Maximin Separation Probability Clustering [*]

Gao Huang[†‡], Jianwen Zhang[§], Shiji Song[†‡], Zheng Chen[§]
[†] Tsinghua National Laboratory for Information Science and Technology (TNList)
[‡] Department of Automation, Tsinghua University, Beijing 100084, China
[§] Microsoft Research, Beijing
huang-g09@mails.tsinghua.edu.cn, jiazhan@microsoft.com
shijis@tsinghua.edu.cn, zhengc@microsoft.com

Abstract

This paper proposes a new approach for discriminative clustering. The intuition is, for a good clustering, one should be able to learn a classifier from the clustering labels with high generalization accuracy. Thus we define a novel metric to evaluate the quality of a clustering labeling, named *Minimum Separation Probability* (MSP), which is a lower bound of the generalization accuracy of a classifier learnt from the clustering labeling. We take MSP as the objective to maximize and propose our approach *Maximin Separation Probability Clustering* (MSPC), which has several attractive properties, such as invariance to anisotropic feature scaling and intuitive probabilistic explanation for clustering quality. We present three efficient optimization strategies for MSPC, and analyze their interesting connections to existing clustering approaches, such as maximum margin clustering (MMC) and discriminative k-means. Empirical results on real world data sets verify that MSP is a robust and effective clustering quality measure. It is also shown that the proposed algorithms compare favorably to state-of-the-art clustering algorithms in both accuracy and efficiency.

Introduction

Clustering is an unsupervised learning paradigm that aims to discover interesting groups in given data sets. It arises in a wide range of contexts, and various clustering algorithms have been proposed in the literature. In recent years, there is a new branch of research efforts on clustering, which connects the objective of clustering with that of classification. Some of the representative approaches include Maximum Margin Clustering (MMC) (Xu et al. 2004; Valizadegan and Jin 2007; Zhang, Tsang, and Kwok 2007; Li et al. 2009; Zhou et al.), Maximum Volume Clustering (MVC) (Niu et al. 2013) and linear discriminative analysis (LDA) based clustering (De la Torre and Kanade 2006; Ding and Li 2007). These methods focus on maximizing discrimination between clusters, thus are usually referred to as *discriminative clustering* (DC). Empirically these methods have shown encouraging results in terms of clustering accuracy. However, there are still two important issues which are not well addressed.

The first issue is the problem of anisotropic feature scaling (AFS), which arises frequently in practice when features correspond to variables with different units. In contrast to supervised classification, clustering algorithms are more susceptible to AFS due to the absence of labels. Supposing we amplify the values of a certain feature by a large factor (e.g., change the unit of time from minute to second), then the maximal margin in MMC, or the maximal volume in MVC, will be dominated by this feature, and the clustering results are very likely to be affected. A common solution to this problem is preprocessing the data so that the values of each feature lie in a proper range. Although effective in practice, data preprocessing is generally performed independently of the clustering process, making it difficult to be optimized.

The second issue is the lack of a consistent measure for evaluating clustering quality. In supervised classification, we can leverage labeled training data to evaluate how confident we are in the learned model. However, it is a challenge to define a quality measure for clustering without true labels. Though the clustering measures of existing clustering algorithms, such as energy in k-means, maximum margin in MMC, entropy or mutual information in information-based clustering, can be used as measures for clustering quality, they are basically problem dependent. For example, some of these measures are not invariant to problem size or are sensitive to feature scaling.

This paper tries to address the aforementioned issues by resorting to the key intuition of discriminative clustering: if a clustering is good, then we should be able to learn a classifier from the clustering labels with high generalization accuracy. However, it is usually difficult to evaluate the generalization performance without true labels. In this paper, we define a novel clustering metric named *Minimum Separation Probability* (MSP) based on the Minimax Probability Machine (MPM) (Lanckriet et al. 2003), and propose a clustering approach by maximizing MSP, which is called *Maximin*[1] *Separation Probability Clustering* (MSPC). It can be

[*] This work was done during Gao Huang's internship at Microsoft Research.

[1] The term *minimax* is used in (Lanckriet et al. 2003) since it considers minimizing the maximum probability of misclassification.

shown that MSPC has close relations to MMC and several other DC approaches. However, MSPC is invariant to AFS, and its objective yields an intuitive explanation for clustering quality. Empirical results demonstrate that MSP has strong positive correlation with clustering accuracy. Additionally, we develop efficient learning algorithms for MSPC, leading to an increase in speed by several orders-of-magnitude over MMC and MVC algorithms. Meanwhile, MSPC achieves impressive clustering accuracy on real data sets.

Our contributions. 1) We introduce a novel clustering approach MSPC, the objective of which provides an effective and robust measure for clustering quality. 2) We propose three algorithms to optimize the proposed model, which are advantageous both in efficiency and clustering accuracy. 3) We study connections between MSPC and several existing algorithms. This also builds a bridge between two important types of DC approaches, namely, MMC and LDA based clustering.

Related Work

Discriminative clustering. Maximum Margin Clustering (MMC) (Xu et al. 2004) extended the large margin principle in SVM to clustering. There are several improvements to MMC, which can handle multi-class clustering or with higher efficiency (Zhang, Tsang, and Kwok 2007; Zhao, Wang, and Zhang 2008; Li et al. 2009). Similarly, the Maximum Volume Clustering (MVC) (Niu et al. 2013) extended the large volume principle to clustering. There are several other related discriminative clustering approaches, which simultaneously perform supervised dimension reduction and clustering (De la Torre and Kanade 2006; Ding and Li 2007; Ye, Zhao, and Wu 2007). We will further discuss the connections between these algorithms and MSPC in the rest sections.

Minimax Probability Machine (MPM) was proposed by Lanckriet et al. (2003) for supervised classification. It learns a classifier by maximizing a lower bound of the generalization accuracy. Recently, Huang et al. (2014) revisited MPM and extended it for transductive learning and semi-supervised learning based on an efficient label-switching strategy. Our paper can be viewed as extending MPM to the unsupervised case where we need to simultaneously learn a clustering labeling and a classifier.

Clustering Quality Measure (CQM) has been studied in the applied statistics literature (Milligan 1981), where it is termed as *cluster validity*. Recently, CQM has been axiomatized for pairwise distance-based clustering by Ackerman and Ben-David (2008). However, they do not connect CQM with the commonly used clustering accuracy. The unsupervised classification accuracy in (Balasubramanian, Donmez, and Lebanon 2011) can also be viewed as a CQM, but it assumes label proportion is known.

Minimax Probability Machine

Considering binary classification, there is a labeled training data set $\{\mathcal{X}, \mathcal{Y}\} = \{\mathbf{x}_i, y_i\}_{i=1}^N (\mathbf{x}_i \in \mathcal{R}^d, y_i \in \{1,2\})$, where N is the number of samples, and d is the dimensionality. The Minimax Probability Machine (MPM) (Lanckriet et al. 2003) assumes that the samples of each class $\mathcal{X}_j (j = 1, 2)$ are independently generated by a random distribution, thus the class-conditional distributions can be modeled by two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. MPM aims to maximize the probabilities that $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively lie on two sides of a hyperplane $\mathcal{H}(\mathbf{w}, b)(\mathbf{w} \in \mathcal{R}^d, b \in \mathcal{R})$, i.e., $\Pr(\mathbf{w}^\top \mathbf{X}_1 + b \leq 0)$ and $\Pr(\mathbf{w}^\top \mathbf{X}_2 + b \geq 0)$. As these probabilities are usually difficult to compute, MPM considers maximizing the worst case separation probability over all possible distributions of $\mathbf{X}_j$ whose mean and covariance $\{\mu_j, \mathbf{\Sigma}_j\}$ match the empirical moments $\{\widehat{\mu}_j, \widehat{\mathbf{\Sigma}}_j\}$ of $\mathcal{X}_j$, leading to the following MPM formulation:

