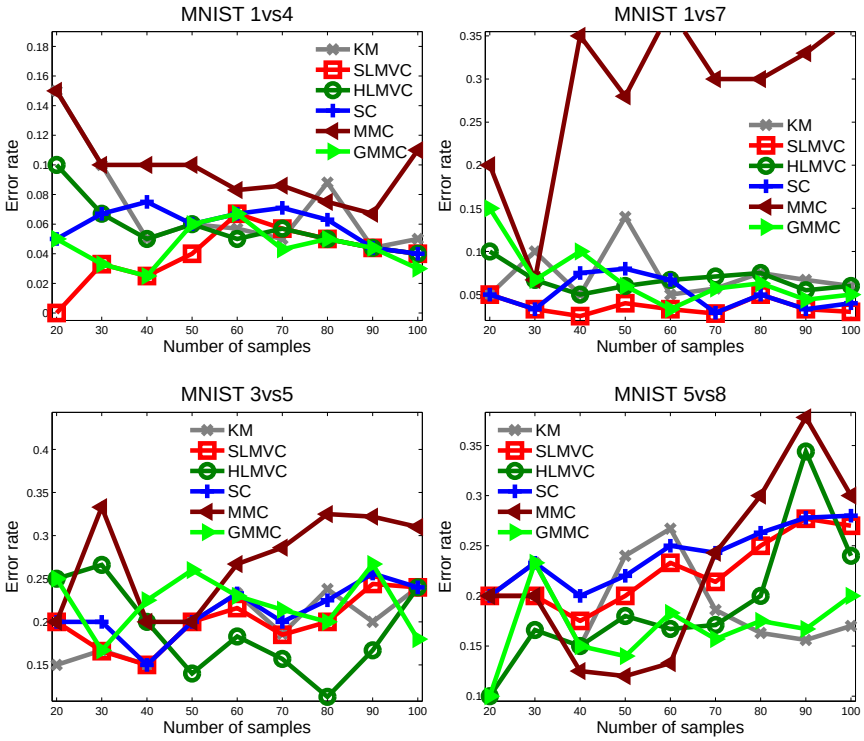

Figure 2: Experimental results on `MNIST` for small n

term which pushes hypotheses away from the coordinate axes and always leads to non-sparse solutions.

The *Approximate Volume Regularization* (AVR) (El-Yaniv et al., 2008) is also strongly connected to MVC. However, the implementations are quite different. In their transductive learning setting, they use KKT conditions to optimize AVR directly, since their $\mathbf{y}$ is a constant and only $\mathbf{h}$ needs to be optimized. In our work, the MVC model involves a combinatorial optimization problem similarly to many clustering and semi-supervised learning models, especially 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 $V(\mathbf{h})$. In order to solve the MVC model, we proposed two MVC algorithms based on SQP and SDP.

5 Experiments

We included six algorithms in experiments: k-means (KM) (Hartigan & Wong, 1979), soft-label and hard-label maximum volume clustering (SL-MVC and HL-MVC), normalized spectral clustering (SC) (Ng et al., 2001), maximum margin clustering (MMC) (Xu et al., 2004) and generalized MMC (GMMC) (Valizadegan & Jin, 2006). The `CVX` package (Grant & Boyd, 2010) was used to solve (5) in SL-MVC and the SDPs involved in HL-MVC, MMC and GMMC.

Three benchmarks were used here: `MNIST` and `USPS` handwritten digits and `20Newsgroups` text datasets that have 784, 256 and 26214 features, respectively.

First we conducted experiments on `MNIST` for small sample size n ($n \leq 100$). The results are reported in Figure 2, where the error is measured by $\frac{1}{n} d_{\mathcal{H}}$. In order to make a fair comparison, we simply took the *normalized graph Laplacian* L as Q in SL-MVC, and the matrix L in both SL-MVC and SC was constructed by k-Nearest Neighbor (k-NN) using cosine similarity and k was the best integer from 3 to 8. Other parameters of SL-MVC were $\epsilon = 10^{-6}$, $\gamma = 0.01$ and $b = 10^{-7}$. On the other hand, Gaussian kernels were used in HL-MVC, MMC and GMMC because cosine similarity did not work, and the square of the kernel width σ^2 was the best value in $\{1, 10, 10^2, 10^3\}$. In HL-MVC, we fixed $C = 10^4$ while $\gamma = 0.01$ for 1vs.4 and 1vs.7, and $\gamma = 100$ for 3vs.5 and 5vs.8. The parameter C of MMC was chosen from $\{1, 10^4\}$. We fixed $C_e = 10^4$ in GMMC, and the parameter C_δ was chosen from $\{1, 10^3, 10^6\}$. In all experiments, we repeated KM and SC five times and reported the best results, while we ran the other four algorithms only once. It could be observed from Figure 2 that two MVC algorithms were comparable with other algorithms.

