# A Novel Separating Hyperplane Classification Framework to Unify Nearest-class-model Methods for High-dimensional Data

Rui Zhu, Ziyu Wang, Naoya Sogi, Kazuhiro Fukui, *Member, IEEE,* and Jing-Hao Xue

***Abstract*—In this paper, we establish a novel separating hyperplane classification (SHC) framework to unify three nearest-class-model methods for high-dimensional data: the nearest subspace method (NSM), the nearest convex hull method (NCHM) and the nearest convex cone method (NCCM). Nearest-class-model methods are an important paradigm for classification of high-dimensional data. We first introduce the three nearest-class-model methods and then conduct dual analysis for theoretically investigating them, to understand deeply their underlying classification mechanisms. A new theorem for the dual analysis of NCCM is proposed in this paper, through discovering the relationship between a convex cone and its polar cone. We then establish the new SHC framework to unify the nearest-class-model methods based on the theoretical results. One important application of this new SHC framework is to help explain empirical classification results: why one class model has better performance than others on certain datasets. Finally, we propose a new nearest-class-model method, the soft NCCM, under the novel SHC framework to solve the overlapping class model problem. For illustrative purposes, we empirically demonstrate the significance of our SHC framework and the soft NCCM through two types of typical real-world high-dimensional data, the spectroscopic data and the face image data.**

***Index Terms*—Classification, convex cone, convex hull, dual analysis, separating hyperplane, subspace.**

## I. Introduction

A Category of popular generative classifiers to classify high-dimensional data is the nearest-class-model methods, also known as the class modelling methods in the chemometrics community or the subspace methods in the machine learning and pattern recognition communities. In the nearest-class-model methods, we construct a class model for each class from the training samples of that class, independently of other classes; a test sample is assigned to the class with the highest similarity between the sample and the class model.

Three class models have been studied in the literature, the principal component (PC) subspace, the convex hull model and the convex cone model. PC subspace is a widely-used class subspace. The PC subspace of a class is built through principal component analysis (PCA) of the training samples of that class, such that a class is represented by a low-dimensional linear subspace spanned by a small number of learning PCs which present the most variable information in the class. Hence the PC subspace has been widely used as a class representation for high-dimensional data. Soft independent modelling of class analogy (SIMCA) [1]–[6] in chemometrics and the mutual subspace method (MSM) [7]–[10] and the nearest subspace classifier (NSC) [11]–[14] in pattern recognition are famous examples of PC-subspace-based classifiers. In SIMCA and NSC, the dissimilarity measure is related to the Euclidean distance between a test sample and a PC subspace; in MSM, it is the canonical angle between them. It is, however, not necessary to use subspaces to represent classes. The geometric convex model representation is another popular class representation approach for classification tasks. The geometric convex model for a class is constructed by a linear combination of class samples, with certain constraints on the linear combination coefficients.

The convex hull model [15]–[19] is one geometric model that attracts a lot of attention recently. Nalbantov et al. [15] propose the nearest convex hull classification, which uses a convex hull model to represent a class and classifies a test sample to the class with the nearest convex hull. The convex hull model of a class is constructed by the convex combination, i.e. the linear combination with nonnegative and sum-to-one constraints on the coefficients, of the training samples of that class. The dissimilarity measure is the Euclidean distance from a test sample to a convex hull [15].

The convex cone model has also been used as class representation for face recognition [14], [20]. A convex cone model is constructed by the conic combinations of the class samples, i.e. the linear combinations with nonnegative coefficients. Kobayashi et al. [20] propose the cone-restricted subspace method, using the angle between a test sample and a convex cone for classification.

Among these three types of models, the PC subspace is a set of vectors that are linear combinations of the PCs with no constraints on the coefficients. Thus the PC subspace covers an infinite area that has weak constraints on the location of a class within its class subspace, which is considered as a loose representation of the class. In contrast, the two geometric convex models provide a restricted area to represent the class by setting constraints on the linear combination coefficients.

R. Zhu is with the Faculty of Actuarial Science and Insurance, City, University of London, London EC1Y 8TZ, UK. E-mail: rui.zhu@city.ac.uk

Z. Wang is with the Department of Security and Crime Science and the Department of Statistical Science, University College London, London WC1E 6BT, UK. E-mail: ziyu.wang.12@ucl.ac.uk

N. Sogi and K. Fukui are with the Department of Computer Science, University of Tsukuba, Tsukuba, Japan. Email: sogi@cvlab.cs.tsukuba.ac.jp; kfukui@cs.tsukuba.ac.jp

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK. E-mail: jinghao.xue@ucl.ac.uk

This work was partially supported by University College London's Security Science Doctoral Training Centre under Engineering and Physical Sciences Research Council (EPSRC) grant EP/G037264/1.

The restricted area is bounded by the class samples that are used to construct the convex models. The convex hull model adopts the convex constraints on the linear combination coefficients. However, the convex constraint is often too tight in the sense that the classes often extend well beyond the convex hulls [17]. Considering the tightness of a model, a convex cone model lies in between a linear subspace model and a convex hull model. A convex cone is more restricted than a linear subspace because of the nonnegative constraints on the coefficients, while is looser than a convex hull because the conic combination constraint is looser than the convex combination constraint.

The geometric convex models have shown superior classification performances to the PC subspace [15], [20]. However, theoretically why and when this will happen is barely studied in literature. *Therefore, in this paper, we aim to theoretically investigate and unify three nearest-class-model classification methods which respectively use the PC subspace, the convex hull and the convex cone. Under the unified framework, we are able to explain why for certain datasets one class model is superior to the others in terms of empirical classification performance. In addition, we aim to develop new nearest-class-model methods under this framework to better classify data with specific properties, e.g. with overlapping class models.* To make the theoretical investigation more straightforward, we use the distance as the dissimilarity measure. In this fashion, the PC subspace representation leads to a nearest subspace method (NSM) [11]; the convex hull model leads to a nearest convex hull method (NCHM) [15]; and the convex cone model leads to a nearest convex cone method (NCCM), which is similar to the method in [20].

We first establish the novel separating hyperplane classification (SHC) framework to unify the nearest-class-model methods through the separating hyperplanes as a common platform. To achieve this, we shall investigate the corresponding hyperplane-based classifiers to NSM, NCHM and NCCM, through the *dual analysis of their minimum distance problems*. We first introduce the dual analysis for NSM and NCHM in literature and then show a new theoretical result of the dual analysis for NCCM through discovering the relationship between a convex cone and its polar cone. This relationship is analogous to that between a subspace and its orthogonal complement. We shall show that the minimum distance from a test sample to a class model is equivalent to the maximum distance from that sample to a hyperplane. Thus for each class model, we can find one separating hyperplane that separates the test sample from the class models. The test sample is then classified to the class with its nearest hyperplane. Therefore different class models are unified by the separating hyperplanes which can be simply described by their normal vectors and biases. However, we note that formulating a nearest class problem using hyperplanes does not bring advantages in computation [21].

Based on the SHC framework, we can then explain empirical classification results by investigating the discriminative abilities of the normal vectors associated with the separating hyperplanes. We shall show that the normal vectors of the separating hyperplanes are of great importance to classification: the more discriminative the normal vectors are, the better the classification.

It is worth noting that our SHC framework is different from the extensions of support vector machine (SVM) based on a pair of separating hyperplanes in one-sided or two-sided best fitting hyperplane classifiers [22], generalised eigenvalue proximal SVM [23] or twin SVM [24]. In [22]–[24], the pair of separating hyperplanes are found for the pair of class models and fixed for all the test samples, making the classification boundary linear for linear kernels. In contrast, the pair of separating hyperplanes in our SHC framework vary with test samples, making the classification boundary nonlinear.

We then propose a new classifier, the soft NCCM, under the SHC framework by imposing proper constraints to solve the overlapping class model problem. In real applications, it is possible to have overlapping class models and the class memberships of the test samples locating in the overlapping area are ambiguous. The new soft NCCM utilises the discriminative between-class information when constructing the class cones and can eliminate the overlapping area between the cones. The test instances locating in the overlapping area can then be better classified unambiguously.

For illustrative purposes, we apply NSM, NCHM, NCCM and soft NCCM to two types of typical high-dimensional data, the spectroscopic data and the face image data. We shall show the effectiveness of the new SHC framework in explaining the empirical classification results on these real data. We shall also show the superior classification performance of the new soft NCCM classifier over other methods to classify these data.

The contributions of our work are threefold.

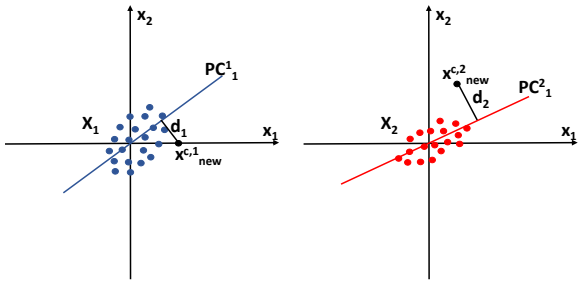
1) We develop new theoretical results of the dual analysis of NCCM, by discovering the relationship between a convex cone and its polar cone.
2) We establish a novel separating hyperplane classification (SHC) framework to unify and easily compare the nearest-class-model methods. Empirically, the new SHC framework can help explain why a class model is superior for certain datasets; and methodologically, it can help to design more sophisticated nearest-class-model methods with better classification performance.
3) We propose a new nearest-class-model method, the soft NCCM, under the SHC framework to solve the overlapping class model problem.