$$\begin{aligned} \max_{p, \mathbf{w} \in \mathcal{R}^d, b \in \mathcal{R}} \quad & p \\ \text{s.t.} \quad & \inf_{\mathbf{X}_1 \sim \{\widehat{\mu}_1, \widehat{\mathbf{\Sigma}}_1\}} \Pr(\mathbf{w}^\top \mathbf{X}_1 + b \geq 0) \geq p, \\ & \inf_{\mathbf{X}_2 \sim \{\widehat{\mu}_2, \widehat{\mathbf{\Sigma}}_2\}} \Pr(\mathbf{w}^\top \mathbf{X}_2 + b \leq 0) \geq p, \end{aligned} \tag{1}$$

where $\widehat{\mu}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in \mathcal{X}_k} \mathbf{x}_i$, $\widehat{\mathbf{\Sigma}}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in \mathcal{X}_k} (\mathbf{x}_i - \widehat{\mu}_k)(\mathbf{x}_i - \widehat{\mu}_k)^\top$ and $N_k = |\mathcal{X}_k|$, with $k = 1, 2$.

A remarkable advantage of considering the worst case distribution is that the constraints in (1) can be converted into second order cone constraints due to the multivariate Chebyshev inequalities (Marshall, Olkin, and others 1960), without making specific assumptions on the distribution form of $\mathbf{X}_k$. As derived in (Lanckriet et al. 2003), the optimization in (1) can be reformulated as

$$\max_{\mathbf{w} \in \mathcal{R}^d} \kappa := \frac{\mathbf{w}^\top (\widehat{\mu}_1 - \widehat{\mu}_2)}{\sqrt{\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_1 \mathbf{w}} + \sqrt{\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_2 \mathbf{w}}}, \tag{2}$$

where $\kappa = \sqrt{p/(1-p)}$. This optimization problem can be solved efficiently by standard second order cone programming (SOCP), or by the iterative least squares approach introduced in (Lanckriet et al. 2003), which has a worst case complexity of $O(d^3)$ [2].

Maximin Separation Probability Clustering

Inspired by MPM, we define a novel metric to measure the quality of a candidate clustering labeling.

Definition 1. *The* minimum separation probability (MSP) *of a clustering labeling* $\mathbf{y}$ *on* $\mathbf{X}$ *is the optimal* p *solved from problem (1) on the pseudo-labeled training set* $\{\mathbf{X}, \mathbf{y}\}$.

The intuition under this metric is, from a good clustering labeling, we should be able to learn a classifier with high generalization accuracy, which is bounded from below by MSP. Note here we are actually assuming that the clusters induced by an arbitrary labeling $\mathbf{y}$ are consisted of i.i.d samples which are generated from independent distributions. This assumption is stronger than that in MPM, but is common in discriminative clustering.

The proposed clustering approach is to find a best clustering labeling by directly maximizing MSP:

$$\max_{\mathbf{w} \in \mathcal{R}^d, \mathbf{y} \in \{1,2\}^N} \kappa := \frac{|\mathbf{w}^\top (\widehat{\mu}_1(\mathbf{y}) - \widehat{\mu}_2(\mathbf{y}))|}{\sum_{k=1,2} \sqrt{\mathbf{w}^\top (\widehat{\mathbf{\Sigma}}_k(\mathbf{y}) + \lambda \mathbf{\Lambda}) \mathbf{w}}} \tag{3}$$

[2] This complexity does not take into account the cost of computing the empirical moments, which has a complexity of $O(Nd^2)$

where $\lambda\mathbf{\Lambda}$ is a regularization term on the covariance matrices. Though the two covariance matrices can be regularized separately with different regularizers, we use a unified one for convenience. In our algorithms, we let $\mathbf{\Lambda} = \mathrm{diag}(\widehat{\mathbf{\Sigma}})$, where $\widehat{\mathbf{\Sigma}}$ is the covariance of the whole training set $\mathcal{X}$. Note that the empirical mean and covariance of both clusters are functions of the label vector $\mathbf{y}$.

We call the formulation (3) *Maximin Separation Probability Clustering* (MSPC). In the subsequent part, we analyze its properties, and introduce optimization methods to solve it.

Invariant to feature scaling The following theorem shows that MSP is invariant to invertible linear transformation on training data, and is thus immune to anisotropic feature scaling.

Theorem 1. *Let $\mathbf{P} \in \mathcal{R}^{d\times d}$ be an invertible matrix. Then the two data sets $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathcal{X}^{\mathbf{P}} = \{\mathbf{P}^\top \mathbf{x}_i\}_{i=1}^N$ have the same optimal MSP.*

Due to limited spaces, all proofs appear in the supplementary material.

Since MSP is not pairwise distance-based CQM, we cannot analyze it under the axiomatic framework proposed by (Ackerman and Ben-David 2008). However, it can be shown that a generalized version of MSP satisfy all the axioms for a valid CQM. We give the detailed discussions in the supplementary material.

Optimization of MSPC

The mixed integer problem (3) is intractable in practice. In this section, we introduce three efficient algorithms to solve it approximately.

Approach 1: Iterative MPM

Inspired by the iterative support vector regression method (IterSVR) (Zhang, Tsang, and Kwok 2007) for MMC, we propose to solve MSPC in a similar fashion. Firstly, k-means clustering is adopted to find an initial labeling. Then we iterate between solving $\mathbf{w}$ and updating $\mathbf{y}$ as the IterSVR algorithm. Note that with fixed labeling, solving $\mathbf{w}$ in MSPC is actually training an ordinary MPM model. We denote this algorithm by MSPC$^{\text{MPM}}$, which is summarized in the first column of Table 1.

This simple algorithm works surprisingly well in our experiments. It also converges fast, with typically 10 iterations before terminating. Moreover, using the updating formula (2) in Appendix B, at each iteration we only need to compute the covariance of a small set instances, i.e., those who are switched from one cluster to the other at this iteration.

Approach 2: Maximizing a lower bound

In this subsection, we derive a lower bound of MSP, enabling the design of an efficient MSPC algorithm. Moreover, this bound allows us to build connections between MSPC and some existing algorithms, as given in the next section.

Let $r_1 \triangleq N_1/N$ and $r_2 \triangleq N_2/N$ denote the ratios of samples assigned to subsets $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. We then have the following theorem which gives a lower bound for the objective in (3).

Theorem 2. *Let $r_1 \vee r_2 \triangleq \max\{r_1, r_2\}$ and $r_1 \wedge r_2 \triangleq \min\{r_1, r_2\}$. Then we have*

$$\kappa^2 \geq \frac{(\mathbf{w}^\top(\widehat{\mu}_1 - \widehat{\mu}_2))^2}{2(r_1 \wedge r_2)^{-1}\mathbf{w}^\top(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{\Lambda})\mathbf{w} - 2(r_1 \vee r_2)(\mathbf{w}^\top(\widehat{\mu}_1 - \widehat{\mu}_2))^2}, \quad (4)$$

and the equality holds when $\mathbf{w}^\top\widehat{\mathbf{\Sigma}}_1\mathbf{w} = \mathbf{w}^\top\widehat{\mathbf{\Sigma}}_2\mathbf{w}$ and $r_1 = r_2$.

In this expression, r_k and $\widehat{\mu}_k$ $(k = 1, 2)$ are functions of the label vector $\mathbf{y}$. When $\mathbf{y}$ is given, maximizing the lower bound in (4) reduces to

$$\max_{\mathbf{w}\in\mathcal{R}^d} \quad \frac{\mathbf{w}^\top(\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^\top\mathbf{w}}{\mathbf{w}^\top(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{\Lambda})\mathbf{w}}. \quad (5)$$

The optimal solution to (5) can be obtained efficiently by finding the largest eigenvector of the generalized eigenvalue problem (GEP): $(\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^\top\mathbf{v} = \gamma(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{\Lambda})\mathbf{v}$.