Next we compared the performances of SL-MVC, SC and GMMC on three benchmarks for large sample size

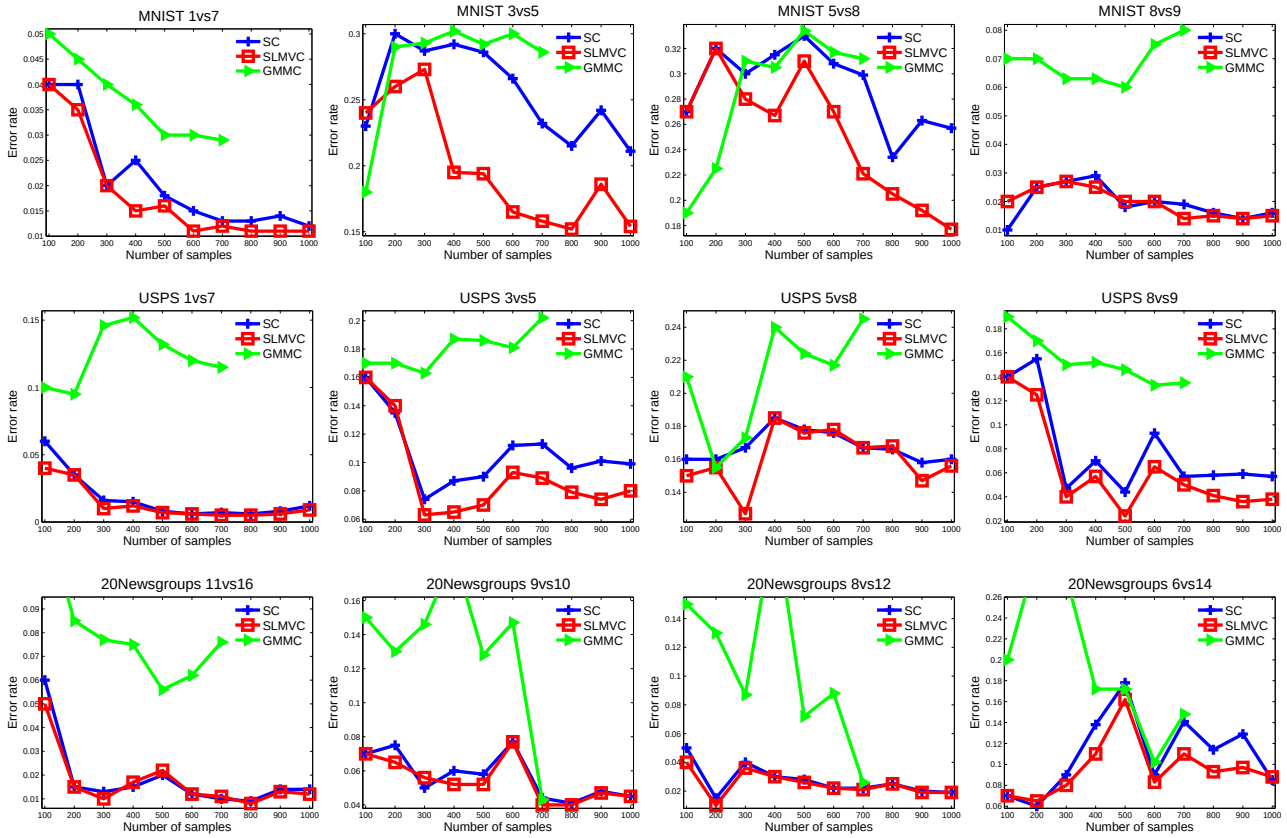

Figure 3: Experimental results on `MNIST`, `USPS` and `20Newsgroups` for large n

n ($100 \leq n \leq 1000$) as reported in Figure 3. We excluded HL-MVC and MMC since they were too slow. The performance of KM was so poor that we excluded it too. We did not test GMMC for some large n if it was unbearably slower than SC and SL-MVC. All parameters here were almost the same as those used in the experiments above, except on `20Newsgroups` the parameter k of k-NN was the best one between 5 and 12 in construction of L for SL-MVC and SC, and we changed $b = 10^{-7}$ to $b = 1$ in SL-MVC to get a looser class balance constraint. The experimental results in Figure 3 illustrate the usefulness of SL-MVC. It often beat GMMC significantly, and almost always had higher or equal accuracy than SC which used the same similarity. In summary, SL-MVC could be a promising alternative to existing spectral and maximum margin clustering algorithms.

