A general model for plane-based clustering with loss function

Zhen Wang, Yuan-Hai Shao, Lan Bai, Chun-Na Li, and Li-Ming Liu

***Abstract*—In this paper, we propose a general model for plane-based clustering. The general model contains many existing plane-based clustering methods, e.g., k-plane clustering (kPC), proximal plane clustering (PPC), twin support vector clustering (TWSVC) and its extensions. Under this general model, one may obtain an appropriate clustering method for specific purpose. The general model is a procedure corresponding to an optimization problem, where the optimization problem minimizes the total loss of the samples. Thereinto, the loss of a sample derives from both within-cluster and between-cluster. In theory, the termination conditions are discussed, and we prove that the general model terminates in a finite number of steps at a local or weak local optimal point. Furthermore, based on this general model, we propose a plane-based clustering method by introducing a new loss function to capture the data distribution precisely. Experimental results on artificial and public available datasets verify the effectiveness of the proposed method.**

***Index Terms*—Unsupervised learning, plane-based clustering, general model, twin support vector clustering, loss function.**

I. Introduction

CLUSTERING, discovering the similarity among the data samples, is one of the most important unsupervised learning topics [1]–[3]. Many approaches assign the samples into the clusters via certain cluster centers [4]–[9]. The plane-based clustering treats the cluster center as a plane, and thus it is able to find the plane-based shape clusters. Moreover, the plane-based clustering can be extended to nonlinear manifold modeling easily to cope with complex data structures. The plane-based clustering has attracted much attention [8], [10]–[16].

The first plane-based clustering, k-plane clustering (kPC) [8], was proposed by O.L. Mangasarian et al., where the discriminative information from within-cluster was considered. Subsequently, the discriminative information from between-cluster has been introduced in plane-based clustering. For instance, proximal plane clustering (PPC) [11] and twin support vector clustering (TWSVC) [12] considered that the cluster center plane should be not only as close as possible to the current cluster samples but also far away from the other clusters. Still later, robust twin support vector clustering (RTWSVC) and fast robust twin support vector clustering (FRTWSVC) were also appeared [16]. Until recently, ramp-based twin support vector clustering (RampTWSVC) [17] was proposed to deal with noise or outliers. So it is interesting to find a cluster center plane by considering the discriminative information both from within-cluster and between-cluster.

Let us notice the close relationship between the cluster problem and the classification problem. In fact, there are the following corresponding relationships between them: PPC corresponds to the generalized eigenvalue proximal support vector machines (GEPSVM) [18], TWSVC to the twin support vector machines (TWSVM) [19], [20], RTWSVC to the L_1-TWSVM [21], FRTWSVC to the L_1 least square TWSVM [22], and RampTWSVC to the best fitting hyperplanes for classification (BFHC) [23]. So it seems like a great way to relate the plane-based clustering to the supervised learning.

Briefly speaking, the supervised learning is essentially based on two concepts "loss function" and "regularization" [21], [22], [24]–[30]. We find that the plane-based clustering can also be established in a similar way. This yields our general model. It is concerned with the new defined loss function from discriminative information [31] and the regularization. The general model iteratively implements two parts: cluster update and cluster assignment. In the cluster update, the new cluster center planes would be obtained by minimizing the loss derived from the current cluster assignment. Besides, in the cluster assignment, each sample would be assigned to the cluster with the least loss. For the general model, it is allowed to select various loss functions and regularization terms, and most of the existing plane-based clustering methods can be regarded as the particular cases with different selections. Furthermore, a new plane-based clustering is derived from the general model. More precisely, following the model, we propose a robust fitting distribution planes clustering (RFDPC) by hiring a new loss function, the ramp loss [23] combined with certain statistics, which owns clear geometric meaning and captures the data distribution.

The main contributions of this paper include:

(i) A general model for plane-based clustering is proposed, in which different loss functions and regularization terms can be chosen, particularly yielding the existing kPC, PPC, TWSVC, RTWSVC, FRTWSVC, RampTWSVC, and etc.

(ii) The cluster update and cluster assignment in the general model is consistent on minimizing the loss of samples, result-

Submitted in January 29, 2019. This work is supported in part by National Natural Science Foundation of China (Nos. 11501310, 61866010, 11871183, and 61703370), in part by Natural Science Foundation of Hainan Province (No. 118QN181), and in part by Scientific Research Foundation of Hainan University (No. kyqd(sk)1804).

Zhen Wang is with School of Mathematical Sciences, Inner Mongolia University, Hohhot, 010021, P.R.China e-mail: wangzhen@imu.edu.cn.

Yuan-Hai Shao (*Corresponding author) is with School of Economics and Management, Hainan University, Haikou, 570228, P.R.China e-mail: shaoyuanhai21@163.com.

Lan Bai is with School of Mathematical Sciences, Inner Mongolia University, Hohhot, 010021, P.R.China e-mail: imubailan@163.com.

Chun-Na Li is with Zhijiang College, Zhejiang University of Technology, Hangzhou, 310024, P.R.China e-mail: na1013na@163.com.

Li-Ming Liu is with School of Statistics, Capital University of Economics and Business, Beijing, 100070, P.R.China. e-mail: llm5609@163.com.

ing in its finite termination at a local or weak local optimal point.
(iii) A new loss function is introduced in the general model with named RFDPC, to cope with outliers, noise, and capture the data distribution more precisely.
(iv) Experiments show the amazing performance of RFDPC compared with the existing plane-based clustering methods.

The rest of this paper is organized as follows. The general model is elaborated in section II. Some plane-based clustering methods are summarised under the general model in section III. A novel plane-based clustering method (RFDPC) is described in section IV. Experiments and conclusions are presented in sections V and VI, respectively.

II. The General Model for Plane-Based Clustering

A. Formulation

Remind the clustering problem with m data samples $\{x_1, x_2, \ldots, x_m\}$ in the n-dimensional real vector space R^n, which is represented by $X \in R^{n\times m}$. Assume that these m samples belong to k clusters with their corresponding labels $y \in \{1, 2, \ldots, k\}$. Our task is to assign the m samples into k clusters, or to give their cluster labels

$$\mathbf{y} = (y(x_1), y(x_2), \ldots, y(x_m))^\top. \tag{1}$$

For partition-based clustering [7], [32]–[34], the usual way is to find the cluster labels $\mathbf{y}$ as well as the k cluster centers. The plane-based clustering treats each cluster center as a plane. The k cluster center planes are described as

$$w_j^\top x + b_j = 0,\ j = 1, \ldots, k, \tag{2}$$

where $w_j \in R^n$ is the weight vector and $b_j \in R$ is the bias term. Consider the deviation of a sample x from the j-th cluster center plane $w_j^\top x + b_j = 0$ ($j = 1, \ldots, k$). For instance, the deviation can be measured by the signed distance of x to the plane as

$$f(x; w_j, b_j) = \frac{w_j^\top x + b_j}{||w_j||}, \tag{3}$$

where $||\cdot||$ denotes L_2 norm. Another simpler way to reduce computation is to hire

$$f(x; w_j, b_j) = w_j^\top x + b_j. \tag{4}$$

Combining these k deviation functions, either (3) or (4), yields the vector function

$$F(x; W, \mathbf{b}) = (f(x; w_1, b_1), f(x; w_2, b_2), \ldots, f(x; w_k, b_k))^\top, \tag{5}$$

where $W = (w_1, \ldots, w_k)$ and $\mathbf{b} = (b_1, \ldots, b_k)^\top$. Thus, the k cluster center planes can be represented as

$$F(x; W, \mathbf{b}) = 0. \tag{6}$$

One of the popular approaches [8], [11], [12] is to find the cluster labels $\mathbf{y}$ and the k cluster center planes (6) iteratively. Start with an initial assignment $\mathbf{y}$. Next, for the given $\mathbf{y}$, find the corresponding $F(x; W, \mathbf{b})$ by establishing and solving an optimization problem. Then, update $\mathbf{y}$ and the vector function $F(x; W, \mathbf{b})$ alternately until certain termination conditions are satisfied.

The key point of this paper is to introduce the loss function into a general optimization problem. For the ith sample x_i with assigned label $y(x_i)$, the ideal case is to find $W^* = (w_1^*, \ldots, w_k^*)$ and $\mathbf{b}^* = (b_1^*, \ldots, b_k^*)^\top$ such that, on one hand the sample x_i lies exactly on the center plane $f(x_i; w_{y(x_i)}^*, b_{y(x_i)}^*) = 0$, and on the other hand, the sample x_i is far away from other center planes (in extremity, $f(x_i; w_j^*, b_j^*) = \pm\infty$, $j \neq y(x_i)$). For the actual situation, the loss of sample x_i should be a measure of the deviation from the ideal case. Therefore, it should consist of two parts: (i) for its own center plane, the loss should depend on the deviation $f(x_i; w_{y(x_i)}, b_{y(x_i)})$ and can be measured by a within-cluster function $J^w(f(x_i; w_{y(x_i)}, b_{y(x_i)}))$, where $J^w(\rho)$ is a function from R to R with the condition $J^w(0) = 0$; (ii) for other center planes, the loss can be measured by a between-cluster function $J^b(f(x_i; w_j, b_j))$ with $j \neq y(x_i)$, where $J^b(\rho)$ is a function from R to R. Thus, for the sample x_i ($i = 1, \ldots, m$), the loss is described by

$$\begin{aligned} &L(y(x_i); F(x_i; W, \mathbf{b})) = c_w J^w(f(x_i; w_{y(x_i)}, b_{y(x_i)})) + \\ &c_b \sum_{\substack{j=1 \\ j\neq y(x_i)}}^{k} J^b(f(x_i; w_j, b_j)), \end{aligned} \tag{7}$$

where c_w and c_b are positive parameters. Furthermore, for the dataset $X = \{x_1, \ldots, x_m\}$, the total loss is

$$L(\mathbf{y}; F(x_1; W, \mathbf{b}), \ldots, F(x_m; W, \mathbf{b})) = \sum_{i=1}^{m} L(y(x_i); F(x_i; W, \mathbf{b})). \tag{8}$$

This leads to the optimization problem for both $\mathbf{y}$ and $(W, \mathbf{b})$ as

$$\min_{\mathbf{y}, W, \mathbf{b}} \ G(\mathbf{y}; W, \mathbf{b}) = L(\mathbf{y}; F(x_1; W, \mathbf{b}), \ldots, F(x_m; W, \mathbf{b})) + ||F||_{\mathcal{F}}, \tag{9}$$

where $||\cdot||_{\mathcal{F}}$ denotes the regularization term in the functional space $\mathcal{F}$.