Thus we can again adopt alternating optimization scheme to optimize the lower bound of κ. Specifically, with fixed $\mathbf{y}$, the optimal $\mathbf{w}$ is obtained by solving a GEP. With fixed $\mathbf{w}$, the optimal label vector can be obtained as follows [3]. We first sort the samples according to the values of their projections on $\mathbf{w}$, i.e., $t_i = \mathbf{w}^\top\mathbf{x}_i$. Then we assign the first N_1 samples to $\mathcal{X}_1$ and the rest to $\mathcal{X}_2$, where N_1 is an integer between 1 and $N - 1$ that maximizes the lower bound in (4). The algorithm is summarized in the second column of Table 1.

Approach 3: A relaxation method

We give a relaxation of the formulation in Approach 2. The relaxed model can be solved without alternating optimization, and is efficient for high dimensional data.

Define a cluster indicator vector $\mathbf{q}$ as $q_i = \sqrt{N_2/(NN_1)}$, if $i \in \mathcal{X}_1$, and $q_i = -\sqrt{N_1/(NN_2)}$, if $i \in \mathcal{X}_2$.

Thus we have $\|\mathbf{q}\| = 1$, and $\mathbf{w}^\top(\widehat{\mu}_1 - \widehat{\mu}_2) = \sqrt{N/(N_1N_2)}\mathbf{w}^\top\mathbf{X}^\top\mathbf{q}$.

Now we relax $\mathbf{q}$ to take real values under the constraint $\|\mathbf{q}\| = 1$, and assume that N_1 and N_2 are constants (consequently, r_1 and r_2 are constants as well). The optimization problem in Approach 2 is approximated by

$$\max_{\mathbf{w}\in\mathcal{R}^d, \|\mathbf{q}\|=1} \quad \frac{\mathbf{w}^\top\mathbf{X}^\top\mathbf{q}\mathbf{q}^\top\mathbf{X}\mathbf{w}}{\mathbf{w}^\top(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{\Lambda})\mathbf{w}}. \quad (6)$$

Without loss of generality, we assume that the data matrix are preprocessed to be centered, i.e., $\mathbf{1}^\top\mathbf{X} = \mathbf{0}$. Then we show that the optimal $\mathbf{q}$ can be obtained by solving an eigenproblem, as stated by the following theorem.

Theorem 3. *An optimal solution of $\mathbf{q}$ to (6) is the largest normalized eignevector of the matrix $\mathbf{I}_N - (\mathbf{I}_N + \widetilde{\mathbf{X}})^{-1}$ (or $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{\Lambda})^{-1}\mathbf{X}^\top$), where $\widetilde{\mathbf{X}} = \mathbf{X}(\lambda\mathbf{\Lambda})^{-1}\mathbf{X}^\top$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix.*

According to Theorem 3, the relaxed cluster indicator vector $\mathbf{q}$ can be optimized conveniently by computing the largest eigenvector of a $N \times N$ matrix. When the optimal $\mathbf{q}^*$ is obtained, we can use the sign of its elements to construct our label vector. Alternatively, we can use the predict vector $\mathbf{X}\mathbf{w}^* = \mathbf{X}\mathbf{X}^\top\mathbf{q}^*$ to obtain our final prediction. In our experiments, the later one is used. We denote the algorithm as MSPC$^{\text{EIG}}$, which is presented in the last column of Table 1.

[3] It can be shown that this process finds the optimal $\mathbf{y}$ with given $\mathbf{w}$, in $O(N \log N + Nd)$ time.

Analysis of the optimization algorithms

Among the three algorithms, MSPC$^{\text{MPM}}$ is the only one that tries to directly optimize the MSP defined in (3), and it usually yields the highest accuracy in our experiments. It has a complexity of $O(l(Nd^2 + d^3))$, where l denotes the number of iterations, Nd^2 comes from the covariance estimation, and d^3 accounts for MPM retaining. Typically, we have $l = 10$, thus the algorithm is very efficient when d is not large.

MSPC$^{\text{GEP}}$ works in a similar fashion as MSPC$^{\text{MPM}}$, but it aims to maximize a lower bound of MSP. Since the lower bound is tight when two clusters are balanced, MSPC$^{\text{GEP}}$ tends to give similar results as MSPC$^{\text{MPM}}$ on balanced data sets. It has a computational complexity of $O(l(Nd^2 + d^\alpha + N \log N))$, where d^α corresponds to solving the GEP in (5) with $2 < \alpha < 3$, and $N \log N$ accounts for the cost of sorting the projections.

MSPC$^{\text{EIG}}$ has a complexity of $O(N^2 d + N^3)$, if $N < d$, or $O(N^\alpha + d^3)$, if $N > d$ ($2 < \alpha < 3$). Thus it is more efficient than MSPC$^{\text{MPM}}$ and MSPC$^{\text{GEP}}$ on high dimensional data where $N \ll d$.

To summarize, MSPC$^{\text{MPM}}$ is the first choice if accuracy is the main concern; MSPC$^{\text{GEP}}$ is especially suited for balanced and medium-dimensional data; MSPC$^{\text{EIG}}$ is the most efficient one on high dimensional data.

Connections with Existing Algorithms

Relationships to MMC

We show that interesting connections exist between MSPC and MMC. Actually, MMC solutions ensures high MSP under certain conditions, and MSPC essentially favors a large *relative margin*.

Theorem 4. *Supposing that the optimal margin learned by MMC is δ, then the MSP (with $\lambda = 0$) is at least $p \geq \frac{1}{2(r_1 \wedge r_2)^{-1}(\sigma_{\max}/\delta)^2 - 2(r_1 \vee r_2) + 1}$, where $\sigma^2_{\max}$ is the largest eigenvalue of the covariance matrix $\widehat{\mathbf{\Sigma}}$, and r_1 and r_2 are the ratios of samples assigned to the two clusters, respectively.*

Theorem 4 states that to ensure high MSP, the *normalized margin* $\delta/\sigma_{\max}$, instead of the absolute margin δ, should be large. For any data set, we could create an arbitrarily large margin by feature scaling, however, the normalized margin cannot be made arbitrary large since $\sigma_{\max}$ is also sensitive to feature scaling. This theorem also explains why data preprocessing is important to MMC: it prevents $\sigma_{\max}$ from growing too large, so that the subsequent margin maximization process could reliably maximize intrinsic separability.

Theorem 5. *Supposing that the two clusters yielded by MSPC are respectively generated from two independent distributions with mean μ_1 and μ_2, and with a MSP ($\lambda = 0$) of p, then there exists a hyperplane $\{\mathbf{w}, b\}$ such that at least $1 - r$ fraction of the points from each distribution satisfy $|\mathbf{w}^\top \mathbf{x}_i + b| \geq \rho\Delta/2$ ($0 \leq \rho \leq 1$), where $\Delta = |\mathbf{w}^\top(\mu_1 - \mu_2)|$, and $r = \frac{p^{-1} - 1}{(p^{-1}-1) + (1-\rho)^2/4}$.*

In MMC, we are usually maximizing a *soft* margin, by pushing *most* of the points away from the margin. From Theorem 5, we see that MSPC also ensures a large fraction (e.g., at least $1 - r$) of the samples are separated by a margin $\rho\Delta$, where Δ is actually the *average margin* between two clusters. If we define the $|\mathbf{w}^\top \mathbf{x}_i + b|/\Delta$ as the *relative margin*, then a high MSP solution ensures a large relative margin with high probability. Unlike the absolute margin in MMC, the relative margin is invariant to feature scaling.