## II. Nearest-class-model methods

In this section we introduce the three nearest-subspace-methods, NSM, NCHM and NCCM, with illustrative examples in a two-dimensional feature space.

### A. PC subspace model: NSM

The nearest subspace method (NSM) models each class by a principal component (PC) subspace which can be obtained by applying the singular value decomposition on the centred training set of one class. A test instance is then classified to the nearest class by comparing its Euclidean distances to the two class subspaces. Fig. 1 shows an illustrative example of NSM in a two-dimensional feature space. The blue and red




(a) The PC subspace for class 1. (b) The PC subspace for class 2.

Fig. 1: An illustrative example of NSM in a 2D space.

straight lines are the PC class subspaces of the two classes, respectively, which are constructed by the first PCs. The Euclidean distances from $\boldsymbol{x}_{new}$ to the two class subspaces are shown as $d_1$ and $d_2$, respectively. In this example, we assign $\boldsymbol{x}_{new}$ to class 1 since $d_1 < d_2$. Note that we use two plots to represent the PC subspaces of the two classes, respectively, in order to achieve better visualisation. The technical details of NSM are described as follows.

**Definition II.1.** *Subspace.* Suppose $S = \{\boldsymbol{x}_i\}_{i=1}^N$ is a subset of $\mathbb{R}^p$. The set $\mathcal{L}(S) = \{\boldsymbol{v} : \boldsymbol{v} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \mid \boldsymbol{x}_i \in S, \alpha_i \in \mathbb{R}\}$, called the subspace generated by $S$, consists of all vectors in $\mathbb{R}^p$ which are linear combinations of vectors in $S$. We also say that the vectors in $S$ span the subspace $\mathcal{L}(S)$.

In the training phase, the nearest subspace method (NSM) builds class subspaces for the classes separately using PCA. We denote $\boldsymbol{X}_k \in \mathbb{R}^{N_k \times p}$ as the training set of class $k$ ($k = 1, 2$ for two-class classification), where $N_k$ is the number of training samples and each row of $\boldsymbol{X}_k$ represents a $p$-dimensional training sample. The PC subspace for the $k$th class can be obtained from applying the reduced singular value decomposition to the column-centred $\boldsymbol{X}_k$: $\boldsymbol{X}_k^c = \boldsymbol{U}_k \boldsymbol{\Lambda}_k \boldsymbol{V}_k^T$, where the rows of $\boldsymbol{U}_k \in \mathbb{R}^{N_k \times q_k}$ denote the normalised PC scores; the columns of $\boldsymbol{V}_k \in \mathbb{R}^{p \times q_k}$ denote the PCs; and $\boldsymbol{\Lambda}_k$ is a diagonal matrix of singular values $\{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{q_k}\}$. The $r_k$-dimensional ($r_k \leq q_k$) PC subspace $\mathcal{L}(\boldsymbol{W}_k)$ is spanned by the first $r_k$ PCs $\boldsymbol{W}_k \in \mathbb{R}^{p \times r_k}$.

In the test phase, a new sample $\boldsymbol{x}_{new} \in \mathbb{R}^{1 \times p}$ is assigned according to the distance from $\boldsymbol{x}_{new}^{c,k}$ to the class subspace $\mathcal{L}(\boldsymbol{W}_k)$, where $\boldsymbol{x}_{new}^{c,k}$ is the centred $\boldsymbol{x}_{new}$ by the mean vector of $\boldsymbol{X}_k$. The distance is defined as the minimum distance from $\boldsymbol{x}_{new}^{c,k}$ to the vectors in $\mathcal{L}(\boldsymbol{W}_k)$:

$$d_k^{\mathcal{L}} = \min_{\boldsymbol{\alpha}_k^{\mathcal{L}}} ||\boldsymbol{x}_{new}^{c,k} - (\boldsymbol{W}_k \boldsymbol{\alpha}_k^{\mathcal{L}})^T||_2, \tag{1}$$

where $\boldsymbol{\alpha}_k^{\mathcal{L}} \in \mathbb{R}^{r_k \times 1}$ contains $r_k$ coefficients associated with the $r_k$ PCs in $\boldsymbol{W}_k$. The minimisation problem (1) has a closed-form solution of $\boldsymbol{\alpha}_k^{\mathcal{L}*} = (\boldsymbol{x}_{new}^{c,k} \boldsymbol{W}_k)^T$. Thus the distance can be written as $d_k^{\mathcal{L}} = ||\boldsymbol{x}_{new}^{c,k} - \boldsymbol{x}_{new}^{c,k} \boldsymbol{P}_k||_2$, where $\boldsymbol{P}_k = \boldsymbol{W}_k \boldsymbol{W}_k^T$ is the projection matrix of the subspace $\mathcal{L}(\boldsymbol{W}_k)$; $\boldsymbol{x}_{new}^{c,k} \boldsymbol{P}_k$ is the projection of $\boldsymbol{x}_{new}^{c,k}$ on $\mathcal{L}(\boldsymbol{W}_k)$. NSM assigns $\boldsymbol{x}_{new}$ to the class with the smallest $d_k^{\mathcal{L}}$:

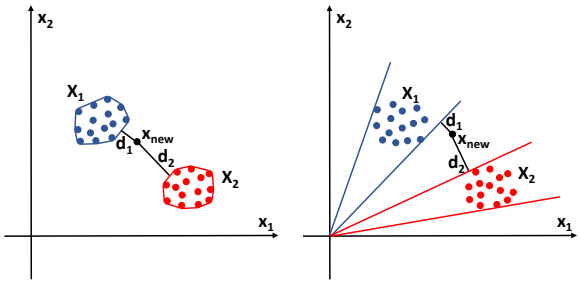
$$\hat{y}^{\mathcal{L}} = \underset{k}{\operatorname{argmin}}\, d_k^{\mathcal{L}}, \tag{2}$$

where $\hat{y}^{\mathcal{L}}$ denotes the predicted label for $\boldsymbol{x}_{new}$ by NSM.

### *B. Geometric convex models: NCHM, NCCM*

Besides the PC subspace, we can also model a class by using a geometric convex model in the training phase. There are two major differences between the PC subspace representation and the geometric convex model representation. First, the PC subspace is spanned by PCs which are the linear combinations of the original features, while the geometric convex model is constructed by the linear combinations of the class samples. To be more specific, the PC subspace is spanned by a set of vectors in $\boldsymbol{W}_k$, which are linear combinations of the original features in $\boldsymbol{X}_k$, i.e. the columns of $\boldsymbol{X}_k$. In contrast, the geometric convex model is for the linear combinations of the rows of $\boldsymbol{X}_k$. Second, since there are no constraints on the linear combination, the PC subspace representation has weak information about the location of the class samples. However, the geometric convex model representation imposes constraints on the linear combination of the training samples, providing more restricted areas for class representation.

Here we introduce the nearest convex hull method (NCHM) and the nearest convex cone method (NCCM), both based on the geometric convex model representation.




(a) NCHM (b) NCCM

Fig. 2: An illustrative example of NCHM and NCCM in a 2D space.

*1) Nearest convex hull method (NCHM):* Nalbantov et al. [15] propose the NCHM, which models each class as a convex hull by using the training instances in that class. An illustrative example of NCHM is shown in a 2D space in Fig. 2(a). The convex hulls of the two classes are shown as the blue and red polygons, respectively. Since $d_1 < d_2$, we assign $\boldsymbol{x}_{new}$ to class 1 in this example. In NCHM, we first define convex hull as follows.

**Definition II.2.** *Convex hull.* Let $S = \{\boldsymbol{x}_i\}_{i=1}^N$ be an arbitrary set in a linear vector space. The convex hull, $ch(S) = \{\boldsymbol{z} : \boldsymbol{z} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \mid \boldsymbol{x}_i \in S,\ 0 \leq \alpha_i \leq 1,\ \sum_{i=1}^{N} \alpha_i = 1\}$, is the smallest convex set containing $S$. In other words, $ch(S)$ is the intersection of all convex sets containing $S$.