6 Conclusions

We proposed a maximum volume clustering model to partition the data into two clusters based on the large volume principle. We elucidated properties of our model and its two approximations (solved by SQP and SDP respectively) from theoretical points of view in detail, and further demonstrated the generality of MVC that it actually includes spectral clustering and maximum margin clustering as special cases. Experiments on benchmarks showed that the proposed MVC approach is very successful in image and text clustering problems.

A key observation in our experiments was that compared with SC and MVC, GMMC worked well when n was small, but performed poorly when n was large. One conjecture is that the volume could capture more structural information of data than the margin for large n and thus better approximate the power of equivalence classes. Another conjecture is that the SDP relaxations of MMC models become looser for large n due to the weak duality, i.e., they are less consistent algorithms when they use SDP to relax the original NP-hard optimizations. In the future work, we will more formally investigate these conjectures.

Acknowledgments

GN is supported by the MEXT scholarship, and MS is supported by the FIRST program. LS is supported by NSF of Jiangsu, China (BK2009233).

References

Boggs, P. T., & Tolle, J. W. (1995). Sequential quadratic programming. *Acta Numerica*, *4*, 1–51.

Cai, D. Text datasets in matlab format. `http://www.cs.uiuc.edu/~dengcai2/Data/TextData.html`.

De Bie, T., & Cristianini, N. (2003). Convex methods for transduction. *NIPS2003*.

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

El-Yaniv, R., & Pechyony, D. (2007). Transductive Rademacher complexity and its applications. *COLT2007*.

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

Grant, M., & Boyd, S. (2010). CVX: Matlab software for disciplined convex programming (web page and software). `http://cvxr.com/cvx`.

Guyon, I., von Luxburg, U., & Williamson, R. C. (2009). Clustering: Science or art? *NIPS 2009 Workshop on Clustering Theory*.

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

Li, Y., Tsang, I. W., Kwok, J. T., & Zhou, Z.-H. (2009). Tighter and convex maximum margin clustering. *AISTATS2009*.

Meila, M., & Shi, J. (2001). A random walks view of spectral segmentation. *AISTATS2001*.

Ng, A., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *NIPS2001*.

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

Roweis, S. Data for MATLAB hackers. `http://cs.nyu.edu/~roweis/data.html`.

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

Valizadegan, H., & Jin, R. (2006). Generalized maximum margin clustering and unsupervised kernel learning. *NIPS2006*.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. Springer Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.

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

Xu, L., Neufeld, J., Larson, B., & Schuurmans, D. (2004). Maximum margin clustering. *NIPS2004*.

Xu, L., & Schuurmans, D. (2005). Unsupervised and semi-supervised multi-class support vector machines. *AAAI2005*.

Zhang, K., Tsang, I. W., & Kwok, J. T. (2007). Maximum margin clustering made practical. *ICML2007*.

Zhao, B., Wang, F., & Zhang, C. (2008a). Efficient maximum margin clustering via cutting plane algorithm. *SDM2008*.

Zhao, B., Wang, F., & Zhang, C. (2008b). Efficient multiclass maximum margin clustering. *ICML2008*.

7 APPENDIX—SUPPLEMENTARY MATERIAL

7.1 Proof of Theorem 3

Proof. Without loss of generality we assume that $\{1,2\} \in \mathcal{I}$ and $\mathbf{h}^* = (\alpha, -\beta, h_3, \ldots, h_n)^\top$ where $\alpha\beta > 0$.