Relationships to LDA-based clustering approaches

Several recent work (De la Torre and Kanade 2006; Ding and Li 2007; Ye, Zhao, and Wu 2007) addressed the advantages of integrating linear discriminant analysis (LDA) and k-means clustering. It can be shown that MSPC is related to the these discriminative clustering approaches.

The objective function (binary clustering case) introduced in (De la Torre and Kanade 2006) is given by

$$\max_{\mathbf{w} \in \mathcal{R}^d, \mathbf{y} \in \{1,2\}^N} (\mathbf{w}^\top \mathbf{S} \mathbf{w})^{-1} \mathbf{w}^\top \mathbf{S}_b(\mathbf{y}) \mathbf{w}, \tag{7}$$

where $\mathbf{S}$ is the *total scatter matrix*, and $\mathbf{S}_b$ is the *between cluster scatter matrix*, which is a function of the label vector $\mathbf{y}$.

It can be observed that for binary clustering, $\mathbf{S}$ and $\mathbf{S}_b$ are respectively proportional to $\widehat{\mathbf{\Sigma}}$ and $(\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^\top$ in our algorithm. Therefore, the two expressions in (4) and (7) are actually equivalent, if we omit the ratios r_j and the regularization term in the former. Thus the algorithm in (De la Torre and Kanade 2006) essentially maximizes an approximated lower bound of MSP (the objective of MSPC$^{\text{EIG}}$).

The objective function in (Ding and Li 2007) is identical to that in (De la Torre and Kanade 2006), except the *total scatter matrix* $\mathbf{S}$ is replaced by the *within cluster scatter matrix* $\mathbf{S}_w$. Since we have the relation $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$, the two objective are essentially equivalent. Thus the algorithm in (Ding and Li 2007) also maximize an approximated lower bound of MSP, but with a different optimization scheme.

The discriminative k-means algorithm proposed in (Ye, Zhao, and Wu 2007) has a similar objective as that in (Ding and Li 2007). It performs kernel k-means clustering with the kernel matrix $\mathbf{I}_N - (\mathbf{I}_N + (\lambda)^{-1}\mathbf{X}\mathbf{X}^\top)^{-1}$. Obviously, this kernel matrix can be obtained from Theorem 3 by setting the regularization matrix $\mathbf{\Lambda}$ to $\mathbf{I}_N$. Though the difference seems trivial, it is critical since the regularizer $\mathbf{\Lambda}$ preserves the property of invariant to AFS in our algorithm.

Empirical Study

Experiment setup All algorithms are implemented in MATLAB$^{\text{TM}}$, and are executed on an Intel i7 Quad Core CPU 3.39GHz machine with 16GB RAM. All of data sets are obtained from the UCI repository, except the handwritten digits data sets *MNIST*[4] and *USPS*. For those data sets originally contain more than two classes, we select their first two classes to create binary clustering tasks, if not explicitly

[4] Available at http://yann.lecun.com/exdb/mnist/

MSPC$^{\text{MPM}}$	MSPC$^{\text{GEP}}$	MSPC$^{\text{EIG}}$
1. Initialize $\mathbf{y}$ by k-means; 2. Fix $\mathbf{y}$ and optimize $\{\mathbf{w}, b\}$ with MPM; 3. Update the label vector $\mathbf{y}$ according to sign$(\mathbf{w}^\top \mathbf{x}_i + b)$; 4. Repeat Step 2 and 3 until converge.	1. Initialize $\mathbf{y}$ by k-means; 2. Fix $\mathbf{y}$ and optimize $\mathbf{w}$ by solving (5); 3. Fix $\mathbf{w}$, and find the optimal $\mathbf{y}$ that maximizes the lower bound of κ; 4. Repeat Step 2 and 3 until converge.	1. Calculate $\mathbf{I}_N - (\mathbf{I}_N + \widetilde{\mathbf{X}})^{-1}$ or $\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{\Lambda})^{-1}\mathbf{X}^\top$; 2. Compute the largest eigenvector $\mathbf{q}^*$; 3. Obtaining the optimal label vector via the sign of $\mathbf{X}\mathbf{X}^\top \mathbf{q}^*$.

Table 1: Summary of the proposed three MSPC algorithms.

ID	Data	N	d	KM	Iter-SVR	LG-MMC	MVC	LDA-KM	Dis-KM	MSPC-MPM	MSPC-GEP	MSPC-EIG
1	*ionosphere*	351	34	28.77	22.51	24.70	**15.67**	28.77	17.95	28.77	29.63	29.63
2	*breast*	683	10	3.81	3.22	3.51	2.93	3.81	3.81	2.93	**2.63**	2.78
3	*australian*	690	14	14.49	**14.06**	**14.06**	17.25	14.49	14.49	14.49	**14.06**	16.67
4	*diabetes*	768	8	33.07	30.86	32.68	29.63	32.94	**28.26**	32.55	31.51	31.77
5	*letter*	1555	16	6.30	5.53	**0.00**	5.66	5.53	5.59	5.59	5.53	8.75
6	*satellite*	2236	36	4.25	6.17	0.76	0.85	4.07	4.20	**0.63**	3.80	1.70
7	*spam*	4601	57	20.04	20.98	18.30	N/A	20.07	19.17	**13.76**	17.19	21.75
8	*mnist3vs8*	13966	784	20.04	20.06	18.12	N/A	18.78	18.78	**17.85**	18.54	23.19
9	*mnist1vs7*	15170	784	4.23	4.23	2.30	N/A	2.52	1.54	**0.93**	1.10	3.70
10	*mnist8vs9*	13783	784	7.34	40.16	7.19	N/A	5.12	4.43	**4.25**	4.67	18.50

Table 2: Clustering error (%) on ten clustering tasks.

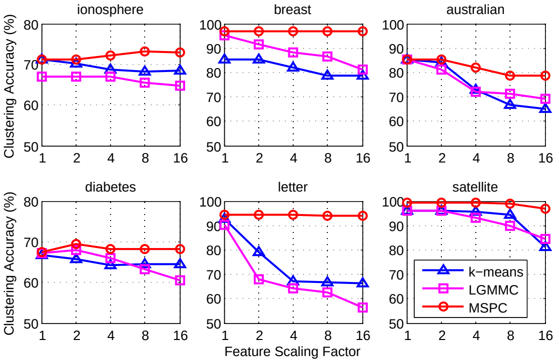

Figure 1: Clustering accuracy (%) with respect to anisotropic feature scaling.

stated. For comparisons to be fair, we normalize each feature of the UCI data sets to the range [-1,1][5].

Baselines and parameter setting The baseline clustering algorithms include two relatively efficient MMC algorithms, IterSVR (Zhang, Tsang, and Kwok 2007) and LGMMC (Li et al. 2009); MVC (Niu et al. 2013), and two related DC algorithms, LDA-KM (Ding and Li 2007) and Dis-KM (Ye, Zhao, and Wu 2007). For IterSVR, LGMMC and MVC, both the regularization parameter C and the Gaussian kernel parameter σ are tuned (linear kernel LGMMC is used for the three *mnist* data sets), following the settings in (Li et al. 2009) and (Niu et al. 2013), i.e., the hyperparameters correspond to highest clustering accuracy are directly selected. For LDA-KM, Dis-KM and the proposed three MSPC algorithms, the regularization coefficient, which is the only hyperparameter, is optimally chosen from the candidate set $[10^{-4}, 10^{-3}, \ldots, 10^{4}]$ based on the clustering accuracy.