Given the training samples $\boldsymbol{X}_k \in \mathbb{R}^{N_k \times p}$ of class $k$, the convex hull built by $\boldsymbol{X}_k$ is the set of vectors $\boldsymbol{z} \in \mathbb{R}^p$: $ch(\boldsymbol{X}_k) = \{\boldsymbol{z} : \boldsymbol{z} = \boldsymbol{X}_k^T \boldsymbol{\alpha}_k^{\mathcal{CH}} \mid 0 \leq \boldsymbol{\alpha}_k^{\mathcal{CH}} \leq 1,\ \mathbf{1}^T \boldsymbol{\alpha}_k^{\mathcal{CH}} = 1\}$, where $\boldsymbol{\alpha}_k^{\mathcal{CH}} \in \mathbb{R}^{N_k \times 1}$ is a vector containing the coefficients associated with the $N_k$ training samples in $\boldsymbol{X}_k$, $0 \leq \boldsymbol{\alpha}_k^{\mathcal{CH}} \leq 1$ means each element are in $[0, 1]$, and $\mathbf{1} \in \mathbb{R}^{N_k \times 1}$ has all elements of one.

Given a new sample $\boldsymbol{x}_{new} \in \mathbb{R}^{1 \times p}$, the distance from $\boldsymbol{x}_{new}$ to the convex hull $ch(\boldsymbol{X}_k)$ of the $k$th class is

$$d_k^{\mathcal{CH}} = \min_{\boldsymbol{\alpha}_k^{\mathcal{CH}}} \ ||\boldsymbol{x}_{new} - (\boldsymbol{X}_k^T \boldsymbol{\alpha}_k^{\mathcal{CH}})^T||_2, \quad \text{s.t. } 0 \leq \boldsymbol{\alpha}_k^{\mathcal{CH}} \leq 1,\ \mathbf{1}^T \boldsymbol{\alpha}_k^{\mathcal{CH}} = 1. \tag{3}$$

Then $\boldsymbol{x}_{new}$ is assigned to the class with the smallest $d_k^{\mathcal{CH}}$:

$$\hat{y}^{\mathcal{CH}} = \underset{k}{\operatorname{argmin}}\, d_k^{\mathcal{CH}}, \tag{4}$$

where $\hat{y}^{\mathcal{CH}}$ denotes the predicted label for $\boldsymbol{x}_{new}$ by NCHM.

*2) Nearest convex cone method (NCCM):* NCCM models each class as a convex cone by using the training instances in that class. We show an illustrative example of NCCM in a 2D space in Fig. 2(b). The convex cones for the two classes are shown as the blue and red triangular area, respectively. Since $d_1 < d_2$, we assign $\boldsymbol{x}_{new}$ to class 1 in this example. In NCCM, we first define convex polyhedral cone as follows.

**Definition II.3.** *Convex polyhedral cone.* A convex polyhedral cone is a convex cone that is generated by a finite number of generators. Let $S = \{\boldsymbol{x}_i\}_{i=1}^N$ be an arbitrary set in a linear vector space. The set, $cc(S) = \{\boldsymbol{z} : \boldsymbol{z} = \sum_{i=1}^N \alpha_i \boldsymbol{x}_i \mid \boldsymbol{x}_i \in S, \alpha_i \geq 0\}$, is the convex polyhedral cone generated by $S$.

Given the training samples $\boldsymbol{X}_k \in \mathbb{R}^{N_k \times p}$ of class $k$, the convex polyhedral cone built by $\boldsymbol{X}_k$ is defined as a set of vectors $\boldsymbol{z} \in \mathbb{R}^p$: $cc(\boldsymbol{X}_k) = \{\boldsymbol{z} : \boldsymbol{z} = \boldsymbol{X}_k^T \boldsymbol{\alpha}_k^{\mathcal{CC}} \mid \boldsymbol{\alpha}_k^{\mathcal{CC}} \geq 0\}$, where $\boldsymbol{\alpha}_k^{\mathcal{CC}} \in \mathbb{R}^{N_k \times 1}$ and $\boldsymbol{\alpha}_k^{\mathcal{CC}} \geq 0$ means each element in $\boldsymbol{\alpha}_k^{\mathcal{CC}}$ is nonnegative. Thus each vector in $cc(\boldsymbol{X}_k)$ is a conical combination of the samples in $\boldsymbol{X}_k$.

To assign a new sample $\boldsymbol{x}_{new} \in \mathbb{R}^{1 \times p}$ to one of the classes, we calculate the distance from $\boldsymbol{x}_{new}$ to $cc(\boldsymbol{X}_k)$:

$$d_k^{\mathcal{CC}} = \min_{\boldsymbol{\alpha}_k^{\mathcal{CC}}} ||\boldsymbol{x}_{new} - (\boldsymbol{X}_k^T \boldsymbol{\alpha}_k^{\mathcal{CC}})^T||_2, \ \text{ s.t. } \boldsymbol{\alpha}_k^{\mathcal{CC}} \geq 0. \tag{5}$$

Then $\boldsymbol{x}_{new}$ is assigned to the class with the minimum $d_k^{\mathcal{CC}}$:

$$\hat{y}^{\mathcal{CC}} = \underset{k}{\operatorname{argmin}}\, d_k^{\mathcal{CC}}, \tag{6}$$

where $\hat{y}^{\mathcal{CC}}$ denotes the predicted label for $\boldsymbol{x}_{new}$ by NCCM.

## III. Dual Analysis of the Minimum Distance Problems

Here we aim to establish a common platform to unify and compare the nearest-class-model methods through dual analysis of the minimum distance problems (1), (3) and (5). By studying the nearest-class-model methods together, we will have better understanding of the classification mechanisms of this important category of classification methods.

Dual analysis of the minimum distance problems enables us to find the separating hyperplanes, making finding the minimum distance from a sample to a class model equivalent to finding the maximum distance from that sample to a separating hyperplane. Different from the Euclidean distances used in the previous section, we discuss more general cases in the normed linear vector space with arbitrary norm in this section. Examples and illustrations for the Hilbert space are also presented for a better geometric understanding.

We first introduce some important theoretical settings in preliminary. Then we show the dual analysis for the three minimum distance problems (1), (3) and (5). The dual analysis for the subspace and the convex hull can be found in [25] and we only show their results in Theorems III.2 and III.3, respectively. In contrast, we provide a novel dual analysis and its proof for the convex cone in Theorem III.4, based on the relationship between a convex cone and its polar cone.

### A. Preliminary

**Definition III.1.** *Normed linear vector space.* A normed linear vector space is a vector space $\mathcal{X}$, on which a real-valued function is defined to map each element $\boldsymbol{x}$ in $\mathcal{X}$ into a real number $||\boldsymbol{x}||$ called the norm of $\boldsymbol{x}$. The norm satisfies the following axioms: 1) $||\boldsymbol{x}|| \geq 0$ for all $\boldsymbol{x} \in \mathcal{X}$, $||\boldsymbol{x}|| = 0$ if and only if $\boldsymbol{x} = 0$; 2) $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$ for each $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$; and 3) $||\alpha \boldsymbol{x}|| = |\alpha| ||\boldsymbol{x}||$ for all scalar $\alpha$ and each $\boldsymbol{x} \in \mathcal{X}$.

**Definition III.2.** *Linear functional.* A transformation from a vector space $\mathcal{X}$ into the space of real scalars is said to be a functional on $\mathcal{X}$. A functional $f$ on a vector space $\mathcal{X}$ is linear if for any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and any two scalars $\alpha$, $\beta$ there holds $f(\alpha \boldsymbol{x} + \beta \boldsymbol{y}) = \alpha f(\boldsymbol{x}) + \beta f(\boldsymbol{y})$.

**Definition III.3.** *The normed dual space.* Let $\mathcal{X}$ be a normed linear vector space. The space of all bounded linear functionals on $\mathcal{X}$ is called the normed dual of $\mathcal{X}$ and is denoted by $\mathcal{X}^*$. The norm of an element $f \in \mathcal{X}^*$ is $||f|| = \sup_{||x|| \leq 1} |f(x)|$.

Following [25], we use $\boldsymbol{x}^*$ to denote the linear functionals and write $\langle \boldsymbol{x}, \boldsymbol{x}^* \rangle$ to denote $f(\boldsymbol{x})$.

**Definition III.4.** *Real inner space.* A real inner space is a real linear vector space $\mathcal{X}$ together with an inner product, which is a map from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{R}$ and denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. The inner product satisfies the following axioms: 1) $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$; 2) $\langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle$; 3) $\langle \lambda \boldsymbol{x}, \boldsymbol{y} \rangle = \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle$; and 4)$\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$; $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x}$ is the origin.

**Definition III.5.** *Real Hilbert space.* A complete real inner space is called a real Hilbert space.

A Hilbert space has the following nice property. If $\boldsymbol{x}^*$ is a bounded linear functional on a Hilbert space $\mathcal{H}$, there exists a unique vector $\boldsymbol{w} \in \mathcal{H}$ such that for all $\boldsymbol{x} \in \mathcal{H}$, $\langle \boldsymbol{x}, \boldsymbol{x}^* \rangle = \langle \boldsymbol{x}, \boldsymbol{w} \rangle$. Moreover, we have $||\boldsymbol{x}^*|| = ||\boldsymbol{w}||$ and every $\boldsymbol{w}$ determines a unique bounded linear functional in this way.

### B. Hyperplane

Based on the above definitions, we define a hyperplane as follows and show some properties of a hyperplane that relates the primal problem with the dual problem.