There must be a bijection ϕ which satisfies $\phi(1) = 2$, $\phi(2) = 1$ and the three requirements of ϕ in *Definition 2*. Consider $\mathbf{h}^\star = (-\beta, \alpha, h_{\phi(3)}, \ldots, h_{\phi(n)})^\top$. Obviously $\|\mathbf{h}^\star\|_1 = \|\mathbf{h}^*\|_1$ and $\|\mathbf{h}^\star\|_2 = \|\mathbf{h}^*\|_2$. Moreover, $\forall i$,

$$\begin{aligned}
&[g(\mathbf{h}^\star)]_i \\
&= \gamma \sum\nolimits_{l=1}^{n} Q_{i,l} h_{\phi(l)} - \eta h_{\phi(i)} - \operatorname{sgn}(h_{\phi(i)}) \\
&= \gamma \sum\nolimits_{k=1}^{n} Q_{i,\phi^{-1}(k)} h_k - \eta h_{\phi(i)} - \operatorname{sgn}(h_{\phi(i)}) \\
&= \gamma \sum\nolimits_{k=1}^{n} Q_{i,\phi(k)} h_k - \eta h_{\phi(i)} - \operatorname{sgn}(h_{\phi(i)}) \\
&= \gamma \sum\nolimits_{k=1}^{n} Q_{\phi(i),k} h_k - \eta h_{\phi(i)} - \operatorname{sgn}(h_{\phi(i)}) \\
&= [g(\mathbf{h}^*)]_{\phi(i)}.
\end{aligned}$$

Hence, $g(\mathbf{h}^\star) = \mathbf{0}$ due to the KKT condition $g(\mathbf{h}^*) = \mathbf{0}$, which means that $\mathbf{h}^\star$ is also a minimum. Similarly we can derive $\mathbf{h}^{\star\top} Q \mathbf{h}^\star = \mathbf{h}^{*\top} Q \mathbf{h}^*$ and thereby we arrive at $G(\mathbf{h}^\star) = G(\mathbf{h}^*)$.

Notice that $d_{\mathcal{H}}(\mathbf{h}^*, \mathbf{h}^\star) \geq 1$, with the only exception $d_{\mathcal{H}}(\mathbf{h}^*, \mathbf{h}^\star) = 0$ when $\operatorname{sgn}(\mathbf{h}^\star) = -\operatorname{sgn}(\mathbf{h}^*)$, i.e., $\forall i, \phi(i) \neq i$ and $\mathcal{I}' = \{\{i, \phi(i)\} | i = 1, \ldots, n\}$. This completes the proof. □

7.2 Proof of Theorem 4

Proof. When $n = 2$ it is trivial that X_n is anisotropic.

Suppose that $n > 2$ and $\mathcal{E}(\mathcal{H}_Q)$ has two principal axes $\mathbf{v}_j$ and $\mathbf{v}_k$ with the same length $1/\sqrt{\lambda_j} = 1/\sqrt{\lambda_k}$. Then there is at least one principal axis $\mathbf{v}_l, l \neq j, k$ about which $\mathcal{E}(\mathcal{H}_Q)$ is rotational along the circle spanned by $\mathbf{v}_j$ and $\mathbf{v}_k$.

From $\forall i, Q_{i,i} = \kappa$ we know that $\mathcal{E}(\mathcal{H}_Q)$ intersects the i-th coordinate axis at $\pm\mathbf{e}_i/\sqrt{\kappa}$ with length $1/\sqrt{\kappa}$, where $\mathbf{e}_i$ is the i-th unit vector of $\mathbb{R}^n$. Now $\mathcal{E}(\mathcal{H}_Q)$ has n principal axes with at most $n-1$ different length but another system of n orthogonal axes with length $1/\sqrt{\kappa}$, so $\mathbf{v}_l$ must be in the form of

$$\mathbf{v}_l = \frac{\bar{\mathbf{v}}_l}{\|\bar{\mathbf{v}}_l\|_2}, \ \bar{\mathbf{v}}_l = \sum_{i=1}^{n} \delta_i \mathbf{e}_i \neq \mathbf{0}, \ \delta_i \in \{-1, 0, 1\}.$$

In other words, $\mathbf{v}_l$ lies on the central direction of certain quadrant of a subspace of $\mathbb{R}^n$ determined by $\mathbf{v}_j$ and $\mathbf{v}_k$. But this is impossible since X_n is not axisymmetric w.r.t. Q.