Clustering Accuracy and efficiency

We report the clustering error on seven UCI data sets and three binary clustering tasks in Table 2, where the best results are shown in bold. An "N/A" indicates an algorithm that fails to finish in a reasonable time (24 hours). It can be observed that MSPC$^{\text{MPM}}$, which directly optimizes the MSP, achieves very encouraging results. It yields the lowest clustering error on the five relatively larger data sets. This validates MSP as a proper clustering measure. In accordance with our analysis in previous sections, the results obtained by MSPC$^{\text{GEP}}$ are similar to that of MSPC$^{\text{MPM}}$ on the balanced data sets, e.g., *letter* and the three binary *mnist* tasks. However, the clustering error given by MSPC$^{\text{EIG}}$ is relatively higher than that obtained by the other two MSPC algorithms.

In Table 3, we report the running time of the DC algorithms on six relatively larger data sets. From the results, one can observe that the proposed three algorithms are several orders of magnitude faster than MMC and MVC. MSPC$^{\text{MPM}}$ is the most efficient algorithm on low dimensional data sets, while it is surpassed by MSPC$^{\text{GEP}}$ on *mnist* tasks which have a relatively larger dimensionality. We attribute the high efficiency of MSPC$^{\text{MPM}}$ and MSPC$^{\text{GEP}}$ to two main reasons: 1) solving for $\mathbf{w}$ is independent of the number of training data when the means and covariances are given, thus it is efficient when d is not large, and 2) the covariances can be updated efficiently by using the formula (2) in Appendix B (the means can be updated similarly). Although some baselines, such as Iter-SVR, also have a linear complexity in N, there is a large constant hidden in it. Note that for extremely high dimensional data with $d \gg N$, we can use MSPC$^{\text{EIG}}$, or use MSPC$^{\text{MPM}}$ with kernel MPM introduced in (Lanckriet et al. 2003).

Sensitivity to anisotropic feature scaling

In order to verify that MSPC is insensitive to AFS, we run MSPC$^{\text{MPM}}$ on six deliberately un-uniformly scaled data sets. At each time, we randomly select one of the features from

[5] Actually, the proposed algorithms performed well on both normalized and unnormalized data, while some baselines degraded significantly without feature normalizing.

ID	Data	Iter-SVR	LG-MMC	MVC	LDA-KM	Dis-KM	MSPC$^{\text{MPM}}$	MSPC$^{\text{GEP}}$	MSPC$^{\text{EIG}}$
5	*letter*	50.2	45.7	178	0.07	0.38	**0.02**	0.23	0.45
6	*satellite*	9.16	140	603	0.10	1.05	**0.02**	0.27	0.95
7	*spam*	42.6	1769	N/A	0.52	5.01	**0.09**	0.10	5.63
8	*mnist3vs8*	6372	1824	N/A	5.53	96.8	2.82	**2.72**	9.25
9	*mnist1vs7*	3602	2602	N/A	3.01	184	1.65	**1.54**	16.6
10	*mnist8vs9*	3504	2990	N/A	4.01	109	2.40	**1.83**	9.88

Table 3: Wall clock time (in seconds) of the clustering algorithms.

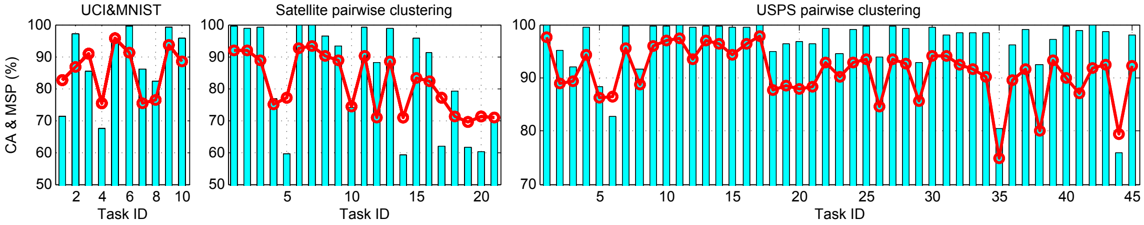

Figure 2: MSP (the circles) and CA (the bars) on three groups of data sets. From left to right: the 10 tasks presented in Table 2; the 21 binary clustering tasks created from *Satellite*; and the 45 binary clustering tasks created from *USPS*(the tasks are ordered by *“0vs1”,“0vs2”,...,“0vs9”,“1vs2”,...,“8vs9”*).

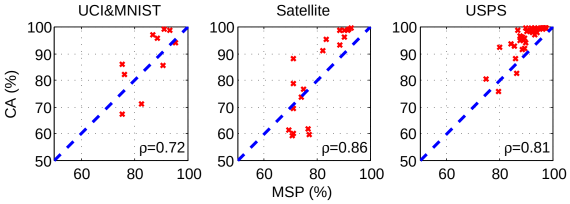

Figure 3: Correlation between CA and MSP.

a data set, and scale its values by a factor of a, where a is chosen from the set $[2, 4, 8, 16]$. Then we run k-means, LGMMC with linear kernel, and MSPC$^{\text{MPM}}$ on it. This experiment is repeated 10 times (we ensure 10 different features are scaled, and for *diabetes* all the eight features are selected and scaled once) on each data set, and the average results obtained by the three algorithms are shown in Fig.1. It is obvious that MSPC$^{\text{MPM}}$ is robust to such feature scaling, while the other two algorithms degrade significantly with a large scaling factor. Note that MSPC$^{\text{MPM}}$ is a local search algorithm and it is initialized with k-means, thus it is not fully immune to AFS. If we can afford to find the global optimum, then MSPC solutions will totally be unaffected by AFS, as guaranteed by Theorem 1.

MSP and clustering accuracy

Estimating clustering accuracy (CA) without ground truth labels is a very challenging task. Since MSP is a problem-independent clustering quality measure, and it has the same value range as CA ([0,1]), it is interesting to see how they correlate with each other empirically. Fig.2 and Fig.3 show the CA and corresponding MSP obtained by MSPC$^{\text{MPM}}$ on the ten tasks presented in Table 2, the 21 binary clustering tasks created from *Satellite*, and the 45 binary clustering tasks created from *USPS*. One can observe that these two variables do correlate with each other to some extent. For example, the correlation coefficient on the 10 UCI and *MNIST* data sets is 0.72, indicating a moderate positive linear relationship between them. If we only consider the later five larger data sets, then the correlation coefficient reaches 0.97, which means an almost perfect linear correlation. Interestingly, the MSP indicates the separability of the three *MNIST* tasks can be ordered as *“1vs7”*>*“8vs9”*>*“3vs8”*, which exactly matches the order given by CA. The results on the binary clustering tasks of *Satellite* and *USPS* also show a high correlation between MSP and CA, with coefficients of 0.86 and 0.81 respectively (shown in Fig.3).

Conclusion

In this paper, we proposed a new clustering metric, minimum separation probability (MSP), which is a lower bound of the generalization accuracy of a classifier learnt from the clustering labeling. Then we proposed three clustering approaches by approximately maximizing MSP. We provided a detailed analysis of the relations between MSPC and several existing discriminative clustering approaches. Empirical results demonstrated that MSP is insensitive to anisotropic feature scaling, and it showed a positive correlation with clustering accuracy on real data sets. The proposed algorithms were impressively faster than MMC and MVC algorithms, and compared favorably to state-of-the-art clustering algorithms in terms of accuracy. Future work could focus on further studying the relations between MSP and clustering accuracy, and extending MSPC for multi-class clustering.

Acknowledgements. Gao Huang and Shiji Song are partially supported by the National Natural Science Foundation of China under Grant Nos. 41427806 and 61273233, the Research Fund for the Doctoral Program of Higher Education under Grant Nos. 20120002110035 and 20130002130010, the Project of China Ocean Association under Grant DY125-25-02, and Tsinghua University Initiative Scientific Research Program under Grant No. 20131089300.