Problem (9) is very similar to the optimization problem in supervised learning, i.e., it consists of the loss and regularization, but their concerns are not the same. For supervised learning, it aims at predicting the unknown samples by minimizing the loss of the training samples. However, for clustering, we focus on minimizing the loss of the given m samples, and the regularization makes this efficient. Based on problem (9), the general model for plane-based clustering is constructed in Model 1.

Model 1 The general model

Input: Dataset X, the within-cluster function $J^w(\rho)$, the between-cluster function $J^b(\rho)$, and the parameters c_w, c_b.
Output: $\mathbf{y}^*$ and $(W^*, \mathbf{b}^*)$.
1. Initialize the sample labels $\mathbf{y}^{(0)} = (y^{(0)}(x_1), \ldots, y^{(0)}(x_m))^\top$.
2. **For** $t = 0, 1, 2, \ldots$, compute $W^{(t)}$, $\mathbf{b}^{(t)}$ and $\mathbf{y}^{(t+1)}$ by the following steps:

(a) Cluster update: For the current $\mathbf{y}^{(t)}$, $(W^{(t)}, \mathbf{b}^{(t)})$ is set to be the solution to the optimization problem

$$\min_{W, \mathbf{b}} \ G(\mathbf{y}^{(t)}; W, \mathbf{b}), \tag{10}$$

where $G(\cdot)$ is given by (9), or equivalently the solutions to k subproblems with $j = 1, \ldots, k$ as follow:

$$\begin{aligned} \min_{w_j, b_j} \; & c_w \sum_{\substack{i=1 \\ y^{(t)}(x_i)=j}}^{m} J^w(f(x_i; w_j, b_j)) + \\ & c_b \sum_{\substack{i=1 \\ y^{(t)}(x_i)\neq j}}^{m} J^b(f(x_i; w_j, b_j)) + ||f(x_i; w_j, b_j)||_{\mathcal{F}}. \end{aligned} \tag{11}$$

(b) Cluster assignment: For the current $(W^{(t)}, \ \mathbf{b}^{(t)})$, the labels $\mathbf{y}^{(t+1)}$ are set to be the solution of the following optimization problem

$$\min_{\mathbf{y}} \; G(\mathbf{y}, W^{(t)}, \mathbf{b}^{(t)}), \tag{12}$$

or equivalently, the labels $\mathbf{y}^{(t+1)}$ are given by

$$y(x_i) = \arg\min_{j} \; L(j; F(x_i; W^{(t)}, \mathbf{b}^{(t)})) \tag{13}$$

with $i = 1, \ldots, m$. If there is a tie, the cluster with the smallest label number is selected.

(c) Repetitiveness check: If $(W^{(t)}, \ \mathbf{b}^{(t)})$ is a solution to problem (10) where $\mathbf{y}^{(t)}$ is replaced by $\mathbf{y}^{(t+1)}$, break the loop and go to step 3.

(d) If the termination condition is satisfied, go to step 3; otherwise, set $t = t + 1$ and back to step 2(a).

3. Set $\mathbf{y}^* = \mathbf{y}^{(t)}$, $W^* = W^{(t)}$, $\mathbf{b}^* = b^{(t)}$.

Remark. In step 1 of Model 1, a common way is to assign the samples into k clusters randomly, resulting in unstable clustering performance. It is preferable to choose some stable initialization techniques, e.g., nearest neighbor graph (NNG) [12], which has been successfully applied to several plane-based clustering methods [12], [16], [17]. In step 2(a), if there are many global solutions can be obtained, the same ones are selected for the same $\mathbf{y}$. However, we may only get a local solution. Note that there has been $(W^{(t-1)}, \ \mathbf{b}^{(t-1)})$ (for $t \geq 1$) before solving problem (10). The local solution in step 2(a) must be not worse than previous solution if a local solution is inevitable. Thus, it is a good choice to hire the $(t-1)$-th local solution as the initial point of the t-th problem in step 2(a), if the assumption is false in step 2(c). In other words, the inequality $G(\mathbf{y}^{(t)}, W^{(t-1)}, \mathbf{b}^{(t-1)}) \geq G(\mathbf{y}^{(t)}, W^{(t)}, \mathbf{b}^{(t)})$ always holds in iteration.

Besides, the functions $J^w(\rho)$ and $J^b(\rho)$ should also be pre-defined. Obviously, it is reasonable to select them with the following properties.

Properties :

(i) $J^w(\rho) = J^w(-\rho)$ and $J^b(\rho) = J^b(-\rho)$.

(ii) $J^w(\rho)$ is monotonically non-decreasing in $[0, \infty)$.

(iii) $J^b(\rho)$ is monotonically non-increasing in $[0, \infty)$.

In this case, we have following theorem.

Theorem II.1. *If the within-cluster function $J^w(\rho)$ and the between-cluster function $J^b(\rho)$ satisfy the above three properties (i)-(iii), then the sample assignment* (13) *can be simplified as*

$$y(x_i) = \arg\min_{j} |f(x_i; w_j, b_j)|, \tag{14}$$

where $|\cdot|$ denotes the absolute value.

Proof. Suppose that, for an arbitrary sample x, l^* is the label of x obtained by (14), i.e., $|f(x; w_{l^*}, b_{l^*})|$ is the smallest one in $\{|f(x; w_1, b_1)|, \ldots, |f(x; w_k, b_k)|\}$, and suppose l is an arbitrary label of x. The objective values of (13) at l^* and l are

$$\begin{aligned} L(l^*; F) &= c_w J^w(f(x; w_{l^*}, b_{l^*})) + c_b J^b(f(x; w_l, b_l)) \\ &+ c_b \sum_{\substack{j=1 \\ j\neq l^*, j\neq l}}^{k} J^b(f(x; w_j, b_j)), \end{aligned} \tag{15}$$

and

$$\begin{aligned} L(l; F) &= c_w J^w(f(x; w_l, b_l)) + c_b J^b(f(x; w_{l^*}, b_{l^*})) \\ &+ c_b \sum_{\substack{j=1 \\ j\neq l^*, j\neq l}}^{k} J^b(f(x; w_j, b_j)), \end{aligned} \tag{16}$$

respectively.

From the properties (i) and (ii), we have $J^w(f(x; w_{l^*}, b_{l^*})) \leq J^w(f(x; w_l, b_l))$. Similarly, we have $J^b(f(x; w_l, b_l)) \leq J^b(f(x; w_{l^*}, b_{l^*}))$ from the properties (i) and (iii). Thus, $L(l^*; F) \leq L(l; F)$ because of the positive c_w and c_b. This implies that l^* corresponds to the smallest objective value of (13). □

Now, we extend the general model to the nonlinear case via a kernel trick [4], [12], [35], [36]. For the nonlinear manifold clustering, the k cluster centers are defined as

$$w_j^\top \phi(x) + b_j = 0, \ j = 1, \ldots, k, \tag{17}$$

where $\phi(\cdot)$ is a pre-defined nonlinear mapping. Thus, the deviation of a sample from a cluster center depends on the nonlinear mapping $\phi(\cdot)$ strictly. Generally, it is not necessary to give the explicit nonlinear mapping $\phi(\cdot)$. Note that the deviation in general model is just considered. There are many kernel tricks to estimate the deviation. For instance, the deviation $f(\phi(x); w_j, b_j)$ can be estimated by $f(K(x, X); w_j, b_j)$, where $K(\cdot, \cdot)$ is a predetermined kernel function [35] and $K(x_1, x_2) =< \phi(x_1), \phi(x_2) >$ ($< \cdot, \cdot >$ denotes the inner product). By selecting an appropriate kernel function, the nonlinear general model can be obtained without any difficulty, so the details are omitted.

B. Analysis

In this subsection, the termination conditions of the above general model are analysed. More exactly, it is concerned with the following three termination conditions.

Termination Conditions :

(i) It happens that there is a repeated overall assignment of samples to clusters, i.e., $\mathbf{y}^{(p)} = \mathbf{y}^{(q)}$ where $p \neq q$ [8], [11].

(ii) It happens that there is a non-decrease in the objective function $G(\cdot)$ [8], [11].

(iii) Both the cases (i) and (ii) happen.

Corresponding to the different meaning of the solution in step 2(a), we have the following two theorems.

Theorem II.2. *Under either termination condition (i) or (ii), the general model terminates in a finite number of steps if the solution in step 2(a) means global solution.*

Proof. The iterations in the general model can be summarized as

$$\begin{aligned}&\mathbf{y}^{(0)} \to (W^{(0)}, \mathbf{b}^{(0)}) \to \mathbf{y}^{(1)} \to (W^{(1)}, \mathbf{b}^{(1)}) \to \cdots \\ &\to \mathbf{y}^{(t)} \to (W^{(t)}, \mathbf{b}^{(t)}) \to \cdots .\end{aligned} \tag{18}$$

Since there are a finite number of ways that the m samples can be assigned to k clusters, there are two integers $t, p > 0$ such that $\mathbf{y}^{(t)} = \mathbf{y}^{(t+p)}$. Therefore, the general model terminates in a finite number of steps under termination condition (i).