Hence all principal axes of $\mathcal{E}(\mathcal{H}_Q)$ have different length, which is exactly what we were to prove. □

7.3 Proof of Theorem 5

Proof. The KKT condition $g(\mathbf{h}) = \mathbf{0}$ implies

$$\mathbf{h} = \hat{Q}\mathbf{y}, \tag{12}$$

where $\mathbf{y} = \operatorname{sgn}(\mathbf{h}), \hat{Q} = (\gamma Q - \eta I)^{-1}$, and the constant $\eta < \gamma\lambda_1$, λ_1 is the smallest eigenvalue of Q. Substitute (12) into $\|\mathbf{h}\|_2 = 1$, and note that $\hat{Q}^\top = \hat{Q}$,

$$(\hat{Q}\mathbf{y})^\top (\hat{Q}\mathbf{y}) = 1 \Longrightarrow \mathbf{y}^\top \hat{Q}^2 \mathbf{y} = 1.$$

All eigenvalues of Q are different and positive, so are all eigenvalues of $\hat{Q}$. Consequently, $\hat{Q}^2$ has a unique spectral decomposition. Let $\hat{Q}^2 = \sum_{i=1}^{n} \mu_i \mathbf{u}_i \mathbf{u}_i^\top$, then $\mathbf{y}^\top \hat{Q}^2 \mathbf{y} = \sum_{i=1}^{n} \mu_i \|\mathbf{u}_i^\top \mathbf{y}\|_2^2$.

We assert that $\forall \mathbf{y}_1, \mathbf{y}_2 \in \{-1, +1\}^n$, the only possibility of $\mathbf{y}_1^\top \hat{Q}^2 \mathbf{y}_1 = \mathbf{y}_2^\top \hat{Q}^2 \mathbf{y}_2$ is either $\mathbf{y}_1 = \mathbf{y}_2$ or $\mathbf{y}_1 = -\mathbf{y}_2$. Otherwise, there exist two nonempty disjoint indices $\mathcal{J}$ and $\mathcal{K}$, such that $\forall j \in \mathcal{J}, k \in \mathcal{K}, [\mathbf{y}_1]_j = -[\mathbf{y}_1]_k = -[\mathbf{y}_2]_j = [\mathbf{y}_2]_k$. Moreover, $\forall i$, $\sum_{j\in\mathcal{J}} [\mathbf{u}_i]_j [\mathbf{y}_1]_j + \sum_{k\in\mathcal{K}} [\mathbf{u}_i]_k [\mathbf{y}_1]_k = \sum_{j\in\mathcal{J}} [\mathbf{u}_i]_j [\mathbf{y}_2]_j + \sum_{k\in\mathcal{K}} [\mathbf{u}_i]_k [\mathbf{y}_2]_k$ since the spectral decomposition of $\hat{Q}^2$ is unique. Hence, $\sum_{j\in\mathcal{J}} [\mathbf{u}_i]_j = \sum_{k\in\mathcal{K}} [\mathbf{u}_i]_k$ for all $i = 1, \ldots, n$. This means that the row rank of the matrix $U = (\mathbf{u}_1 | \ldots | \mathbf{u}_n)$ is $n-1$, which contradicts the linear independence of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.

Therefore, all minima of (4) are equivalent w.r.t. $d_{\mathcal{H}}$. □

7.4 Proof of Lemma 7

Proof. For any $\mathbf{h} \in \widetilde{\mathcal{H}}_Q$, $\exists \boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\mathbf{h} = U\boldsymbol{\alpha}$, where U consists of n orthonormal eigenvectors of Q, and $\|\boldsymbol{\alpha}\|_2 = 1$ since $\|\mathbf{h}\|_2 = 1$ and $U^\top U = I$. This $\mathbf{h} = U\boldsymbol{\alpha}$ is a UL decomposition (El-Yaniv & Pechyony, 2007) since U has only information about unlabeled samples. Each column of U has unit length, and thus $\|U\|_{\text{Fro}}^2 = n$. The first part of the bound comes from *inequality (20)* in El-Yaniv and Pechyony (2007).

Another UL decomposition is shown in (12). The equation (12) holds for η^* since it holds for any constant η smaller than γ times the smallest eigenvalue of Q. It is also a kernel UL decomposition since the matrix $\hat{Q}$ is symmetric positive definite. Then, the other part of the bound is derived from *inequality (20)* and *inequality (23)* in El-Yaniv and Pechyony (2007) with $\mu_1 = \sqrt{n}$ and $\mu_2 = \sqrt{\mu}$, respectively. □