References

Ackerman, M., and Ben-David, S. 2008. Measures of clustering quality: A working set of axioms for clustering. In *NIPS*.

Balasubramanian, K.; Donmez, P.; and Lebanon, G. 2011. Unsupervised supervised learning II: Margin-based classification without labels. *The Journal of Machine Learning Research* 12:3119–3145.

De la Torre, F., and Kanade, T. 2006. Discriminative cluster analysis. In *ICML*.

Ding, C., and Li, T. 2007. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *ICML*.

Huang, G.; Song, S.; Xu, Z.; and Weinberger, K. 2014. Transductive minimax probability machine. In *ECML/PKDD*, 579–594.

Lanckriet, G. R.; Ghaoui, L. E.; Bhattacharyya, C.; and Jordan, M. I. 2003. A robust minimax approach to classification. *The Journal of Machine Learning Research* 3:555–582.

Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2009. Tighter and convex maximum margin clustering. In *AISTATS*.

Marshall, A. W.; Olkin, I.; et al. 1960. Multivariate chebyshev inequalities. *The Annals of Mathematical Statistics* 31(4):1001–1014.

Milligan, G. W. 1981. A review of monte carlo tests of cluster analysis. *Multivariate Behavioral Research* 16(3):379–407.

Niu, G.; Dai, B.; Shang, L.; and Sugiyama, M. 2013. Maximum volume clustering: A new discriminative clustering approach. *The Journal of Machine Learning Research* 14(1):2641–2687.

Valizadegan, H., and Jin, R. 2007. Generalized maximum margin clustering and unsupervised kernel learning. In *NIPS*.

Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D. 2004. Maximum margin clustering. In *NIPS*.

Ye, J.; Zhao, Z.; and Wu, M. 2007. Discriminative k-means for clustering. In *NIPS*.

Zhang, K.; Tsang, I. W.; and Kwok, J. T. 2007. Maximum margin clustering made practical. In *ICML*.

Zhao, B.; Wang, F.; and Zhang, C. 2008. Efficient multiclass maximum margin clustering. In *ICML*.

Zhou, G.-T.; Lan, T.; Vahdat, A.; and Mori, G. Latent maximum margin clustering. In *NIPS*.

Supplementary of Maximin Separation Probability Clustering [*]

Gao Huang[†‡], Jianwen Zhang[§], Shiji Song[†‡], Zheng Chen[§]
[†] Tsinghua National Laboratory for Information Science and Technology (TNList)
[‡] Department of Automation, Tsinghua University, Beijing 100084, China
[§] Microsoft Research, Beijing
huang-g09@mails.tsinghua.edu.cn, jiazhan@microsoft.com
shijis@tsinghua.edu.cn, zhengc@microsoft.com

Appendix A

Proof of Theorem 1

Proof. Let $\widehat{\mu}_j$ and $\widehat{\mathbf{\Sigma}}_j$ be the empirical mean and covariance of cluster $\mathcal{X}_j$ correspond to an arbitrary labeling $\mathbf{y}$ on the original data set $\mathcal{X}$. It is easy to verify that for the same labeling $\mathbf{y}$, the corresponding cluster $\mathcal{X}_j^{\mathbf{P}}$ of the transformed data set $\mathcal{X}^{\mathbf{P}}$ has a mean and covariance of $\mathbf{P}^\top\mu_j$ and $\mathbf{P}^\top\widehat{\mathbf{\Sigma}}_j\mathbf{P}$, respectively. Similarly, we have $\mathbf{\Lambda}^{\mathbf{P}} = \mathbf{P}^\top\mathbf{\Lambda}\mathbf{P}$, where $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^{\mathbf{P}}$ are respectively the regularization matrix for $\mathcal{X}$ and $\mathcal{X}^P$. Thus $\{\mathbf{P}^{-1}\mathbf{w}, \mathbf{y}\}$ always lead to the same objective value of κ on $\mathcal{X}^P$ as that yielded by $\{\mathbf{w}, \mathbf{y}\}$ on $\mathcal{X}$, for all $\mathbf{w} \in \mathcal{R}^d$ and $\mathbf{y} \in \{1,2\}^N$. It follows that if $\{\mathbf{w}^*, \mathbf{y}^*\}$ maximizes κ on $\mathcal{X}$, then $\{\mathbf{P}^{-1}\mathbf{w}^*, \mathbf{y}^*\}$ maximizes κ on $\mathcal{X}^{\mathbf{P}}$. Noticing that $\kappa = \sqrt{p/(1-p)}$, we have that the two data sets have the same optimal MSP. ■

Discussion of generalized MSP

Ackerman and Ben-David (2008) set up a set of axioms for general clustering quality measures (CQMs). Thus it is interesting to verify whether the proposed MSP satisfies the requirements of a valid CQM. However, these axioms are restricted to pairwise distance-based clustering approaches, making it difficult to directly use them to analyze MSP. Therefore, we give a generalized form of MSP, which can then be fitted into this framework. Actually, if we view $|\mathbf{w}^\top\mathbf{x}_i - \mathbf{w}^\top\mathbf{x}_j|$ as the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and replace it by the general expression $s(\mathbf{x}_i, \mathbf{x}_j)$, where $s(\cdot,\cdot)$ is a distance function over $\mathcal{X}$, then we obtain the following generalized MSP (GMSP):

$$\widetilde{p} = \frac{\widetilde{\kappa}^2}{1+\widetilde{\kappa}^2}, \quad \widetilde{\kappa} = \frac{\frac{1}{N_1 N_2}\sum_{\mathbf{x}_i\in\mathcal{X}_1}\sum_{\mathbf{x}_j\in\mathcal{X}_2} s(\mathbf{x}_i,\mathbf{x}_j)}{\sum\limits_{k=1,2}\frac{1}{N_k}\sqrt{\sum_{\mathbf{x}_i\in\mathcal{X}_k}\sum_{\mathbf{x}_j\in\mathcal{X}_k} s^2(\mathbf{x}_i,\mathbf{x}_j)}}. \quad (1)$$

We next show that GMSP is a valid clustering quality measure (CQM) under the framework of (Ackerman and Ben-David 2008), which contains the following axioms.

Definition 1. ***(Scale Invariance)*** *A quality measure m satisfies* scale invariance *if for every clustering C of $(\mathcal{X}, s)$, and every positive α, $m(C,\mathcal{X},s) = m(C,\mathcal{X},\alpha s)$.*

Definition 2. ***(Consistency)*** *A quality measure m satisfies* consistency *if for every clustering C over $(\mathcal{X}, s)$, whenever s' is a C* consistent variant *of s, then $m(C,\mathcal{X},s') \geq m(C,\mathcal{X},s)$.*

Definition 3. ***(Richness)*** *A quality measure m satisfies* richness *if for each non-trivial clustering C of $\mathcal{X}$, there exists a distance function s over X such that $C = \operatorname{argmax} m(C,\mathcal{X},s)$.*

Proposition 1. *The GMSP $\widetilde{p}(\mathbf{y},\mathcal{X},s)$ satisfies all the three CQM axioms defined above.*

Proof. We show GMSP $\widetilde{p}(\mathbf{y},\mathcal{X},s)$ satisfies the three axioms respectively.

1) Scale invariance Let s' be a distance function which satisfies $s'(\mathbf{x}_i,\mathbf{x}_j) = \alpha s(\mathbf{x}_i,\mathbf{x}_j)$ $(\alpha \in \mathcal{R}^+)$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$. It is straightforward to verify that $\widetilde{\kappa}(\mathbf{y},\mathcal{X},s) = \widetilde{\kappa}(\mathbf{y},\mathcal{X},s')$. Therefore, we have $\widetilde{p}(\mathbf{y},\mathcal{X},s) = \widetilde{p}(\mathbf{y},\mathcal{X},s')$, showing that GMSP is scale invariant.