Moreover, the corresponding $(W^{(t)}, \mathbf{b}^{(t)})$ and $(W^{(t+p)}, \mathbf{b}^{(t+p)})$ are the global solutions to the same optimization problem (10). Thus, we have $G(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)}) = G(\mathbf{y}^{(t+p)}; W^{(t+p)}, \mathbf{b}^{(t+p)})$. Note that the global solution in step 2(a) guarantees that the objective $G(\mathbf{y}; W, \mathbf{b})$ is non-increasing in iteration. Then we have

$$\begin{aligned}&G(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)}) = G(\mathbf{y}^{(t+1)}; W^{(t)}, \mathbf{b}^{(t)}) = G(\mathbf{y}^{(t+1)}; \\ &W^{(t+1)}, \mathbf{b}^{(t+1)}) = \cdots = G(\mathbf{y}^{(t+p)}; W^{(t+p)}, \mathbf{b}^{(t+p)}).\end{aligned} \tag{19}$$

Therefore, the general model terminates in a finite number of steps under termination condition (ii). □

Theorem II.3. *Suppose the number of the local solutions or the local optimal values to the problem* $\min\limits_{W,\mathbf{b}} G(\mathbf{y}; W, \mathbf{b})$ *is finite. Under termination condition (iii), the general model terminates in a finite number of steps if the solution in step 2(a) means local solution.*

Proof. Consider sequence (18). Since there are a finite number of ways that the m samples can be assigned to k clusters, we can find a subsequence of $\mathbf{y}$ from (18) in which the elements are the same. Based on the assumptions, there are two integers $t, p > 0$ such that $G(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)}) = G(\mathbf{y}^{(t)}; W^{(t+p)}, \mathbf{b}^{(t+p)})$, where $\mathbf{y}^{(t)}$ belongs to the above subsequence of $\mathbf{y}$. Since the objective G is non-increasing in the iteration, it is invariable from the step t to $t+p$. Therefore, the general model terminates before or at step $t + p$ under termination condition (iii). □

Generally speaking, the general model may also terminate in a finite number of steps with other termination conditions. However, the termination point obtained by the general model would be very different under different termination conditions. In fact, when the general model terminates, there should not be any other available points which make the objective function $G(\cdot)$ decrease. To study the convergence of the general model further, we introduce two definitions.

Definition II.1. *(***Local optimal point** *by O.L. Mangasarian in [8]) Point* $(\mathbf{y}^*; W^*, \mathbf{b}^*)$ *is defined as the local optimal point to the function* $G(\mathbf{y}; W, \mathbf{b})$ *if* $\mathbf{y}^*$ *is the global solution to the problem* $\min\ G(\mathbf{y}; W^*,\ \mathbf{b}^*)$*, and meanwhile* $(W^*,\ \mathbf{b}^*)$ *is the global solution to the problem* $\min\ G(\mathbf{y}^*; W,\ \mathbf{b})$.

Definition II.2. *(***Weak local optimal point***) Point* $(\mathbf{y}^{**}; W^{**}, \mathbf{b}^{**})$ *is defined as the weak local optimal point to the function* $G(\mathbf{y}; W, \mathbf{b})$ *if* $\mathbf{y}^{**}$ *is the global solution to the problem* $\min\ G(\mathbf{y}; W^{**},\ \mathbf{b}^{**})$*, and meanwhile* $(W^{**},\ \mathbf{b}^{**})$ *is a local solution to the problem* $\min\ G(\mathbf{y}^{**}; W,\ \mathbf{b})$.

Now, we have the following two theorems.

Theorem II.4. *The general model with termination condition (i) or (ii) terminates at a local optimal point if the solution in step 2(a) means global solution.*

Proof. From the proof of Theorem 2.2, there is a finite number $t > 0$ such that equations (19) hold. Thus, the point $(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)})$ is a local optimal point and $(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)}) = (\mathbf{y}^{(t+1)}; W^{(t+1)}, \mathbf{b}^{(t+1)})$. Then, the general model terminates at step $t+1$ under termination condition (i) or (ii). □

Theorem II.5. *Suppose the number of the local solutions or the local optimal values to the problem* $\min\limits_{W,\mathbf{b}} G(\mathbf{y}; W, \mathbf{b})$ *is finite. The general model with termination condition (iii) terminates at a weak local optimal point if the solution in step 2(a) means local solution.*

Proof. From the proof of Theorem 2.3, there are two finite integers $t, p > 0$ such that $G(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)}) = G(\mathbf{y}^{(t)}; W^{(t+p)}, \mathbf{b}^{(t+p)})$. Due to the non-increase of the objective G in the iteration, equations (19) hold and $\mathbf{y}^{(t)} = \mathbf{y}^{(t+p)}$. Note that $\mathbf{y}^{(t+1)}$ is the global solution to the problem $\min\limits_{\mathbf{y}} G(\mathbf{y}; W^{(t)}, \mathbf{b}^{(t)})$, and $G(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)})$ also attains the same optimal value. It shows that $\mathbf{y}^{(t)}$ is also the global solution to the above problem. Thus, $\mathbf{y}^{(t)} = \mathbf{y}^{(t+1)}$ holds because of the uniqueness of the assigned labels guaranteed in step 2(b), and then $(\mathbf{y}^{(t)}; W^{(t)}, \mathbf{b}^{(t)})$ is a weak local optimal point. Therefore, the conclusion holds by Theorem 2.3. □

III. Reorganization of the plane-based clustering methods

In this section, we show that the general model yields current plane-based clustering methods by selecting different deviation formations and loss functions.

A. kPC

kPC [8] is the first plane-based clustering method. It starts with a random assignment of the samples. Then, for the jth cluster ($j = 1, \ldots, k$), its cluster center (2) requires the samples be along with it by solving the following problem

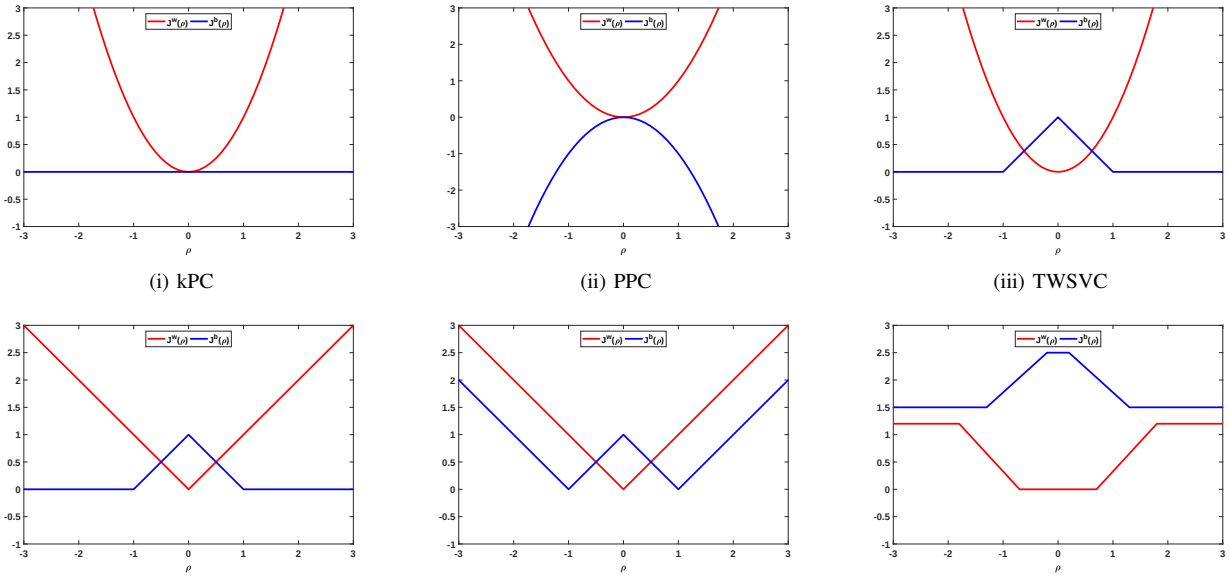
$$\begin{aligned}\min_{w_j, b_j}\quad &\sum_{\substack{i=1\\ y(x_i)=j}}^{m} (w_j^\top x_i + b_j)^2. \\ \text{s.t.}\quad &||w_j|| = 1.\end{aligned} \tag{20}$$

When the k cluster centers are obtained, the samples are reassigned to k clusters by

$$y(x_i) = \arg\min_j \frac{|w_j^\top x_i + b_j|}{||w_j||}. \tag{21}$$

The cluster centers and the samples' labels are updated alternately until termination condition (i) or (ii) is satisfied.

To organize kPC by the general model, we select $f(x; w_j, b_j) = \frac{w_j^\top x + b_j}{||w_j||}$ and hire the within-cluster function $J_1^w(\rho) = \rho^2$ and between-cluster function $J_1^b(\rho) = 0$ with



(i) kPC (ii) PPC (iii) TWSVC

(iv) RTWSVC (v) FRTWSVC (vi) RampTWSVC

Fig. 1. Within-cluster function $J^w(\rho)$ and between-cluster function $J^b(\rho)$ selected in different plane-based clustering methods, where $c_w = c_b = 1,\ \Delta = 0.3$, and $s = -0.2$.

$c_w = 1$ (see Fig. 1(i)). Thus, the loss of the sample x_i $(i = 1, \ldots, m)$ is

$$L_1(y(x_i); F(x_i; W, \mathbf{b})) = (w_{y(x_i)}^\top x_i + b_{y(x_i)})^2. \tag{22}$$

Without any difficulty, we can use the general model to generate a plane-based clustering method by using the loss function (22). By setting $\tilde{w}_j = w_j/||w_j||$ and $\tilde{b}_j = b_j/||w_j||$, problem (11) solved in the general model is equivalent to problem (20) in kPC. Since $J_1^w(\rho)$ and $J_1^b(\rho)$ satisfy the conditions of Theorem 2.1, it is easy to conclude that kPC is consistent with the general model by the loss function (22). The global solution to problem (20) can be obtained by solving an eigenvalue problem, and we immediately conclude that kPC finitely terminates at a local optimal point by Theorem 2.4 (this finite termination has been proven by Mangasarian, see Theorem 7 in [8]).