**Definition III.6.** *Hyperplane.* A hyperplane $H$ in a linear vector space $\mathcal{X}$ is a maximal proper linear variety, that is, a linear variety $H$ such that $H \neq \mathcal{X}$, and if $V$ is any linear variety containing $H$, then either $V = \mathcal{X}$ or $V = H$.

**Proposition 1** ( [25]). *Let $H$ be a hyperplane in a linear vector space $\mathcal{X}$. Then there is a linear functional $\boldsymbol{x}^*$ on $\mathcal{X}$ and a constant $c$ such that $H = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle = c\}$. Conversely, if $\boldsymbol{x}^*$ is a nonzero linear functional on $\mathcal{X}$, the set $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle = c\}$ is a hyperplane in $\mathcal{X}$. $H$ is closed for every $c$ if and only if $\boldsymbol{x}^*$ is continuous.*

As shown in Proposition 1, hyperplanes have a close relationship with linear functionals. Thus the primal problem can be transformed to the dual problem by using the hyperplane as a media.

For a closed hyperplane $H$, we define two closed half-spaces: the negative half-space $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle \leq c\}$ and the positive half-space $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle \geq c\}$. The distance from a point to a hyperplane is of great importance in dual analysis, thus we introduce it in Theorem III.1.

**Theorem III.1** ( [26]). *Let $\boldsymbol{x}_e$ be an element in a real normed linear space $\mathcal{X}$ and let $d$ denote its distance from the hyperplane $H$: $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle = c\}$. Then, $d = \inf_{\boldsymbol{h} \in H} ||\boldsymbol{x}_e - \boldsymbol{h}|| = \frac{|\langle \boldsymbol{x}_e, \boldsymbol{x}^* \rangle - c|}{||\boldsymbol{x}^*||}$.*

### C. Dual analysis for NSM, NCHM and NCCM

*1) Dual analysis of the minimum distance problem in NSM:* In NSM, the separating hyperplane between an instance $\boldsymbol{x}_e$ and a subspace $\mathcal{M}$ is found based on the orthogonal complement $\mathcal{M}^\perp$ of $\mathcal{M}$, which is stated in Theorem 7. To make the theoretical settings clear, we first define the orthogonal complement of a subspace as follows.

**Definition III.7.** *Orthogonal complement.* Let $\mathcal{M}$ be a subset of a normed linear space $\mathcal{X}$. The orthogonal complement $\mathcal{M}^\perp$ of $\mathcal{M}$ consists of all elements $\boldsymbol{x}^* \in \mathcal{X}^*$ orthogonal to every vector in $\mathcal{M}$.

**Theorem III.2** ( [25]). *Let $\boldsymbol{x}_e$ be an element in a real normed linear space $\mathcal{X}$ and let $d$ denote its distance from the subspace $\mathcal{M}$. Suppose the orthogonal complement of $\mathcal{M}$ is $\mathcal{M}^\perp$. Then,*

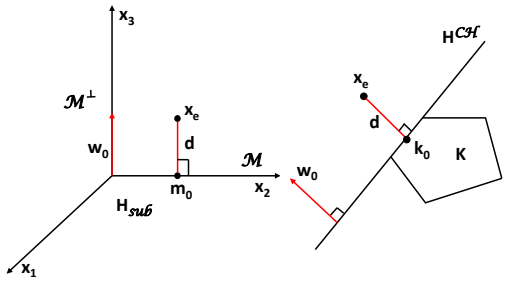
$$d = \inf_{\boldsymbol{m} \in \mathcal{M}} ||\boldsymbol{x}_e - \boldsymbol{m}|| = \max_{||\boldsymbol{x}^*|| \leq 1, \boldsymbol{x}^* \in \mathcal{M}^\perp} \langle \boldsymbol{x}_e, \boldsymbol{x}^* \rangle, \quad (7)$$

*where the maximum on the right is achieved for some $\boldsymbol{x}_0^* \in \mathcal{M}^\perp$.*

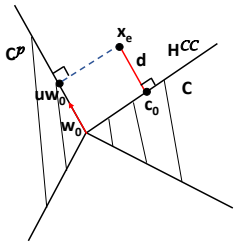
*If the infimum on the left is achieved for some $\boldsymbol{m}_0 \in \mathcal{M}$, then $\boldsymbol{x}_0^*$ is aligned with $\boldsymbol{x}_e - \boldsymbol{m}_0$, i.e. $\langle \boldsymbol{x}_e - \boldsymbol{m}_0, \boldsymbol{x}_0^* \rangle = ||\boldsymbol{x}_e - \boldsymbol{m}_0|| ||\boldsymbol{x}_0^*||$.*

Based on Theorem III.1, the right-hand side of (7) can be explained as the maximum distance from $\boldsymbol{x}_e$ to the hyperplane $H_{sub} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle = 0 \mid \boldsymbol{x}^* \in \mathcal{M}^\perp\}$, since the maximum is achieved when $||\boldsymbol{x}^*|| = 1$. Thus Theorem III.2 can be understood as: The minimum distance from a point $\boldsymbol{x}_e$ to the subspace $\mathcal{M}$ is equivalent to the maximum distance from $\boldsymbol{x}_e$ to the hyperplane $H_{sub}$.

For a better geometric understanding, we discuss Theorem III.2 in the Hilbert space. For each $\boldsymbol{x}^*$, we can find a unique $\boldsymbol{w} \in \mathcal{H}$ which is the normal vector of $H_{sub}$. Replace $\boldsymbol{x}^*$ by $\boldsymbol{w}$, the right-hand side of (7), i.e. $\langle \boldsymbol{x}_e, \boldsymbol{w} \rangle$, still denotes the distance from $\boldsymbol{x}_e$ to $H_{sub}$ since the maximum is achieved for $||\boldsymbol{w}|| = ||\boldsymbol{x}^*|| = 1$. We also have $\langle \boldsymbol{x}_e - \boldsymbol{m}_0, \boldsymbol{w}_0 \rangle = ||\boldsymbol{x}_e - \boldsymbol{m}_0|| ||\boldsymbol{w}_0||$, thus $\boldsymbol{x}_e - \boldsymbol{m}_0 = \mu \boldsymbol{w}_0$ ($\mu > 0$). For any vector $\boldsymbol{m} \in \mathcal{M}$, $\langle \boldsymbol{x}_e - \boldsymbol{m}_0, \boldsymbol{m} \rangle = \langle \mu \boldsymbol{w}_0, \boldsymbol{m} \rangle = \mu \langle \boldsymbol{w}_0, \boldsymbol{m} \rangle = 0$, as $\boldsymbol{w}_0 \in \mathcal{M}^\perp$. This indicates that $\boldsymbol{x}_e - \boldsymbol{m}_0$ has the same direction as $\boldsymbol{w}_0$ and $\boldsymbol{x}_e - \boldsymbol{m}_0$ is perpendicular to $\mathcal{M}$.




(a) Theorem III.2 of NSM (b) Theorem III.3 of NCHM




(c) Theorem III.4 of NCCM

Fig. 3: Illustrative examples of (a) Theorem III.2 of NSM, (b) Theorem III.3 of NCHM and (c) Theorem III.4 of NCCM.

Fig. 3(a) illustrates an example of Theorem III.2. Suppose $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are the orthogonal bases for $\mathcal{R}^3$. Assume $\mathcal{M}$ is the subspace spanned by $\boldsymbol{x}_2$. Thus $\mathcal{M}^\perp$ is the subspace spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$. Suppose $\boldsymbol{x}_e$ lies in the subspace spanned by $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. Then the minimum distance from $\boldsymbol{x}_e$ to $\mathcal{M}$ is achieved at the point $\boldsymbol{m}_0$; and the maximum distance from $\boldsymbol{x}_e$ to any subspaces with their normal vectors in $\mathcal{M}^\perp$ is attained when $\boldsymbol{w}_0$ has the same direction as $\boldsymbol{x}_3$; the subspace associated with this maximum distance is denoted by $H_{sub}$, which is a plane spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, as illustrated in Fig. 3(a). That is, we can find that these two distances are the same, both equal to $d$. The hyperplane with the normal vector $\boldsymbol{w}_0$ is actually the subspace spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The vector $\boldsymbol{x}_e - \boldsymbol{m}_0$ has the same direction as $\boldsymbol{w}_0$. This result is clear with simple geometry, if we treat $\boldsymbol{m}_0$ as the orthogonal projection of $\boldsymbol{x}_e$ on the subspace $\mathcal{M}$.

*2) Dual analysis of the minimum distance problem in NCHM:* In NCHM, the maximum distance between $\boldsymbol{x}_e$ and a separating hyperplane that separates $\boldsymbol{x}_e$ and a convex hull $K$ is achieved when the separating hyperplane is a supporting hyperplane of $K$. The details are shown in Theorem 8.