2) Consistency Let s' be a consistent variant of s, i.e., $s'(\mathbf{x}_i,\mathbf{x}_j) \leq s(\mathbf{x}_i,\mathbf{x}_j)$, if $y_i = y_j$, and $s'(\mathbf{x}_i,\mathbf{x}_j) \geq s(\mathbf{x}_i,\mathbf{x}_j)$, if $y_i \neq y_j$. It follows that

$$\sum_{\mathbf{x}_i\in\mathcal{X}_1}\sum_{\mathbf{x}_j\in\mathcal{X}_2} s'(\mathbf{x}_i,\mathbf{x}_j) \geq \sum_{\mathbf{x}_i\in\mathcal{X}_1}\sum_{\mathbf{x}_j\in\mathcal{X}_2} s(\mathbf{x}_i,\mathbf{x}_j),$$
$$\sum_{\mathbf{x}_i\in\mathcal{X}_1}\sum_{\mathbf{x}_j\in\mathcal{X}_1} s'(\mathbf{x}_i,\mathbf{x}_j) \leq \sum_{\mathbf{x}_i\in\mathcal{X}_1}\sum_{\mathbf{x}_j\in\mathcal{X}_1} s(\mathbf{x}_i,\mathbf{x}_j),$$
$$\sum_{\mathbf{x}_i\in\mathcal{X}_2}\sum_{\mathbf{x}_j\in\mathcal{X}_2} s'(\mathbf{x}_i,\mathbf{x}_j) \leq \sum_{\mathbf{x}_i\in\mathcal{X}_2}\sum_{\mathbf{x}_j\in\mathcal{X}_2} s(\mathbf{x}_i,\mathbf{x}_j).$$

Therefore, we have $\widetilde{p}(\mathbf{y},\mathcal{X},s') \geq \widetilde{p}(\mathbf{y},\mathcal{X},s)$, showing that GMSP is consistent.

3) Richness Given a valid clustering labeling $\mathbf{y}'$ over $\mathcal{X}$, we have $\mathcal{X}_j \neq \emptyset, j = 1,2$. Define a distance function as $s(\mathbf{x}_i,\mathbf{x}_j) = 1$ for all $y'_i = y'_j$, and $s(\mathbf{x}_i,\mathbf{x}_j) = 2$ for all $y'_i \neq y'_j$. Then we can show that $\mathbf{y}' = \operatorname{argmax}_{\mathbf{y}} \widetilde{p}(\mathbf{y},\mathcal{X},s)$ for this distance function s. Actually, we have $\widetilde{\kappa}(\mathbf{y}',\mathcal{X},s) = 1$, and $\widetilde{\kappa}(\mathbf{y},\mathcal{X},s) < 1$ for all $\mathbf{y} \neq \mathbf{y}'$, since any $\mathbf{y} \neq \mathbf{y}'$ makes the nonnegative numerator and denominator of $\widetilde{\kappa}$ respectively less than 2 and larger than 2. Therefore, $\mathbf{y}'$ maximizes $\widetilde{\kappa}(\mathbf{y},\mathcal{X},s)$, and thus maximizes $\widetilde{p}(\mathbf{y},\mathcal{X},s)$. ■

Proposition 1 shows that GMSP is a valid CQM under the framework of (Ackerman and Ben-David 2008). This new

[*] This work was done during Gao Huang's internship at Microsoft Research.

CQM, i.e., GMSP, may inspire novel clustering algorithms based on pairwise distances. However, in this paper we will focus on MSP, and leave the study on GMSP to future work.

Appendix B

Proof of Theorem 2

Proof. The square of the denominator in expression of κ can be bounded from above by

$$\begin{aligned}
&\frac{1}{2}\left(\sqrt{\mathbf{w}^\top(\widehat{\boldsymbol{\Sigma}}_1+\lambda\boldsymbol{\Lambda})\mathbf{w}}+\sqrt{\mathbf{w}^\top(\widehat{\boldsymbol{\Sigma}}_2+\lambda\boldsymbol{\Lambda})\mathbf{w}}\right)^2\\
&\le \mathbf{w}^\top(\widehat{\boldsymbol{\Sigma}}_1+\lambda\boldsymbol{\Lambda})\mathbf{w}+\mathbf{w}^\top(\widehat{\boldsymbol{\Sigma}}_2+\lambda\boldsymbol{\Lambda})\mathbf{w}\\
&\le \tfrac{r_1\vee r_2}{r_1r_2}(r_1\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_1\mathbf{w}+r_2\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_2\mathbf{w})+\tfrac{\lambda(r_1\vee r_2)}{r_1r_2}\mathbf{w}^\top\boldsymbol{\Lambda}\mathbf{w}\\
&= \tfrac{r_1\vee r_2}{r_1r_2}\mathbf{w}^\top\left(\widehat{\boldsymbol{\Sigma}}-r_1r_2(\widehat{\mu}_1-\widehat{\mu}_2)(\widehat{\mu}_1-\widehat{\mu}_2)^\top\right)\mathbf{w}\\
&\quad+\tfrac{\lambda(r_1\vee r_2)}{r_1r_2}\mathbf{w}^\top\boldsymbol{\Lambda}\mathbf{w}\\
&= (r_1\wedge r_2)^{-1}\mathbf{w}^\top\left(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda}\right)\mathbf{w}\\
&\quad-(r_1\vee r_2)(\mathbf{w}^\top(\widehat{\mu}_1-\widehat{\mu}_2))^2,
\end{aligned} \tag{2}$$

where the second inequality holds since $(r_1\vee r_2)\ge 1/2$ and $(r_1\wedge r_2)\le 1/2$, and all the matrices are semi-positive definite. The first equality follows from the following property:

$$\widehat{\boldsymbol{\Sigma}}=r_1\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_1\mathbf{w}+r_2\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_2+r_1r_2(\widehat{\mu}_1-\widehat{\mu}_2)(\widehat{\mu}_1-\widehat{\mu}_2)^\top. \tag{3}$$

From (2), we obtain the lower bound for κ^2, and equality holds when $\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_1\mathbf{w}=\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_2\mathbf{w}$ and $r_1=r_2$. ■

Proof of Theorem 3

Proof. First we notice that the optimal $\mathbf{w}$ corresponds to the largest eigenvector of the GEP

$$\mathbf{X}^\top\mathbf{q}\mathbf{q}^\top\mathbf{X}\mathbf{v}=\gamma(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})\mathbf{v}, \tag{4}$$

where $\mathbf{v}$ and γ are the eigenvector and eigenvalue respectively.

Denote $\theta=(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})^{-\frac{1}{2}}\mathbf{X}^\top\mathbf{q}$, and let $\mathbf{U}\in\mathcal{R}^d$ be a unitary matrix whose first column is $\theta/\|\theta\|$. It can be verified that the matrix $\mathbf{R}=(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})^{-\frac{1}{2}}\mathbf{U}$ satisfies

$$\mathbf{R}^\top(\mathbf{X}^\top\mathbf{q}\mathbf{q}^\top\mathbf{X})\mathbf{R}=\boldsymbol{\Gamma}, \tag{5}$$

$$\mathbf{R}^\top(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})\mathbf{R}=\mathbf{I}_d, \tag{6}$$

where $\boldsymbol{\Gamma}$ is a all-zero matrix except $\boldsymbol{\Gamma}_{11}=\|\theta\|^2$.