It is worth to notice that kPC only considers the discriminative information from within-cluster. The following PPC was proposed by introducing the discriminative information from between-cluster.

B. PPC

The procedure of PPC [10], [11] is similar to kPC, where the only difference is the stage of reconstructing the cluster centers. PPC considers the samples from the current cluster should close to its cluster center, and meanwhile the samples from different clusters should be far away from it. The jth $(j = 1, \ldots, k)$ cluster center plane is obtained by solving following problem

$$\begin{aligned} \min_{w_j, b_j} \quad & \sum_{\substack{i=1 \\ y(x_i)=j}}^{m} (w_j^\top x_i + b_j)^2 - c \sum_{\substack{i=1 \\ y(x_i)\neq j}}^{m} (w_j^\top x_i + b_j)^2 \\ \text{s.t.} \quad & ||w_j|| = 1, \end{aligned} \tag{23}$$

where c is a positive parameter.

Similarly, to organize PPC by the general model, we select $f(x; w_j, b_j) = \frac{w_j^\top x + b_j}{||w_j||}$ and hire the functions $J_2^w(\rho) = \rho^2$ and $J_2^b(\rho) = -\rho^2$ with $c_w = 1,\ c_b = c$ (see Fig. 1(ii)). Thus, the loss of the sample x_i $(i = 1, \ldots, m)$ is

$$\begin{aligned} L_2(y(x_i); F(x_i; W, \mathbf{b})) = (w_{y(x_i)}^\top x_i + b_{y(x_i)})^2 - \\ c \sum_{\substack{j=1 \\ j\neq y(x_i)}}^{k} (w_j^\top x_i + b_j)^2. \end{aligned} \tag{24}$$

Obviously, $J_2^w(\rho)$ and $J_2^b(\rho)$ satisfy the conditions of Theorem 2.1. Therefore, PPC can be regarded as the general model by using the loss function (24). Since the global solution to problem (23) can be obtained by solving an eigenvalue problem, we can immediately conclude that PPC finitely terminates at a local optimal point by Theorem 2.4, which was not provided previously. By the loss function (24), it can be seen that PPC uses L_2 norm to measure the discriminative information from between-cluster, which may be sensitive with noise or outliers.

C. TWSVC

To reduce the influence of the noise and outliers, TWSVC [12] makes the samples from different clusters far away from the cluster center to a certain distance. The jth $(j = 1, \ldots, k)$ cluster center is considered from following problem

$$\begin{aligned} \min_{w_j, b_j, \xi_i} \quad & \sum_{\substack{i=1 \\ y(x_i)=j}}^{m} (w_j^\top x_i + b_j)^2 + c \sum_{\substack{i=1 \\ y(x_i)\neq j}}^{m} \xi_i \\ \text{s.t.} \quad & |w_j^\top x_i + b_j| \geq 1 - \xi_i, \xi_i \geq 0, y(x_i) \neq j, \\ & i = 1, \ldots, m, \end{aligned} \tag{25}$$

where $\xi_i \in R$ is a slack variable.

By selecting $f(x; w_j, b_j) = w_j^\top x + b_j$ and hiring the functions $J_3^w(\rho) = \rho^2$ and $J_3^b(\rho) = (1 - |\rho|)_+$ with $c_w = 1$, $c_b = c$ (see Fig. 1(iii)), the loss of the sample x_i $(i = 1, \ldots, m)$ is

$$L_3(y(x_i); F(x_i; W, \mathbf{b})) = (w_{y(x_i)}^\top x_i + b_{y(x_i)})^2 + c \sum_{\substack{j=1 \\ j \neq y(x_i)}}^{k} (1 - |w_j^\top x_i + b_j|)_+, \tag{26}$$

where $(\cdot)_+$ replaces the negative value with zero. Obviously, $J_3^w(\rho)$ and $J_3^b(\rho)$ satisfy the conditions of Theorem 2.1. Therefore, TWSVC can be regarded as our general model by using the loss function (26) except a slight difference in the solution to problem (25), which is obtained independently. It is worth to mention that if TWSVC is implemented by the general model strictly, it would terminate in a finite number of steps at a weak local optimal point by Theorem 2.5.

D. Extensions on TWSVC

There are several extensions on TWSVC due to its stable performance. For instance, RTWSVC [16] replaces L_2 norm with L_1 norm in the within-cluster function, resulting in decreasing the influence of the noise and outliers further. Another extension FRTWSVC [16] uses a least squares formation to accelerate the learning speed. The third extension RampTWSVC [17] introduces the ramp loss function into TWSVC to further decrease the influence of the noise and outliers from both within-cluster and between-cluster. They construct the cluster centers by different optimization problems. By selecting $f(x; w_j, b_j) = w_j^\top x + b_j$, we summarize their within-cluster, between-cluster and loss functions (see Fig. 1(iv)-(vi)) as follows.

RTWSVC:

$$J_4^w(\rho) = |\rho|, \quad J_4^b(\rho) = (1 - |\rho|)_+,$$
$$L_4(y(x_i); F(x_i; W, \mathbf{b})) = |w_{y(x_i)}^\top x_i + b_{y(x_i)}| + c \sum_{\substack{j=1 \\ j \neq y(x_i)}}^{k} (1 - |w_j^\top x_i + b_j|)_+;$$

FRTWSVC:

$$J_5^w(\rho) = |\rho|, \quad J_5^b(\rho) = |1 - |\rho||,$$
$$L_5(y(x_i); F(x_i; W, \mathbf{b})) = |w_{y(x_i)}^\top x_i + b_{y(x_i)}| + c \sum_{\substack{j=1 \\ j \neq y(x_i)}}^{k} |1 - |w_j^\top x_i + b_j||;$$

and RampTWSVC:

$$J_6^w(\rho) = \begin{cases} 0 & \text{if } |\rho| \leq 1 - \Delta \\ |\rho| - 1 + \Delta & \text{if } 1 - \Delta < |\rho| < 2 - \Delta - s \\ 1 - s & \text{if } |\rho| \geq 2 - \Delta - s \end{cases},$$

$$J_6^b(\rho) = \begin{cases} 2 + 2\Delta & \text{if } |\rho| \leq -s \\ -|\rho| + 2 + 2\Delta - s & \text{if } -s < |\rho| < 1 + \Delta \\ 1 + \Delta - s & \text{if } |\rho| \geq 1 + \Delta \end{cases},$$

$$L_6(y(x_i); F(x_i; W, \mathbf{b})) = J_6^w(f_{y(x_i)}(x_i)) + c \sum_{\substack{j=1 \\ j \neq y(x_i)}}^{k} J_6^b(f_j(x_i)), \tag{27}$$

where $\Delta \in [0, 1]$, $s \in (-1, 0]$ are the user defined constants.

By substituting these loss functions (27) into our general model, it is easy to get the optimization problems of RTWSVC, FRTWSVC and RampTWSVC. In theory, RTWSVC, FRTWSVC and RampTWSVC would terminate in a finite number of steps at the weak local optimal points if they are implemented by the general model strictly. The details are omitted.

IV. RFDPC

In this section, we introduce a new loss function fluctuated with the dataset, and then propose our robust fitting distribution planes for clustering (RFDPC) based on the general model.

Let us start from the efficient RampTWSVC [17]. Its ability to reduce the influence of the noise and outliers is manifested in Fig. 1. However, for the case of the samples from the same distribution, RampTWSVC may obtain very different cluster centers, leading bias from the data distribution. For instance, in Fig. 2, there are two groups of samples from $N(0, 1)$ (i.e., left and right three columns). RampTWSVC obtains two centers, depicted by solid blue lines in Fig. 2(b), are very different from each other.

To capture the data distribution, we introduce the 1-order and 2-order statistics [37] of the cluster into the within-cluster function and propose a new within-cluster function as

$$J_7^w(f(x; w_j, b_j)) = J_6^w(f(x; w_j, b_j)) + \frac{\gamma_1}{c_w} \bar{f}(x; w_j, b_j)^2 + \frac{\gamma_2}{c_w} \tilde{f}(x; w_j, b_j), \tag{28}$$

where γ_1, γ_2 are positive parameters. $\bar{f}(x; w_j, b_j) = \frac{1}{|N|} f(x; w_j, b_j)$ and $\tilde{f}(x; w_j, b_j) = \frac{1}{|N|-1}(f(x; w_j, b_j) - \frac{1}{|N|} \sum_{y(x_i) \in N} f(x_i; w_j, b_j))^2$, where N is the index set of the jth cluster that x belongs to and $|N|$ denotes the sample number of this cluster. In other words, $\bar{f}(x; w_j, b_j)$ and $\tilde{f}(x; w_j, b_j)$ are the corresponding parts in the mean and variance of the jth cluster with $j = 1, \ldots, k$. The additional statistics in (28) mean that a sample x assigned to a cluster would lead additional losses: (i) loss derived from the mean deviation, i.e., the deviation of the sample from the statistical center; (ii) loss derived from the variance of deviation, i.e., the deviation proportionality. Minimizing these statistics would make the cluster center close to the highest density region and the samples be uniformly distributed along with the cluster center. Fig. 2(c) shows the result by new function (28).

Then, by setting the between-cluster function $J_7^b(\rho) = J_6^b(\rho)$, the loss function of RFDPC becomes

$$L_7(y(x_i), F(x_i)) = c_w J_7^w(f(x_i; w_{y(x_i)}, b_{y(x_i)})) + c_b \sum_{\substack{j=1 \\ j \neq y(x_i)}}^{k} J_7^b(f(x_i; w_j, b_j)), \tag{29}$$

where $f(x; w_j, b_j) = w_j^\top x + b_j$.