**Theorem III.3** ( [25]). *Let $\boldsymbol{x}_e$ be a point in a real normed*

*vector space $X$ and let $d > 0$ denote its distance from the convex set $K$ having support functional $h$, i.e. $h(\boldsymbol{x}^*) = \sup_{\boldsymbol{k}\in K}\langle \boldsymbol{k}, \boldsymbol{x}^*\rangle$. Then*

$$d = \inf_{\boldsymbol{k}\in K} ||\boldsymbol{x}_e - \boldsymbol{k}|| = \max_{||\boldsymbol{x}^*||\leq 1} [\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle - h(\boldsymbol{x}^*)], \quad (8)$$

*where the maximum on the right is achieved by some $\boldsymbol{x}_0^* \in \mathcal{X}^*$.*

*If the infimum on the left is achieved by some $\boldsymbol{k}_0 \in K$, then $\boldsymbol{x}_0^*$ is aligned with $\boldsymbol{x}_e - \boldsymbol{k}_0$, i.e. $\langle \boldsymbol{x}_e - \boldsymbol{k}_0, \boldsymbol{x}_0^*\rangle = ||\boldsymbol{x}_e - \boldsymbol{k}_0||||\boldsymbol{x}_0^*||$.*

The right-hand side of (8) can be understood as the maximum distance from $\boldsymbol{x}_e$ to the hyperplane $H^{\mathcal{CH}} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = h(\boldsymbol{x}^*)\}$. Thus Theorem III.3 indicates that the minimum distance from $\boldsymbol{x}_e$ to the convex hull is equivalent to the maximum distance from $\boldsymbol{x}_e$ to the hyperplane $H^{\mathcal{CH}}$.

In the Hilbert space, we can find a unique $\boldsymbol{w}_0 \in \mathcal{H}$ for $\boldsymbol{x}_0^*$. Since $\boldsymbol{x}_0^*$ is aligned with $\boldsymbol{x}_e - \boldsymbol{k}_0$, $\boldsymbol{x}_e - \boldsymbol{k}_0 = \mu \boldsymbol{w}_0$ ($\mu > 0$) and $\boldsymbol{x}_e - \boldsymbol{k}_0$ has the same direction as $\boldsymbol{w}_0$.

Fig. 3(b) shows an intuitive example of Theorem III.3 in $\mathbb{R}^2$. The minimum distance from $\boldsymbol{x}_e$ to $K$ is achieved at point $\boldsymbol{k}_0$, which lies on the nearest face of $K$ to $\boldsymbol{x}_e$. The maximum distance between $\boldsymbol{x}_e$ and $H^{\mathcal{CH}}$ that separates $\boldsymbol{x}_e$ and $K$ is achieved when the nearest face of $K$ to $\boldsymbol{x}_e$ is in $H^{\mathcal{CH}}$. The normal vector $\boldsymbol{w}_0$ is perpendicular to $H^{\mathcal{CH}}$ and has the same direction as $\boldsymbol{x}_e - \boldsymbol{k}_0$.

*3) Dual analysis of the minimum distance problem in NCCM:* Inspired by the relationship between $\mathcal{M}$ and $\mathcal{M}^{\perp}$ used in Theorem III.2, we apply the relationship between a convex cone and its polar cone to the dual analysis of (5) to obtain the separating hyperplane for NCCM in Theorem III.4. We first introduce the definition of a polar cone and then show Theorem III.4 and its proof.

**Definition III.8.** *Polar cone.* Given a convex polyhedral cone $C$ in a normed space $\mathcal{X}$, the set $C^p = \{\boldsymbol{x}^* \in \mathcal{X}^* : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle \leq 0,\ \forall \boldsymbol{x} \in C\}$ is called the polar cone of $C$.

If $\boldsymbol{x}_e$ is an interior point of $C$, then $d = 0$, which is a trivial case. Thus in the following theorem, we discuss the case when $\boldsymbol{x}_e$ is not an interior point of $C$ with $d > 0$.

**Theorem III.4.** *Let $\boldsymbol{x}_e$ be an element in a real normed linear space $\mathcal{X}$. Let $d > 0$ denote the distance from $\boldsymbol{x}_e$ to the convex cone $C$. Then,*

$$d = \inf_{\boldsymbol{c}\in C} ||\boldsymbol{x}_e - \boldsymbol{c}|| = \max_{||\boldsymbol{x}^*||\leq 1, \boldsymbol{x}^*\in C^p} \langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle,$$

*where the maximum on the right is achieved for some $\boldsymbol{x}_0^* \in C^p$.*

*If the infimum on the left is achieved for some $\boldsymbol{c}_0 \in C$, then $\boldsymbol{x}_0^*$ is aligned with $\boldsymbol{x}_e - \boldsymbol{c}_0$, i.e. $\langle \boldsymbol{x}_e - \boldsymbol{c}_0, \boldsymbol{x}_0^*\rangle = ||\boldsymbol{x}_e - \boldsymbol{c}_0||||\boldsymbol{x}_0^*||$.*

*Proof.* We first show that there exist some $\boldsymbol{x}^* \in C^p$ with the hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = 0\}$ being able to separate $\boldsymbol{x}_e$ and $C$. The two closed half-spaces associated with the hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = 0\}$ are $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle \geq 0\}$ and $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle \leq 0\}$. When $\boldsymbol{x}^* \in C^p$, $\langle \boldsymbol{c}, \boldsymbol{x}^*\rangle \leq 0$ for $\boldsymbol{c} \in C$, and $C$ is in the negative half-space. Since $\boldsymbol{x}_e$ is not an interior point of $C$, we can find some $\boldsymbol{x}^* \in C^p$ such that $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle \geq 0$ and $\boldsymbol{x}_e$ is in the positive half-space. Thus $\boldsymbol{x}_e$ and $C$ lie in opposite half-spaces determined by the hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = 0\}$ with $\boldsymbol{x}^* \in C^p$.

Let $S(\epsilon)$ be the sphere centred at $\boldsymbol{x}_e$ of radius $\epsilon$. For $\boldsymbol{x}^* \in C^p$ having $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle \geq 0$ and $||\boldsymbol{x}^*|| = 1$, let $\epsilon^*$ be the supremum of the $\epsilon$'s for which the hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = 0\}$ separates $C$ and $S(\epsilon)$. It is clear that $0 \leq \epsilon^* \leq d$. Also $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle = \epsilon^*$ when $||\boldsymbol{x}^*|| = 1$. Thus, for every $\boldsymbol{x}^* \in C^p$ having $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle \geq 0$ and $||\boldsymbol{x}^*|| = 1$, we have $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle \leq d$.

On the other hand, since $C$ contains no interior point of $S(d)$, there is a hyperplane separating $C$ and $S(d)$, and thus an $\boldsymbol{x}_0^* \in C^p$ such that $\langle \boldsymbol{x}_e, \boldsymbol{x}^*\rangle = d$.

To prove the alignment statement, suppose $\boldsymbol{c}_0 \in C$ and $||\boldsymbol{x}_e - \boldsymbol{c}_0|| = d$. Since $\boldsymbol{c}_0 \in C$, $\langle \boldsymbol{c}_0, \boldsymbol{x}_0^*\rangle \leq 0$ and $\langle \boldsymbol{x}_e - \boldsymbol{c}_0, \boldsymbol{x}_0^*\rangle \geq \langle \boldsymbol{x}_e, \boldsymbol{x}_0^*\rangle = d$. However, according to the Cauchy-Schwarz inequality, $\langle \boldsymbol{x}_e - \boldsymbol{c}_0, \boldsymbol{x}_0^*\rangle \leq ||\boldsymbol{x}_e - \boldsymbol{c}_0||||\boldsymbol{x}_0^*|| = d$. Thus $\langle \boldsymbol{x}_e - \boldsymbol{c}_0, \boldsymbol{x}_0^*\rangle = ||\boldsymbol{x}_e - \boldsymbol{c}_0||||\boldsymbol{x}_0^*|| = d$ and $\boldsymbol{x}_0^*$ is aligned with $\boldsymbol{x}_e - \boldsymbol{c}_0$. □

Theorem III.4 indicates that the minimum distance between $\boldsymbol{x}_e$ and $C$ is equivalent to the maximum distance between $\boldsymbol{x}_e$ and the hyperplane $H^{\mathcal{CC}} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle = 0 \mid \boldsymbol{x}^* \in C^p, ||\boldsymbol{x}^*|| = 1\}$ that separates $\boldsymbol{x}_e$ and $C$.