Therefore, $\mathbf{R}$ contains all the eigenvectors of (4), and $\|\theta\|^2$ is the largest eigenvalue. According to the definition of θ, we have $\|\theta\|^2=\mathbf{q}^\top\mathbf{X}(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^\top\mathbf{q}$. Hence, the optimal $\mathbf{q}$ that maximizes the largest eigenvalue of (4) is the largest eigenvector of $\mathbf{X}(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^\top$.

Since we assume the data are centered, the covariance matrix can be expressed by $\widehat{\boldsymbol{\Sigma}}=\mathbf{X}^\top\mathbf{X}$. Finally, by the Woodbury identity, we have that

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^\top=\mathbf{I}_N-(\mathbf{I}_N+\mathbf{X}(\lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^\top)^{-1}. \tag{7}$$

■

Appendix C

Proof of Theorem 4

Proof. Given an optimal margin δ, there exists a hyperplane $\{\mathbf{w},b\}$ such that

$$\begin{cases}(\mathbf{w}^\top\mathbf{x}+b)/\mathbf{w}^\top\mathbf{w}\ge\delta/2, & \forall i\in\mathcal{X}_1\\(\mathbf{w}^\top\mathbf{x}+b)/\mathbf{w}^\top\mathbf{w}\le-\delta/2, & \forall i\in\mathcal{X}_2\end{cases} \tag{8}$$

Without loss of generality, we assume $\|\mathbf{w}\|=1$, thus we have

$$\mathbf{w}^\top\widehat{\mu}_1=\mathbf{w}^\top\Sigma_{\mathbf{x}_i\in\mathcal{X}_1}\mathbf{x}_i/N_1\ge-b+\delta/2,$$
$$\mathbf{w}^\top\widehat{\mu}_2=\mathbf{w}^\top\Sigma_{\mathbf{x}_i\in\mathcal{X}_2}\mathbf{x}_i/N_2\le-b-\delta/2.$$

This leads to

$$\mathbf{w}^\top(\widehat{\mu}_1-\widehat{\mu}_2)\ge\delta. \tag{9}$$

Meanwhile, for any $\|\mathbf{w}\|=1,\mathbf{w}\in\mathcal{R}^d$, the following inequality holds

$$\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{w}\le\sigma_{\max}^2, \tag{10}$$

where $\sigma_{\max}^2$ is the largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$.

Combining (9), (10) and the expression in Theorem 2, and omit the regularization term, we obtain

$$\kappa^2\ge\frac{\delta^2}{2(r_1\wedge r_2)^{-1}\sigma_{\max}^2-2(r_1\vee r_2)\delta^2}. \tag{11}$$

Substitute the above expression into $p=\kappa^2/(1+\kappa^2)$ yields the conclusion.

Proof of Theorem 5

Lemma 1(Marshall, Olkin, and others 1960; Lanckriet et al. 2003) Let $\mathbf{z}$ be a random vector with finite first and second order moments μ_Z and $\boldsymbol{\Sigma}_Z$, and $\mathcal{S}$ a given convex set, then

$$\sup_{\mathbf{z}\sim(\mu_{\mathbf{z}},\boldsymbol{\Sigma}_{\mathbf{z}})}\Pr(\mathbf{z}\in\mathcal{S})=1/(1+s^2), \tag{12}$$

where $s^2=\inf_{\mathbf{z}\in\mathcal{S}}(\mathbf{z}-\mu_{\mathbf{z}})^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}(\mathbf{z}-\mu_{\mathbf{z}})$, and the supremum is taken over all distributions of $\mathbf{z}$ with the given mean μ_Z and covariance $\boldsymbol{\Sigma}_Z$.

Proof. Let $\{\mathbf{w},b\}$ be the optimal hyperplane obtained by MSPC. Without loss of generality, we assume $\mathbf{w}^\top(\mu_1-\mu_2)>0$, and denote $\Delta=\mathbf{w}^\top(\mu_1-\mu_2)$ and $b'=\Delta/2-\mathbf{w}^\top\mu_1$ (or equivalently, $b'=-\Delta/2-\mathbf{w}^\top\mu_2$). Notice that generally $b\neq b'$.

Since we assume that $\mathbf{x}_i\in\mathcal{X}_1$ and $\mathbf{x}_i\in\mathcal{X}_2$ are respectively drawn from two independent distributions, their corresponding predicts $\{y|y_i=\mathbf{w}^\top\mathbf{x}_i+b',\mathbf{x}_i\in\mathcal{X}_1\}$ and $\{y|y_i=\mathbf{w}^\top\mathbf{x}_i+b',\mathbf{x}_i\in\mathcal{X}_2\}$ can be modeled by two independent random variables Y_1 and Y_2 respectively, whose first and second order moments are be expressed by

$$\begin{gathered}\widehat{\mu}_{Y_1}=\mathbf{w}^\top\widehat{\mu}_1+b'=\Delta/2,\quad\widehat{\mu}_{Y_2}=\mathbf{w}^\top\widehat{\mu}_2+b'=-\Delta/2,\\ \widehat{\sigma}_{Y_1}^2=\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_1\mathbf{w},\quad\widehat{\sigma}_{Y_2}^2=\mathbf{w}^\top\widehat{\boldsymbol{\Sigma}}_2\mathbf{w},\end{gathered} \tag{13}$$

where $\widehat{\mu}_j$ and $\widehat{\boldsymbol{\Sigma}}_j$ are the mean and covariance of $\mathcal{X}_j(j=1,2)$, respectively.

Using Lemma 1, and let $\mathcal{S} = \{Y_1 \leq \rho\Delta/2\}$ $(0 \leq \rho \leq 1)$, we have

$$\sup_{Y_1 \sim (\widehat{\mu}_{Y_1}, \widehat{\sigma}_{Y_1})} \Pr(Y_1 \leq \rho\Delta/2) = 1/(1+s^2), \tag{14}$$

where the minimum distance is $s = (1-\rho)\Delta/(2\widehat{\sigma}_{Y_1})$.

Thus we have

$$\inf_{Y_1 \sim (\widehat{\mu}_{Y_1}, \widehat{\sigma}_{Y_1})} \Pr(Y_1 \geq \rho\Delta/2) = \frac{(1-\rho)^2\Delta^2}{4\widehat{\sigma}_{Y_1}^2 + (1-\rho)^2\Delta^2}, \tag{15}$$

From the MPM formulation, we have that

$$\widehat{\sigma}_{Y_1} = \sqrt{\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_1 \mathbf{w}} = \Delta/\kappa - \sqrt{\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_2 \mathbf{w}} \leq \Delta/\kappa. \tag{16}$$

Combining (15) and (16), we have

$$\Pr(Y_1 \geq \rho\Delta/2) \geq \frac{(1-\rho)^2}{4/\kappa^2 + (1-\rho)^2}, \tag{17}$$

This means that at least $1 - r = (1-\rho)^2/(4\kappa^{-2} + (1-\rho)^2)$ fraction of the points in $\mathcal{X}_1$ satisfy $|\mathbf{w}^\top \mathbf{x}_i + b| \geq \rho\Delta/2$. Similarly, we can show that at least $1 - r$ fraction of the points in $\mathcal{X}_2$ satisfy $|\mathbf{w}^\top \mathbf{x}_i + b| \geq \rho\Delta/2$ as well. Using the relation $p = \kappa^2/(1+\kappa^2)$, we conclude the proof. ■

References

Ackerman, M., and Ben-David, S. 2008. Measures of clustering quality: A working set of axioms for clustering. In *NIPS*.

Lanckriet, G. R.; Ghaoui, L. E.; Bhattacharyya, C.; and Jordan, M. I. 2003. A robust minimax approach to classification. *The Journal of Machine Learning Research* 3:555–582.

Marshall, A. W.; Olkin, I.; et al. 1960. Multivariate chebyshev inequalities. *The Annals of Mathematical Statistics* 31(4):1001–1014.