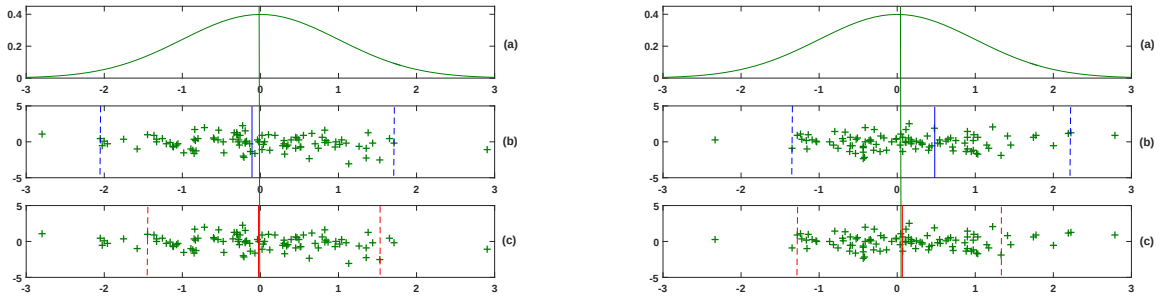


Fig. 2. Illustration of RampTWSVC and RFDPC on two groups of 100 samples from the same data distribution. The vertical green line is the mean of data samples. (a) is the image of normal distribution $N(0,1)$, (b) is 100 samples from the distribution (where the solid blue lines are the centers by $J_6^w(\rho)$ used in RampTWSVC), and (c) is 100 samples the same as (b) (where the solid red lines are the centers by $J_7^w(\rho)$ used in our RFDPC).

By introducing a L_2 regularization term, the subproblem in step 2(a) is considered as

$$\begin{aligned}\min_{w_j,b_j} \quad & c_w \sum_{\substack{i=1 \\ y^{(t)}(x_i)=j}}^{m} J_7^w(f(x_i;w_j,b_j))+ \\ & c_b \sum_{\substack{i=1 \\ y^{(t)}(x_i)\neq j}}^{m} J_7^b(f(x_i;w_j,b_j)) + \tfrac{1}{2}(||w_j||^2+b_j^2),\end{aligned} \tag{30}$$

and its local solution can be obtained by the concave-convex procedure (CCCP) [38].

It should be pointed out that the cluster assignment (13) can be replaced by the simplified assignment (14), though the function $J_7^w(\rho)$ does not satisfy properties (i)-(iii).

Theorem IV.1. *In RFDPC, the sample assignment* (13) *can be simplified as* (14).

Proof. Suppose l^* is the label of an arbitrary sample x obtained by (14), and l is an arbitrary label of x. From the proof of Theorem 2.1, we just need to prove $J_7^w(f(x;w_{l^*},b_{l^*})) \leq J_7^w(f(x;w_l,b_l))$ and $J_7^b(f(x;w_l,b_l)) \leq J_7^b(f(x;w_{l^*},b_{l^*}))$.

Note that γ_1 and γ_2 are positive parameters. Since smaller $|f|$ leads smaller $\bar{f}^2$ and smaller $\tilde{f}$, and since $J_6^w(\rho)$ is non-decreasing in $[0,\infty)$, the inequality $J_7^w(f(x;w_{l^*},b_{l^*})) \leq J_7^w(f(x;w_l,b_l))$ holds. Noticing that $J_7^b(\rho)$ satisfies properties (i)-(iii), the inequality $J_7^b(f(x;w_l,b_l)) \leq J_7^b(f(x;w_{l^*},b_{l^*}))$ holds. Therefore, the conclusion is obtained. □

In addition, our RFDPC hires termination condition (iii), and thus it terminates in a finite number of steps at a weak local optimal point by Theorem 2.5.

V. Experimental results

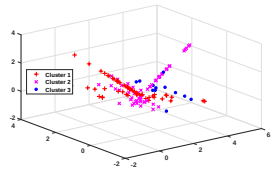
In this section, we analyze the performance of our RFDPC compared with some state-of-the-art partition-based clustering methods on several artificial and benchmark datasets. All the methods were implemented by MATLAB2017 on a PC with an Intel Core Duo Processor (double 4.2 GHz) with 16GB RAM. In the experiments, we used the metrics accuracy (AC) [12] and mutual information (MI) [39] to measure the performance of these methods.

On the synthetic data, we tested the ability of the plane-based clustering methods to capture the plane-based data distribution. The synthetic data in R^3 consists of three classes, where one class is on a plane and the other two classes are on two lines, respectively. The details of the synthetic data are shown in Table I. We sampled four groups from the synthetic data which include 120, 100, 80 and 60 samples, respectively. Then, the plane-based clustering methods, including kPC [8], PPC [10], [11], TWSVC [12], RTWSVC [16], FRTWSVC [16], RampTWSVC [17] and our RFDPC, were implemented on these four groups, where the parameters c, c_1 and c_2 were set to 0.1, Δ was set to 0.3, and s was set to -0.2. The clustering results were depicted in Fig. 3. It can be seen from Fig. 3 that (i) kPC and TWSVC cannot capture these plane-based clusters; (ii) PPC obtains a plane constructed by the two lines frequently; (iii) RTWSVC and FRTWSVC capture the three clusters on group G1, but both of them lose a cluster when the number of samples decreases; (iv) RampTWSVC always finds three clusters inaccurately; (iv) our RFDPC finds the three clusters exactly. Thus, our RFDPC captures the plane-based clusters more precisely than other methods on the synthetic datasets.

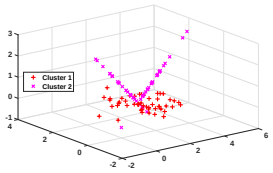
TABLE I
Details of the synthetic datasets

Dataset	Group			
	G1	G2	G3	G4
No. of samples	120	100	80	60
No. of dimensions	3	3	3	3
Distribution	**Class**			
	1	2	3	
Coordinate x	$\mathcal{N}$(1,1)	$\mathcal{N}$(3,1)	$\mathcal{N}$(2,1)	
Coordinate y	1	1	$\mathcal{N}$(1,1)	
Coordinate z	-x+1	x-1	0	

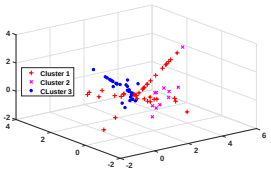
To exhibit the relationship between sample and its cluster center, the deviation statistics of the samples from their cluster center planes were depicted in Fig. 4, where '-' denotes the 1-order statistics $\bar{f}$ of each cluster and '×' denotes the 2-order statistics $\pm\tilde{f}$ of each cluster. A cluster that only has a '×' in Fig. 4 means its 1-order and 2-order statistics are out of the figure window. It is obvious that the 2-order statistics of deviation of kPC, PPC, and TWSVC are far from their 1-order statistics, and hence they cannot find the three plane-based clusters exactly. The cluster samples lie on their cluster



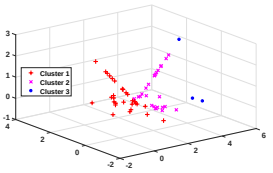
(i) kPC (Group G1)



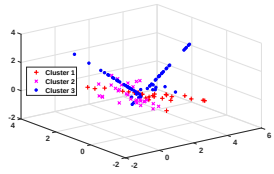
(ii) kPC (Group G2)



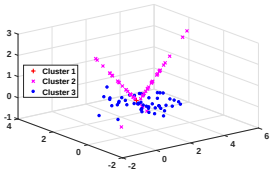
(iii) kPC (Group G3)



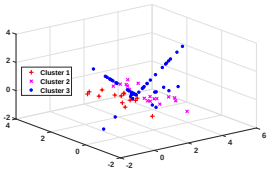
(iv) kPC (Group G4)



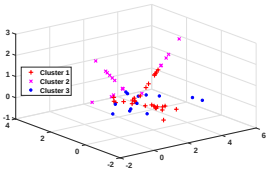
(v) PPC (Group G1)



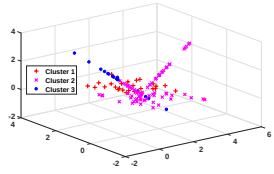
(vi) PPC (Group G2)



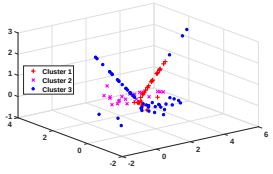
(vii) PPC (Group G3)



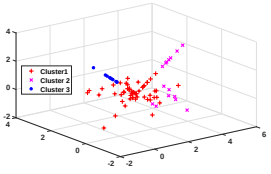
(viii) PPC (Group G4)



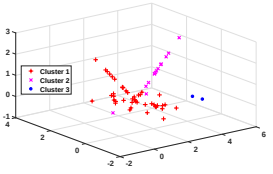
(ix) TWSVC (Group G1)



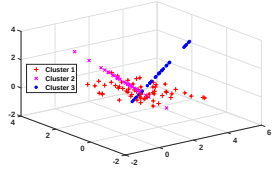
(x) TWSVC (Group G2)



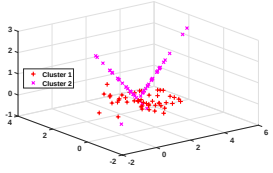
(xi) TWSVC (Group G3)



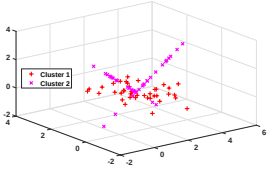
(xii) TWSVC (Group G4)



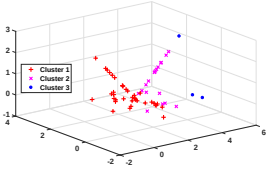
(xiii) RTWSVC (Group G1)



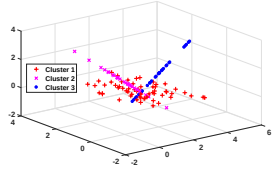
(xiv) RTWSVC (Group G2)



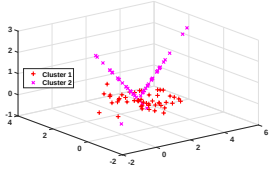
(xv) RTWSVC (Group G3)