In the Hilbert space, we can find a unique $\boldsymbol{w}_0 \in \mathcal{H}$ for $\boldsymbol{x}_0^*$. Substituting $\boldsymbol{w}_0$ with $\boldsymbol{x}_0^*$, we can get $\langle \boldsymbol{x}_e, \boldsymbol{w}_0\rangle = d$. Also $\langle \boldsymbol{x}_e - \boldsymbol{c}_0, \boldsymbol{w}_0\rangle = ||\boldsymbol{x}_e - \boldsymbol{c}_0||||\boldsymbol{w}_0|| = d$. The equality holds when $\boldsymbol{x}_e - \boldsymbol{c}_0 = \mu \boldsymbol{w}_0$ ($\mu > 0$). Thus we can get the following two conclusions. First, $\langle \boldsymbol{c}_0, \boldsymbol{w}_0\rangle = 0$, which indicates that $\boldsymbol{c}_0$ and $\boldsymbol{w}_0$ are orthogonal. Second, $\boldsymbol{x}_e = \boldsymbol{c}_0 + \mu \boldsymbol{w}_0$, which indicates that $\boldsymbol{x}_e$ can be decomposed to $\boldsymbol{c}_0 \in C$ and $\mu \boldsymbol{w} \in C^p$. These two conclusions indicates that the orthogonal decompositions of $\boldsymbol{x}_e$ to $C$ and $C^p$ are $\boldsymbol{c}_0$ and $\mu \boldsymbol{w}_0$, respectively. Based on the Moreau's theorem in the Hilbert space [27], $\boldsymbol{c}_0$ and $\mu \boldsymbol{w} \in C^p$ are the projections of $\boldsymbol{x}_e$ on $C$ and $C^p$, respectively.

Fig. 3(c) illustrates Theorem III.4 in $\mathbb{R}^2$. The minimum distance $d$ from $\boldsymbol{x}_e$ to $C$ is achieved by $\boldsymbol{c}_0$, which is the orthogonal projection of $\boldsymbol{x}_e$ to the nearest face of $C$ to $\boldsymbol{x}_e$. The maximum distance from $\boldsymbol{x}_e$ to $H^{\mathcal{CC}}$ is achieved when $H^{\mathcal{CC}}$ contains the nearest face of $C$ to $\boldsymbol{x}_e$. It is obvious that the distance from $\boldsymbol{x}_e$ to this $H^{\mathcal{CC}}$ is also $d$. The normal vector associated with this hyperplane is $\boldsymbol{w}_0$, which has the same direction as $\boldsymbol{x}_e - \boldsymbol{c}_0$; the point $\mu \boldsymbol{w}_0$ is the orthogonal projection of $\boldsymbol{x}_e$ to $C^p$.

## IV. Unify the Nearest-Class-Model Methods

In Section IV-A, we propose the novel separating hyperplane classification (SHC) framework based on the theoretical discussion in Section III. The nearest-class-model methods can be unified under the SHC framework with different set of constraints on the normal vectors $\boldsymbol{w}$ and the bias $b$.

The SHC framework has two advantages. First, we can explain the empirical classification performance by analysing the discriminative abilities of the normal vectors. Second, we can design new nearest-class-model methods by imposing appropriate constraints to the framework based on the properties of the data. We show an example of designing a new soft NCCM classifier under the SHC framework, to solve the overlapping class model problem in Section IV-B.

### A. A novel separating hyperplane classification (SHC) framework

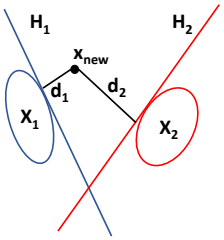



Fig. 4: The separating hyperplane classification framework.

The dual analysis enables us to explain the classification schemes of NSM, NCCM and NCHM from the separating hyperplane point of view. Theorems III.2, III.3 and III.4 indicate that the three methods all classify a test sample by using separating hyperplanes associated with each class. We illustrate a binary classification case in Fig. 4. The red and blue ellipses represent the two class models, respectively, and the red and blue lines represent the separating hyperplanes between a new instance $\boldsymbol{x}_{new}$ and the class models, respectively. $\boldsymbol{x}_{new}$ is classified by comparing its distance to the two separating hyperplanes.

Based on the separating hyperplanes, we can derive a new separating hyperplane classification (SHC) framework for different class representation models and distances with arbitrary norms: First, for the $k$th class, we obtain

$$\max_{c_k, ||\boldsymbol{x}_k^*||=1} \quad d_k = \langle \boldsymbol{x}_{new}, \boldsymbol{x}_k^* \rangle - c_k$$
$$\text{s.t. constraint}(\boldsymbol{x}_k^*, c_k), \tag{9}$$

where $\boldsymbol{x}_k^*$ and $c_k$ are the two parameters to define the separating hyperplane $H_k = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{x}_k^* \rangle = c_k\}$ between $\boldsymbol{x}_{new}$ and the $k$th class model, and constraint$(\boldsymbol{x}_k^*, c_k)$ denotes constraints on $\boldsymbol{x}_k^*$ and $c_k$. Then, $\boldsymbol{x}_{new}$ is assigned to the class $k$ with the minimum $d_k$.

This SHC framework for two-class classification can be explained as follows. For each test sample, we find a pair of separating hyperplanes that separate the test sample and the two class models, respectively. The test sample is then assigned to the class with the minimum distance from that sample to the corresponding hyperplane.

In the SHC framework, the normal vectors of the separating hyperplanes play important roles in classification. Theorems III.2, III.3 and III.4 suggest that the dual function $\boldsymbol{x}_0^*$ that determines the separating hyperplane is aligned with the vector $\boldsymbol{x}_{new} - \boldsymbol{x}_0$, where $\boldsymbol{x}_0$ is the nearest point to $\boldsymbol{x}_{new}$ in the class model. In the Hilbert space, this means that the normal vector of the separating hyperplane is parallel with $\boldsymbol{x}_{new} - \boldsymbol{x}_0$. The norm of $\boldsymbol{x}_{new} - \boldsymbol{x}_0$ is defined as the distance from $\boldsymbol{x}_{new}$ to the class model. Thus the discriminative information contained in the direction of $\boldsymbol{x}_{new} - \boldsymbol{x}_0$, which is also the direction of the associated normal vector of the hyperplane, is vital to classification. The more the discriminative information contained in the normal vector, the higher the classification accuracy. In other words, to get better classification, constraints should be specified to make the normal vector contain more discriminative information.

In NSM, NCHM and NCCM, the Euclidean norm $||\cdot||_2$ is used. We summarise constraint$(\boldsymbol{x}_k^*, c_k)$ for NSM, NCHM and NCCM in Table I. Note that $\boldsymbol{x}_k^*$ is replaced by $\boldsymbol{w}_k$. For NSM, $\boldsymbol{w}_k$ has a closed-form solution of $\boldsymbol{x}_{new}^{c,k} - \boldsymbol{x}_{new}^{c,k}\boldsymbol{P}_k$, where $\boldsymbol{x}_{new}^{c,k}$ is the centred $\boldsymbol{x}_{new}$ by the column mean of $\boldsymbol{X}_k$.

TABLE I: constraint$(\boldsymbol{x}_k^*, c_k)$ for NSM, NCHM and NCCM.

| NSM | NCHM | NCCM |
|---|---|---|
| $\langle \boldsymbol{x}_{new}^{c,k}\boldsymbol{P}_k, \boldsymbol{w}_k \rangle = 0$<br>$c_k = 0$ | $\langle \boldsymbol{x}_{new}, \boldsymbol{w}_k \rangle \geq c_k$<br>$\langle \boldsymbol{x}_i^k, \boldsymbol{w}_k \rangle \leq c_k$ | $\langle \boldsymbol{x}_{new}, \boldsymbol{w}_k \rangle \geq 0$<br>$\langle \boldsymbol{x}_i^k, \boldsymbol{w}_k \rangle \leq 0$<br>$c_k = 0$ |
| $\boldsymbol{P}_k$ denotes the projection matrix for class $k$.<br>$\boldsymbol{x}_i^k \in \mathbb{R}^{1\times p}$ denotes the $i$th row in $\boldsymbol{X}_k$. | | |

### B. A novel soft nearest-convex-cone method (soft NCCM)

Besides the constraints listed in Table I, other constraints can also be specified based on the properties of the dataset and the requirements from the user, to extend further. In this section, we show an example of designing a new nearest-class-model method under the SHC framework, to better classify data with overlapping class models.

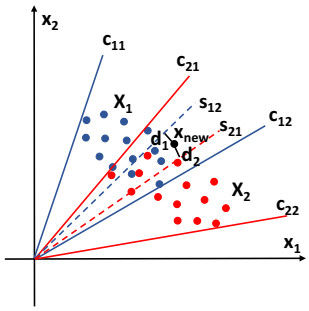



Fig. 5: An illustrative example of soft NCCM in a 2D space.

When the class models overlap, the class memberships of the test instances locating in the overlapping area are ambiguous and cannot be determined by the nearest-class-model methods. This is because the distances from those instances to class models are all zeros and we cannot find a hyperplane to separate them with the class models. We illustrate this situation in Fig. 5. The original convex cone models are shown by the triangular areas constructed by the solid lines: the blue solid lines $c_{11}$ and $c_{12}$ form the convex cone for the first class while the red solid lines $c_{21}$ and $c_{22}$ form the convex cone for the second class. We can observe a large overlapping area between the two convex cones. The instances located between $c_{21}$ and $c_{12}$ cannot be clearly classified to a specific class because of the overlapping problem.