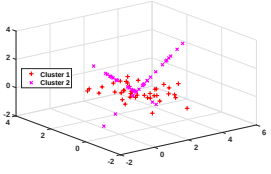
(xvi) RTWSVC (Group G4)



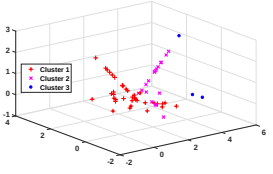
(xvii) FRTWSVC (Group G1)



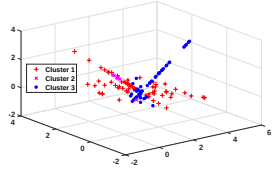
(xviii) FRTWSVC (Group G2)



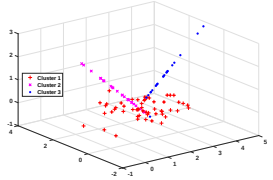
(xix) FRTWSVC (Group G3)



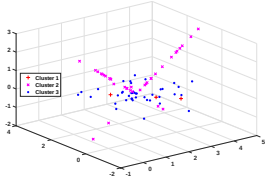
(xx) FRTWSVC (Group G4)



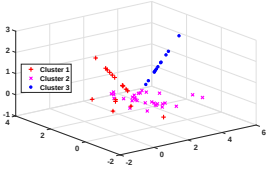
(xxi) RampTWSVC (Group G1)



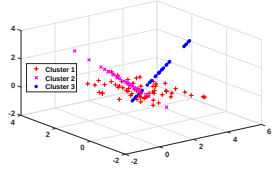
(xxii) RampTWSVC (Group G2)



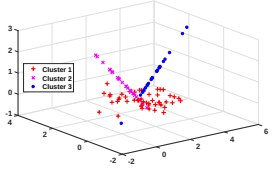
(xxiii) RampTWSVC (Group G3)



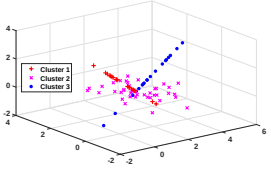
(xxiv) RampTWSVC (Group G4)



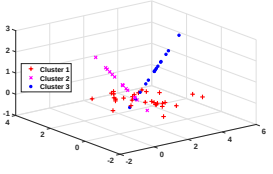
(xxv) RFDPC (Group G1)



(xxvi) RFDPC (Group G2)



(xxvii) RFDPC (Group G3)



(xxviii) RFDPC (Group G4)

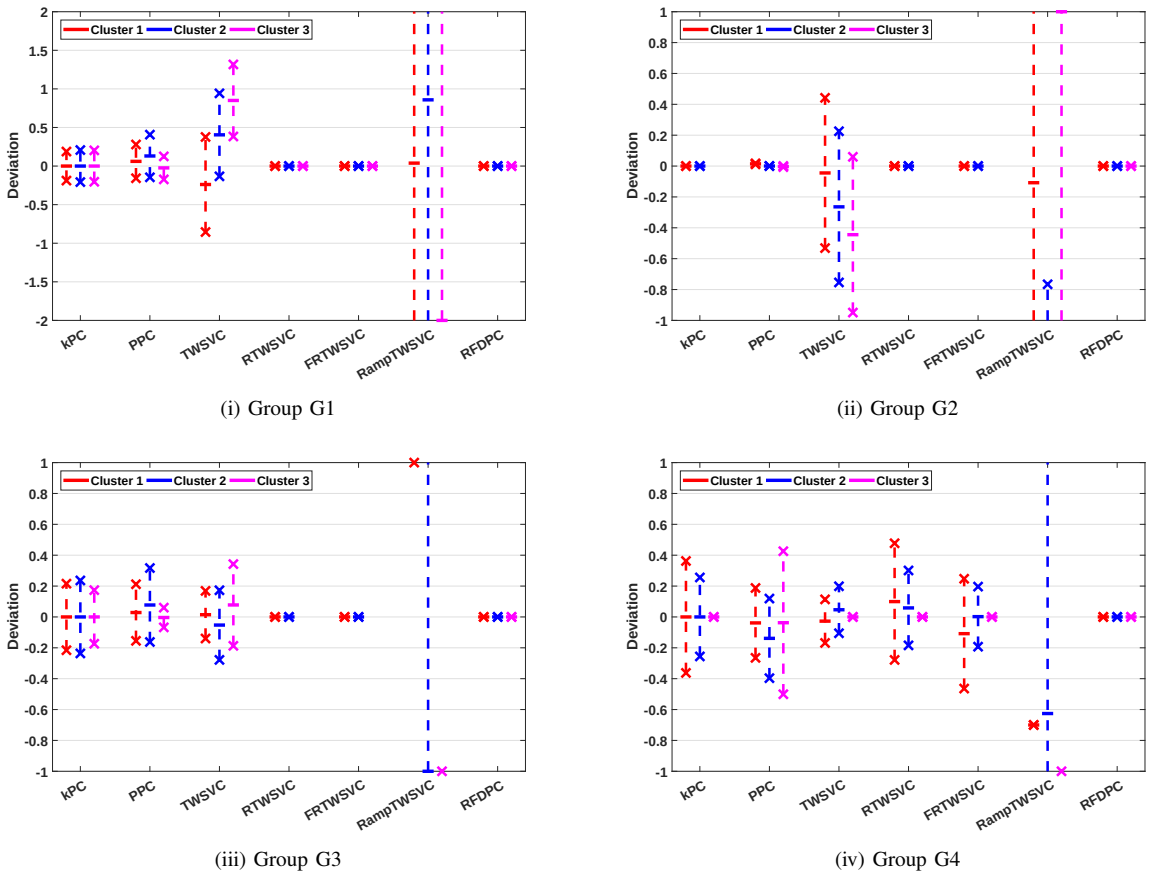
Fig. 3. Plane-based clustering methods applied to four groups of the synthetic datasets, where groups G1, G2, G3 and G4 include 120, 100, 80 and 60 samples, respectively.

TABLE II
AC AND MI OF THE PLANE-BASED CLUSTERING METHODS ON THE SYNTHETIC DATASETS

Group	kPC AC(%)/MI(%)	PPC AC(%)/MI(%)	TWSVC AC(%)/MI(%)	RTWSVC AC(%)/MI(%)	FRTWSVC AC(%)/MI(%)	RampTWSVC AC(%)/MI(%)	RFDPC AC(%)/MI(%)
G1	60.74/27.18	62.20/29.63	56.57/27.55	**100.0/100.0**	**100.0/100.0**	72.82/33.97	**100.0/100.0**
G2	87.37/67.10	87.39/67.34	62.79/28.42	87.37/67.10	87.37/67.10	84.38/56.71	**100.0/100.0**
G3	56.42/13.13	61.14/40.67	58.96/21.39	87.34/67.21	87.34/67.21	65.09/23.83	**98.13/94.35**
G4	58.87/24.09	58.81/11.19	61.53/30.08	57.06/16.79	57.18/21.07	73.45/48.10	**100.0/100.0**

TABLE III
AC AND MI OF THE CLUSTERING METHODS ON BENCHMARK DATASETS FOR LINEAR CASE

Data k:m×n	kmeans AC(%) MI(%)	kPC AC(%) MI(%)	PPC AC(%) MI(%)	TWSVC AC(%) MI(%)	RTWSVC AC(%) MI(%)	FRTWSVC AC(%) MI(%)	RampTWSVC AC(%) MI(%)	RFDPC AC(%) MI(%)
Compound	86.29±4.05	72.04	73.90	75.97	80.44	82.52	77.07	**89.01**
6:399×2	**70.44**±7.14	34.71	38.85	50.44	56.63	48.18	48.37	67.38
Dermatology	69.76±0.77	60.50	70.36	71.93	60.50	60.50	72.67	**93.47**
6:366×34	11.47±2.15	29.65	3.48	10.17	28.95	28.59	24.42	**76.48**
Ecoli	82.19±2.68	33.11	66.46	85.74	34.33	34.33	79.42	**91.03**
8:336×7	56.84±4.42	8.61	9.65	58.45	10.42	10.42	43.35	**70.37**
Glass	65.58±3.22	55.73	**66.75**	66.62	57.59	57.40	62.77	62.37
6:214×9	**35.76**±2.23	22.55	8.54	17.83	17.69	18.20	20.95	12.61
Iris	84.57±6.86	67.54	60.95	91.24	92.67	94.95	86.79	**98.25**
3:150×4	70.47±9.10	25.41	12.04	82.53	82.31	86.97	71.71	**94.86**
Pathbased	74.85±0.09	66.49	74.57	73.94	76.30	76.30	65.73	**79.14**
3:300×2	51.46±0.16	30.17	50.92	47.90	54.63	54.63	28.21	**59.21**
Zoo	87.49±1.96	54.12	84.06	88.83	54.12	54.12	90.22	**95.47**
7:101×16	71.93±3.15	34.23	55.56	72.93	32.15	32.15	**76.98**	71.79
Aggregation	91.91±0.69	79.19	79.00	88.49	82.82	84.10	80.71	**95.97**
7:788×2	81.24±0.75	48.84	48.43	63.52	64.23	60.70	52.36	**84.82**
R15	**98.21**±0.67	92.15	92.00	93.76	93.07	92.91	81.76	96.92
15:600×2	**93.35**±2.40	64.86	59.28	73.64	73.49	67.33	47.37	86.80
Vehicle	63.24±2.21	62.03	62.77	51.00	65.23	65.00	58.59	**68.42**
4:846×18	17.73±0.82	3.25	1.28	9.02	12.63	12.07	14.84	**22.19**
Vowel	**86.62**±0.61	82.93	84.10	83.28	84.23	83.60	80.94	84.18
11:528×10	**45.64**±1.94	11.39	10.60	11.57	24.28	7.73	25.93	34.20
Echocardiogram	66.41±7.92	52.81	56.66	56.10	75.01	75.01	71.84	**85.79**
2:131×10	24.79±17.27	0.54	2.99	1.35	39.64	39.64	35.46	**58.50**
Haberman	49.91±0.02	49.84	60.95	61.89	61.57	62.21	60.95	**64.96**
2:306×3	0.04±0.04	0.07	0.74	2.28	**9.00**	4.44	0.74	8.70
Heartc	51.04±0.00	50.12	50.23	50.67	59.21	59.21	50.75	**66.57**
2:303×14	1.39±0.00	0.05	0.14	0.90	13.98	13.98	1.14	**25.41**
Heartstatlog	51.45±0.07	50.04	50.35	50.81	51.40	51.40	51.82	**59.40**
2:270×13	1.87±0.07	0.20	0.15	0.63	1.63	1.67	2.40	**17.08**
Hepatitis	62.77±3.03	55.56	71.90	66.27	67.02	**73.66**	67.02	69.38
2:155×19	0.29±0.13	0.96	14.93	0.17	7.18	**17.36**	1.95	6.09
Hourse	50.15±0.00	51.34	**54.15**	50.15	51.34	51.34	52.12	51.98
2:300×26	1.24±0.00	0.55	**5.39**	1.24	0.55	0.55	0.46	0.25
Housevotes	78.83±0.15	63.77	68.77	75.83	71.40	71.40	79.61	**83.64**
2:435×16	48.07±0.38	34.16	27.27	44.66	39.36	39.36	50.15	**56.38**
Sonar	50.22±0.18	49.80	49.99	50.43	51.26	50.06	**51.62**	**51.62**
2:208×60	0.74±0.28	0.01	0.23	0.64	2.06	0.67	**4.05**	2.72
Spect	52.97±0.00	65.86	50.67	65.86	50.88	50.58	67.17	**67.61**
2:267×44	**8.48**±0.00	0.51	0.51	0.51	0.35	0.34	1.15	1.17
Spectf	53.95±2.31	49.49	50.51	51.93	41.93	51.39	53.20	**59.62**
2:88×44	**16.09**±6.43	0.19	1.67	6.34	4.00	3.03	6.51	15.15
Pimaindian	55.07±0.00	51.74	54.50	57.99	53.97	54.82	55.07	**60.33**
2:768×8	2.67±0.00	0.23	0.09	5.64	0.21	0.58	0.95	**9.51**
Tictactoe	50.90±0.47	50.08	96.71	55.26	59.84	59.84	55.82	**96.91**
2:958×27	0.69±0.14	0.69	87.89	0.97	13.48	13.48	1.95	**88.49**
AC-win	2	0	2	0	0	1	1	18
MI-win	6	0	1	0	1	1	2	12
Both-win	2	0	1	0	0	1	1	12