To address this problem, we propose a novel nearest-class-model method, the soft NCCM classifier, by imposing proper constraints into the optimisation problem (9). In Fig. 5, by using the soft NCCM, we expect to get two separating hyperplanes shown by the two dashed lines, $s_{12}$ and $s_{21}$, for

the first class and the second class, respectively. It is clear that we actually reduce the areas of the convex cones by pushing the overlapping boundaries $c_{12}$ and $c_{21}$ towards $\boldsymbol{x}_{new}$ and obtain the new boundaries $s_{12}$ and $s_{21}$, respectively. The resulting 'soft' convex cones of the two classes are constructed by the blue lines $c_{11}$ and $s_{12}$ for the first class and the red lines $c_{22}$ and $s_{21}$ for the second class. Thus $\boldsymbol{x}_{new}$ can be then classified by comparing $d_1$ and $d_2$ to the two separating hyperplanes $s_{12}$ and $s_{21}$, respectively.

In soft NCCM, we design the constraints to achieve the following two aims: first, the test instances in the overlapping area can be classified unambiguously, and second, the discriminative between-class information is utilised to make separating hyperplanes better for classification. The optimisation problem is written as follows:

$$\begin{aligned} \max_{||\boldsymbol{w}_k||_2=1} \quad & d_k^{\mathcal{SCC}} = \boldsymbol{w}_k^T \boldsymbol{x}_{new}, \\ \text{s.t.} \quad & \boldsymbol{w}_k^T \boldsymbol{x}_{new} \geq 0, \\ & \boldsymbol{w}_k^T \boldsymbol{x}_i^k \leq \xi_i, \quad i = 1, 2, \ldots, N_k, \\ & \boldsymbol{w}_k^T \boldsymbol{x}_j^{-k} \geq -\xi_j, \quad j = 1, 2, \ldots, N_{-k}, \\ & \xi_i \geq 0 \; \forall i, \xi_j \geq 0 \; \forall j, \sum_i \xi_i + \sum_j \xi_j \leq C, \end{aligned} \tag{10}$$

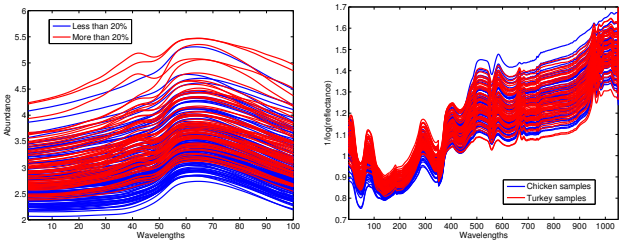
where the subscript $k$ denotes the $k$th class while $-k$ denotes all other classes, i.e. $N_k$ is the number of training samples in the $k$th class and $N_{-k}$ is the number of training samples in all classes except for the $k$th class.

To achieve the first aim, we introduce slack variables $\xi_i$, allowing some of the training instances from the $k$th class to locate on the same side of the separating hyperplane as $\boldsymbol{x}_{new}$. In this way, we can find a hyperplane that can separate $\boldsymbol{x}_{new}$ and the convex cone class model with tolerance of errors, even when $\boldsymbol{x}_{new}$ locates in the convex cone. Thus there is no overlapping area when we use the separating hyperplanes to classify $\boldsymbol{x}_{new}$, and an unambiguous class membership can be obtained. To achieve the second aim, we propose the third constraint which utilises the discriminative information from other classes and makes the training instances from the $k$th class and those from all other classes locate on different sides of the separating hyperplane corresponding to the $k$th class.

## V. Experiments

For illustration, we apply NSM, NCHM, NCCM and soft NCCM to two types of high-dimensional data, the spectroscopic data and the face image data, in Sections V-A and V-B, respectively. For each type of data, we first show the classification results of the four nearest-class-model methods. The classification performances of a popular classification method for high-dimensional data, support vector machine (SVM), are also recorded to show the effectiveness of the nearest-class-model methods. We then analyse why a class model performs better than others, by exploring the discriminative abilities of the normal vectors based on the SHC framework.

### A. The spectroscopic data




(a) The fat dataset (b) The meat dataset

Fig. 6: The spectroscopic datasets.

*1) Datasets:* We use two high-dimensional spectroscopic datasets, the fat dataset and the meat dataset, in the following experiments.

The fat dataset [28] measures the spectra of finely chopped meat, which can be downloaded from http://lib.stat.cmu.edu/datasets/tecator. Each spectrum is measured at 100 wavelengths. The dataset contains 193 spectra, with 122 meat samples of less than 20% fat and 71 samples of larger than 20% fat. Fig. 6(a) shows the spectra of the fat dataset.

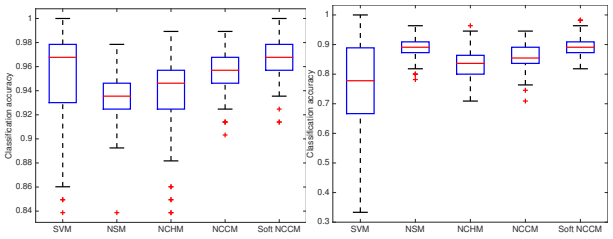
For the fat dataset, a training set contains 100 randomly selected samples, with 35 samples of less than 20% fat and 35 samples of larger than 20% fat, and a test set contains the remaining samples.

The meat dataset contains 55 chicken and 54 turkey meat spectra measured at 1051 wavelengths. We use the first 350 wavelengths ranging from 400 to 1100 nm, following the suggestion in Arnalds et al. [29]. Fig. 6(b) shows the spectra of the meat dataset.

For the meat dataset, a training set contains 27 chicken samples and 27 turkey samples, and a test set contains 28 chicken samples and 27 turkey samples.

*2) Experiment settings:* In NSM, the dimensions of the two class subspaces are tuned by 10-fold cross-validation on the training set. The dimensions are chosen to minimise the classification error. In NCHM, the optimisation problem (3) is solved using the 'cvx' package in MATLAB. In NCCM, the optimisation problem (5) is solved using the 'lsqnonneg' function in MATLAB. In soft NCCM, the optimisation problem (10) and the parameter $C$ is tuned by 10-fold cross-validation from $[10^{-1}, 1, 10]$. In SVM, the linear kernel is adopted, because it is usually recommended for high-dimensional data [30]. We randomly split the data to a training set and a test set 100 times and the experiments are performed on all training/test splits. The classification accuracies of all the experiments are recorded and depicted in boxplots.

*3) Classification results:* The classification accuracies of SVM, NSM, NCHM, NCCM and soft NCCM for the two datasets are shown in Fig. 7. It is clear that soft NCCM can provide the best classification performances for both datasets. In both cases, NCHM performs worse than NCCM, which suggests that the convex hull class model might be too tight for the spectroscopic data and the convex cone class model can be a better choice. In addition, soft NCCM can provide better classification accuracies than NCCM, which suggests



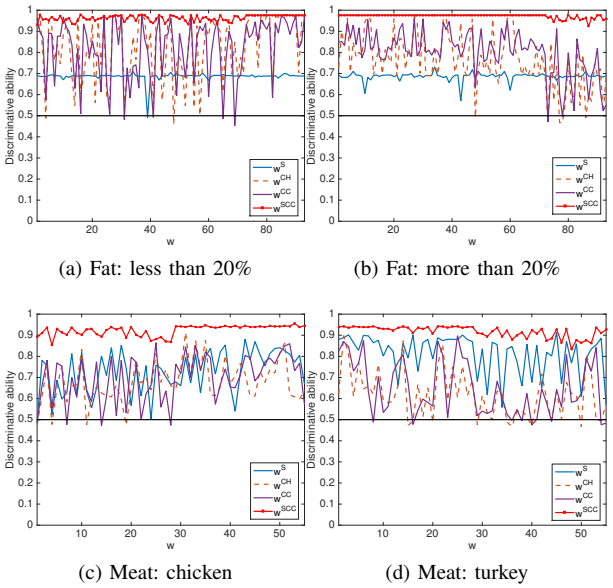

(a) The fat dataset (b) The meat dataset

Fig. 7: The classification accuracies of SVM, NSM, NCHM, NCCM and soft NCCM on the two spectroscopic datasets.

the effectiveness of the constraints that we propose in (10).

For the fat data, SVM and soft NCCM have the best median accuracies. However, it is obvious that soft NCCM has a much smaller variance in classification accuracies than SVM. The classification performances of NSM, NCHM are worse than that of NCCM, which suggests that convex cone is a better class model than PC subspace and convex hull for this dataset.

For the meat data, SVM performs the worst with a large variation. Soft NCCM has a similar median to NSM while less extreme low accuracies than NSM. Comparing the classification performances of NSM, NCHM and NCCM, we can state that PC subspace is a better class model for this dataset compared with the geometric class models.




(a) Fat: less than 20% (b) Fat: more than 20%

(c) Meat: chicken (d) Meat: turkey

Fig. 8: The discriminative abilities, denoted by $\boldsymbol{w}^S$, $\boldsymbol{w}^{CH}$, $\boldsymbol{w}^{CC}$ and $\boldsymbol{w}^{SCC}$, of the normal vectors of NSM, NCHM, NCCM and soft NCCM, respectively, for the two spectroscopic datasets.