(i) Group G1

(ii) Group G2

(iii) Group G3

(iv) Group G4

Fig. 4. Deviation statistics of the samples from their cluster center planes on the synthetic datasets, where the clustering methods include kPC, PPC, TWSVC, RTWSVC, FRTWSVC, RampTWSVC and RFDPC, '-' denotes the 1-order statistics and '$\times$' denotes the 2-order statistics. If the symbol '$\times$' for a method is on the lower bound or upper bound, it means the deviations of this cluster are out of the region. If there is not any symbol for a cluster, it means there is not such a cluster for this method.

center planes by RTWSVC and FRTWSVC on groups G1, G2 and G3, but they fail to find the 3rd cluster on groups G2 and G3. RampTWSVC has great fluctuation, and thus it cannot capture the three plane-based clusters exactly. Accordingly, our RFDPC captures the plane-based clusters well by adding additional statistics. The quantitative measurements were reported in Table II, and the highest ones were bold. Apparently, our RFDPC owns the highest performance on the four groups than other plane-based clustering methods, which is consistent with the previous observations.

In the following experiment, we implemented the above methods and kmeans [32] on several benchmark datasets [40] for linear and nonlinear cases. Typically, Δ was set to 0.3, s was set to -0.2, and $\gamma_1 = \gamma_2$. Other parameters in these methods were selected from $\{2^i | i = -8, -7, \ldots, 7\}$. For nonlinear case, Gaussian kernel [10], [41], [42] $K(x_1, x_2) = \exp\{-\mu||x_1 - x_2||^2\}$ was used and its parameter μ was selected from $\{2^i | i = -10, -9, \ldots, 5\}$. The random initialization was fused for kmeans, and the NNG initialization [12] was fused for the rest plane-based clustering methods to obtain stable performances. We reported AC and MI of these methods in Tables III and IV for linear and nonlinear cases, respectively. Thereinto, kmeans was implemented ten times, and then the mean value and standard deviation were computed and reported. The highest ACs or MIs are bold, and the numbers of datasets with highest AC, MI and both are also shown in these tables. From Table III, it can be seen that our RFDPC outperforms other methods on most of the datasets. Our RFDPC has the highest AC on 18/23 datasets, the highest MI on 12/23 datasets, and both of them on 12/23 datasets. Moreover, our RFDPC is comparable with the methods that own the highest AC or MI on most of the rest datasets. Table IV has similar results to that of Table III and confirms the observation from Table III. To exhibit the cluster center planes obtained by these plane-based clustering methods, we depicted the deviation statistics on the datasets "Haberman", "Iris", "Pathbased" and "Vehicle" in Fig. 5 as instances. Obviously, our RFDPC has a small and tight 2-order deviation statistics around the 1-order statistics, which improves the performance of plane-based clustering significantly.

VI. Conclusions

A general model for plane-based clustering has been proposed by introducing loss function and regularization. It has been shown that the general model terminates in a finite number of steps at the local or weak local optimal points theoretically. The existing plane-based clustering methods, including kPC, PPC, TWSVC, RTWSVC, FRTWSVC and

TABLE IV
AC AND MI OF THE CLUSTERING METHODS ON BENCHMARK DATASETS FOR NONLINEAR CASE

Data k:m×n	kmeans AC(%) MI(%)	kPC AC(%) MI(%)	PPC AC(%) MI(%)	TWSVC AC(%) MI(%)	RTWSVC AC(%) MI(%)	FRTWSVC AC(%) MI(%)	RampTWSVC AC(%) MI(%)	RFDPC AC(%) MI(%)
Compound	84.84±4.11	90.16	70.32	90.25	90.16	90.16	89.73	**91.65**
6:399×2	70.65±6.84	**72.24**	16.97	71.60	72.24	72.24	64.07	61.57
Dermatology	71.66±1.26	72.66	70.62	72.60	72.60	72.60	72.90	**74.44**
6:366×34	17.84±3.67	18.00	3.65	18.00	18.00	18.00	**26.79**	20.16
Ecoli	79.93±1.24	82.49	69.13	88.29	82.49	82.68	83.01	**90.34**
8:336×7	49.31±2.28	57.79	16.46	**62.21**	57.79	57.57	49.97	50.97
Glass	69.27±1.45	69.04	66.82	70.10	69.04	69.04	70.77	**71.41**
6:214×9	37.50±2.09	**41.42**	7.35	23.42	41.42	41.42	29.18	30.81
Iris	87.63±8.09	91.24	59.47	91.24	91.24	91.24	94.95	**97.40**
3:150×4	76.26±9.85	79.15	13.93	79.15	79.15	79.15	**86.23**	85.59
Pathbased	**96.11**±0.18	76.29	59.94	80.57	76.29	76.29	91.92	93.92
3:300×2	**88.23**±0.40	57.51	11.60	65.87	57.51	57.51	79.83	82.28
Zoo	87.14±3.39	90.63	89.52	90.63	90.63	90.63	**91.25**	91.15
7:101×16	70.79±5.39	77.99	72.90	77.99	77.99	77.99	**79.70**	70.27
Echocardiogram	71.14±0.82	55.04	56.66	56.66	55.04	55.04	71.84	**83.23**
2:131×10	32.41±0.53	0.85	2.73	2.73	0.85	0.85	28.53	**51.66**
Haberman	60.61±0.30	63.21	61.26	**63.55**	63.21	63.21	63.21	**63.55**
2:306×3	0.23±0.13	4.97	0.75	**5.55**	4.97	4.97	4.97	5.41
Heartc	50.76±0.09	51.37	51.26	50.50	51.37	51.37	52.44	**53.24**
2:303×14	1.74±0.31	2.19	1.68	0.62	2.19	2.19	3.52	**5.31**
Heartstatlog	50.83±0.41	53.00	51.54	50.92	53.00	53.00	**54.91**	54.22
2:270×13	1.88±0.54	3.79	1.64	0.81	3.79	3.79	**6.98**	6.25
Hepatitis	65.35±1.58	66.27	67.79	67.79	66.27	66.27	67.02	**69.38**
2:155×19	1.04±0.68	0.29	2.01	2.01	0.29	0.29	0.29	**5.35**
Hourse	52.27±0.25	52.12	53.05	51.71	52.12	52.12	52.12	**54.55**
2:300×26	0.68±0.57	0.46	2.30	0.50	0.46	0.46	0.46	**3.92**
Housevotes	79.79±0.94	75.50	75.83	**91.21**	75.50	75.50	80.68	79.96
2:435×16	46.91±1.87	42.09	46.38	**72.31**	42.09	42.09	48.86	48.74
Sonar	50.16±0.28	51.62	52.66	52.22	51.62	51.62	**54.52**	**54.52**
2:208×60	0.39±0.39	4.24	4.08	5.43	4.24	4.24	6.64	**6.77**
Spect	60.68±4.79	66.73	68.06	68.06	66.73	66.73	68.98	**71.87**
2:267×44	3.38±3.72	0.17	2.35	2.35	0.17	0.17	**17.69**	10.96
Spectf	63.87±0.94	50.16	**74.18**	50.16	50.16	50.16	62.03	70.76
2:88×44	21.84±1.52	3.88	**49.34**	3.88	3.88	3.88	20.54	34.36
AC-win	1	0	1	2	0	0	3	12
MI-win	1	2	1	3	0	0	5	5
Both-win	1	0	1	1	0	0	2	5

RampTWSVC, are consistent with this general model. Furthermore, a new plane-based clustering method (RFDPC) based on the general model has been proposed. Experimental results on the synthetic and public available datasets have indicated that our RFDPC can capture the data distribution more precisely. For practical convenience, the corresponding RFDPC Matlab code has been uploaded upon http://www.optimal-group.org/Resources/Code/RFDPC.html. In the future work, it is interesting to find more efficient loss functions and generalization terms in the general model to suit for specific clustering purpose.

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] M. Aldenderfer and R. Blashfield, *Cluster Analysis*. Los Angeles: Sage Publications, 1985.

[3] X. Pei, C. Chen, and W. Gong, "Concept factorization with adaptive neighbors for document clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 343–352, 2018.

[4] I. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," *The tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 551–556, 1988.

[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[6] X. Huang, Y. Ye, and H. Zhang, "Extensions of kmeans-type algorithms: a new clustering framework by integrating intracluster compactness and intercluster separation." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1433–1446, 2014.

[7] P. Bradley, O. Mangasarian, and W. Street, "Clustering via concave minimization," *Advances in Neural Information Processing Systems*, vol. 9, pp. 368–374, 1997.

[8] P. Bradley and O. Mangasarian, "k-plane clustering," *Journal of Global Optimization*, vol. 16, no. 1, pp. 23–32, 2000.

[9] P. Tseng, "Nearest q-flat to m points," *Journal of Optimization Theory and Applications*, vol. 105, no. 1, pp. 249–252, 2000.

[10] L. Liu, Y. Guo, Z. Wang, Z. Yang, and Y. Shao, "k-proximal plane clustering," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 5, pp. 1537–1554, 2017.

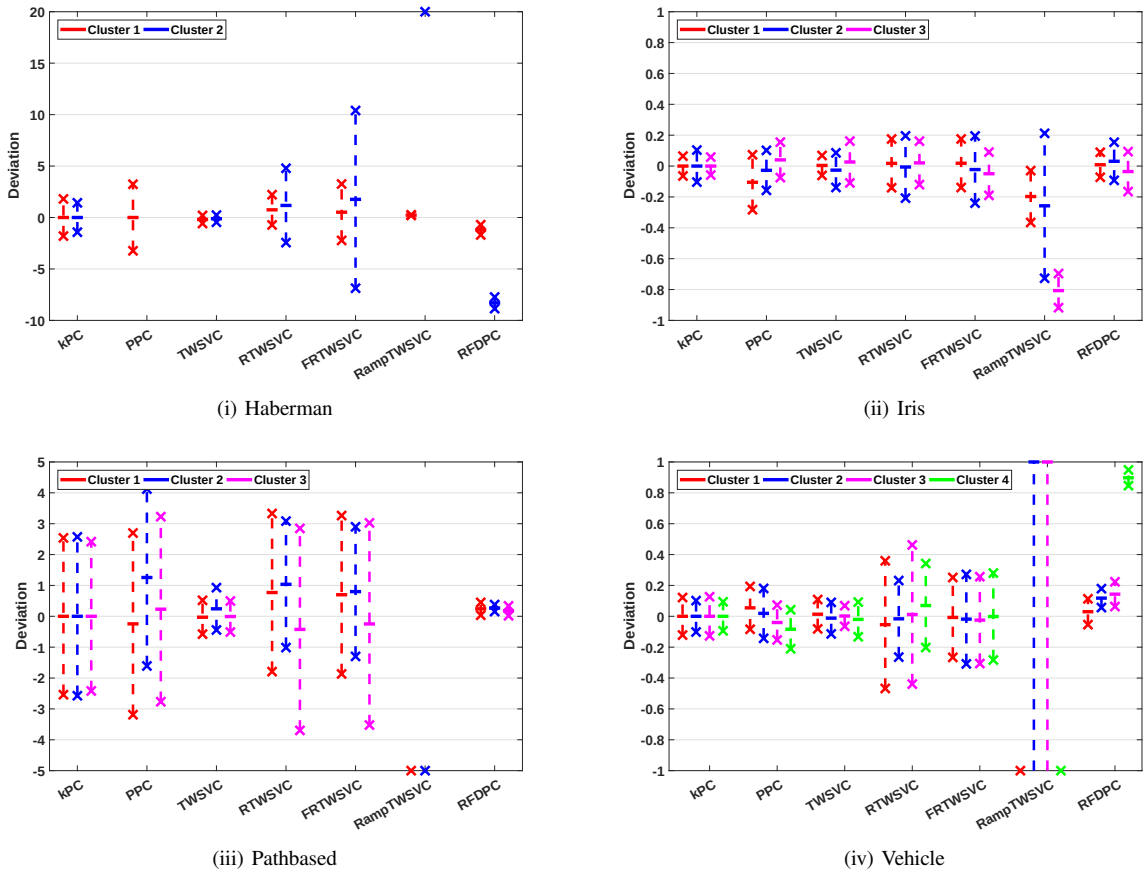
[11] Y. Shao, L. Bai, Z. Wang, X. Hua, and N. Deng, "Proximal plane clustering via eigenvalues," *Procedia Computer Science*, vol. 17, pp. 41–47, 2013.

[12] Z. Wang, Y. Shao, L. Bai, and N. Deng, "Twin support vector machine for clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2583–2588, 2015.

[13] Z. Yang, H. Wu, C. Li, and Y. Shao, "Least squares recursive projection twin support vector machine for multi-class classification," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 3, pp. 411–426, 2016.

[14] R. Khemchandani, A. Pal, and S. Chandra, "Fuzzy least squares twin support vector clustering," *Neural Computing and Applications*, vol. 29, no. 2, pp. 553–563, 2018.

[15] Z. Yang, Y. Guo, C. Li, and Y. Shao, "Local k-proximal plane cluster-



(i) Haberman (ii) Iris

(iii) Pathbased (iv) Vehicle

Fig. 5. Deviation statistics of the samples from their cluster center planes on four benchmark datasets, where the symbolic descriptions can be found in Fig. 4.

ing," *Neural Computing and Applications*, vol. 26, no. 1, pp. 199–211, 2015.

[16] Q. Ye, H. Zhao, Z. Li, X. Yang, S. Gao, T. Yin, and N. Ye, "L1-norm distance minimization-based fast robust twin support vector k-plane clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4494–4503, 2018.

[17] Z. Wang, X. Chen, C. Li, and Y. Shao, "Ramp-based twin support vector clustering," *arXiv preprint, arXiv:1812.03710*, 2018.

[18] O. Mangasarian and E. Wild, "Multisurface proximal support vector classification via generalize eigenvalues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 69–74, 2006.

[19] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[20] Y. Shao, C. Zhang, X. Wang, and N. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962 – 968, 2011.

[21] X. Peng, D. Xu, L. Kong, and D. Chen, "L1-norm loss based twin support vector machine for data recognition," *Information Sciences*, vol. 340, pp. 86–103, 2016.

[22] Y. Q. . Y. N. Gao, S., "1-norm least squares twin support vector machines," *Neurocomputing*, vol. 74, no. 17, pp. 3590–3597, 2011.

[23] H. Cevikalp, "Best fitting hyperplanes for classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1076–1088, 2017.

[24] V. Vapnik, *Statistical Learning Theory*, H. Simon, Ed. Wiley-Interscience, New York, USA, 1998.

[25] J. Zhu, S. Rosset, R. Tibshirani, and T. Hastie, "1-norm support vector machines," *In Advances in Neural Information Processing Systems*, pp. 49–56, 2004.

[26] X. Huang, L. Shi, and J. A. Suykens, "Support vector machine classifier with pinball loss," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 984–997, 2014.

[27] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process Letter*, vol. 9, no. 3, pp. 293–300, 1999.

[28] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song, "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[29] X. Sun, Y. Chen, Y. Shao, C. Li, and C. Wang, "Robust nonparallel proximal support vector machine with lp-norm regularization," *IEEE Access*, vol. 6, pp. 20 334–20 347, 2018.

[30] X. Li, Q. Lu, Y. Dong, and D. Tao, "Robust subspace clustering by cauchy loss function," *IEEE Transactions on Neural Networks and Learning Systems*, vol. doi:10.1109/TNNLS.2018.2876327, 2018.

[31] C. Hou, F. Nie, D. Yi, and D. Tao, "Discriminative embedded clustering: A framework for grouping high-dimensional data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1287–1299, 2015.

[32] A. Jain and R. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[33] S. Elavarasi, J. Akilandeswari, and B. Sathiyabhama, "A survey on partition clustering algorithms," *International Journal of Enterprise Computing and Business Systems*, vol. 1, p. 1, 2011.

[34] G. Patel, V. Dabhi, and H. Prajapati, "Study and analysis of particle swarm optimization for improving partition clustering," *In Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pp. 218–225, 2015.

[35] R. Khemchandani, Jayadeva, and S. Chandra, "Optimal kernel selection in twin support vector machines," *Optimization Letters*, vol. 3, pp. 77–88, 2009.

[36] Z. Wang, Y. Shao, L. Bai, C. Li, L. Liu, and N. Deng, "Insensitive stochastic gradient twin support vector machines for large scale problems," *Information Sciences*, vol. 462, pp. 114–131, 2018.

[37] R. Johnson and P. Kuby, *Elementary statistics*. Cengage Learning, 2011.

[38] A. Yuille and A. Rangarajan, "The concave-convex procedure (cccp),"

Advances in Neural Information Processing Systems, vol. 2, pp. 1033–1040, 2002.

[39] A. Kraskov, H. Stogbauer, R. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, p. 278, 2005.

[40] C. Blake and C. Merz, *UCI Repository for Machine Learning Databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[41] B. Schölkopf and A. Smola, *Learning with kernels*. Cambridge: MA:MIT Press, 2002.

[42] B. Schölkopf and J. Alexander, *Advances in Kernel Methods-Support Vector Learning*. MA:MIT Press, 1998.