*4) Analysis of classification results:* Section V-A3 shows that different data prefer different class models. To understand this pattern, we compare the normal vectors of pairs of separating hyperplanes of the four methods. As discussed in Section IV-A, the more discriminative the normal vectors are, the higher the classification accuracy.

Here we measure the discriminative ability of the normal vectors by the classification accuracies of linear discriminant analysis (LDA). More specifically, for each test instance, we have two normal vectors associated with the two separating hyperplanes for the two classes, respectively. We project all the instances to each normal vector and apply LDA on the projected instances based on 100 random training/test splits. The mean classification accuracies are recorded for the two normal vectors. We repeat this procedure for all test instances from one training/test split in the previous section.

We show in Fig. 8 the mean classification accuracies for the normal vectors of NSM, NCHM, NCCM and soft NCCM, $\boldsymbol{w}^S$, $\boldsymbol{w}^{CH}$, $\boldsymbol{w}^{CC}$ and $\boldsymbol{w}^{SCC}$. The horizontal line indices the normal vectors for the test instances and the vertical line denotes the discriminative abilities. The black solid line shows the classification accuracy of 0.5, which is a threshold indicating with and without discriminative ability.

Obviously, the normal vectors of soft NCCM, $\boldsymbol{w}^{SCC}$, has the best discriminative abilities with the highest curves for both classes and both datasets, which is consistent with its best classification performances on the two datasets. For the fat data, $\boldsymbol{w}^S$ of NSM has much lower discriminative abilities in most cases, which is also consistent with its worst classification performance compared with other methods. For the meat data, NSM has better classification performance than NCHM and NCCM and this is also shown in the turkey meat class in Fig. 8(d): $\boldsymbol{w}^S$ has a higher curve than $\boldsymbol{w}^{CH}$ and $\boldsymbol{w}^{CC}$.

### B. The face image data

*1) Dataset:* To further show the effectiveness of the proposed SHC framework and the new soft NCCM, we also apply the methods to another popular type of high-dimensional data, the face image data. We use the extended Yale face database B [11] as an exemplar. The database contains 38 individuals, each with around 64 near frontal images under different illuminations. Each image has a frontal face cropped from the original image and is resized to 32×32 pixels. Fig. 9 shows the 64 face images of one individual. Here we take the first ten individuals in the experiments for illustration.

Fig. 9: Example face images in the Yale face database B.

*2) Experiment settings:* We randomly split the dataset to a training set containing 80% of the data and a test set containing 20% of the data. We repeat all experiments for 20 random training/test splits and record the corresponding classification accuracies. The experiment settings for the classification methods are the same as those for the spectroscopic data.

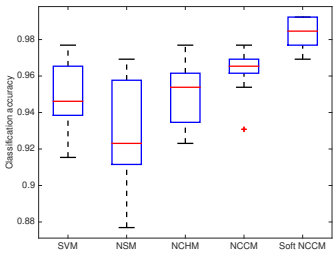



Fig. 10: The classification accuracies of SVM, NSM, NCHM, NCCM and soft NCCM on the face image data.

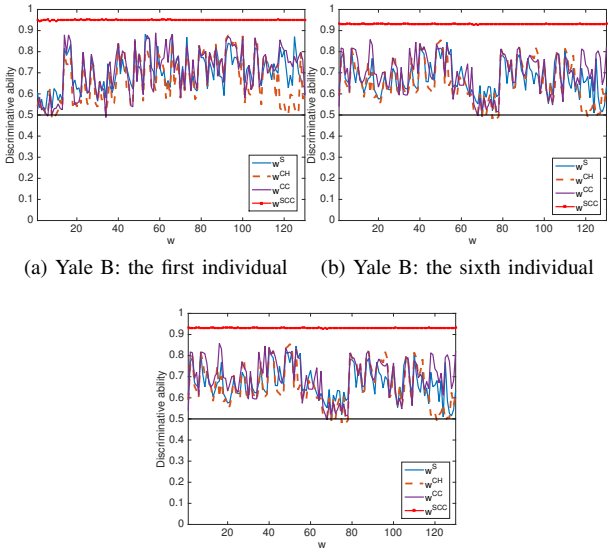
*3) Classification results:* Fig. 10 shows the boxplots for the classification accuracies of the five classifiers. All methods have median accuracies over 0.9, which shows their effectiveness to classify face image data. NSM has the lowest box and the largest variation among all methods. SVM is competitive with NCHM, while NCHM has a higher median accuracy.

The two cone-based methods, NCCM and soft NCCM, show the best classification accuracies. This is reasonable because the frontal face images under various illumination conditions can be effectively represented by an illumination cone [11]. Soft NCCM performs even better than NCCM, which shows the advantage of considering the between-class information.

*4) Analysis of classification results:* To show the discriminative ability of the normal vector, we make a slight change to the multi-class case here compared with the binary case in the spectroscopic data. We first project all the data to the direction of the normal vector and then apply LDA to do binary classification: one is for the class associated with that normal vector and the other is for other classes. Given one normal vector, we repeat this process nine times for all other nine classes and take the average of the mean classification accuracies as the discriminative ability of that normal vector. The reason for this is that, based on the separating hyperplane corresponding to one class, it is hard to achieve multi-class classification. It is natural to use this separating hyperplane to distinguish between the corresponding class and the other classes. All other settings to analyse the classification results are the same as those for the spectroscopic data.

For the ten classes tested in the experiments, we can find the discriminative abilities of the ten normal vectors. Figs. 11(a) and 11(b) show the discriminative abilities of the normal vectors corresponding to the first and sixth individuals in one training/test split, respectively, for example. The normal vectors of soft NCCM have the highest discriminative abilities. We can also observe that the curves for $\boldsymbol{w}^{CC}$ are slightly above those for $\boldsymbol{w}^{S}$ and $\boldsymbol{w}^{CH}$ in most cases. To visualise the ten plots together, we also plot the means of the discriminative abilities of the ten normal vectors in Fig. 11(c). The pattern is roughly the same as that in Figs. 11(a) and 11(b).

To sum up, we can draw two conclusions from the experiments on both the high-dimensional spectroscopic and face image datasets. First, the new soft NCCM that solves the overlapping class model problem has the best classification performances over all compared methods. Second, the discriminative ability of the normal vector is associated with the classification performance of nearest-class-model methods, which demonstrates the effectiveness of the new SHC framework in explaining the classification results.




(a) Yale B: the first individual (b) Yale B: the sixth individual

(c) Yale B: the mean discriminative abilities

Fig. 11: The discriminative abilities of two normal vectors and the mean discriminative abilities of NSM, NCHM, NCCM and soft NCCM for the Yale face database B.

## VI. Conclusions and Future Work

In this paper, we establish a new separating hyperplane classification (SHC) framework to unify three nearest-class-model methods for high-dimensional data: NSM, NCHM and NCCM. The SHC framework is established on the theoretical results from the dual analysis of the three methods. We show a new theorem for the dual analysis of NCCM by discovering the relationship between a convex cone and its polar cone.

Based on this novel SHC framework, we can explain why one class model is good to classify a specific dataset by showing the discriminative ability of the normal vectors of the separating hyperplanes. The higher the discriminative abilities of the normal vectors, the higher the classification accuracy of one method. The experiment results also demonstrate this argument. In addition, we propose a new soft NCCM under the SHC framework to solve the overlapping class model problem. The experiments on both spectroscopic data and face image data show the superior classification performance of the new soft NCCM over other nearest-class-model methods.

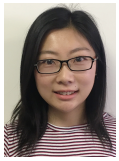
Our future work includes: 1) investigating and unifying more class models, such as the affine hull [31] and hyperdisk [17], [32] class models; 2) unifying the nearest-class-model methods based on the transformation applied on the convex sets; 3) solving the overlapping class model problem for NCHM under the SHC framework; and 4) designing more powerful discriminative ability measures to better visualise the difference between the normal vectors of diffrent methods.

## ACKNOWLEDGEMENTS

We sincerely thank the four anonymous reviewers for their insightful and critical comments, which largely improved the theoretical and empirical presentations of our work, and stimulated our proposal of soft NCCM.

## REFERENCES

[1] S. Wold, "Pattern recognition by means of disjoint principal components models," *Pattern Recognition*, vol. 8, no. 3, pp. 127–139, 1976.
[2] P. Pořízka, J. Klus, A. Hrdlička, J. Vrábel, P. Škarková, D. Prochazka, J. Novotnỳ, K. Novotnỳ, and J. Kaiser, "Impact of laser-induced breakdown spectroscopy data normalization on multivariate classification accuracy," *Journal of Analytical Atomic Spectrometry*, vol. 32, no. 2, pp. 277–288, 2017.
[3] Y. Lee, S.-H. Han, and S.-H. Nam, "Soft independent modeling of class analogy (SIMCA) modeling of laser-induced plasma emission spectra of edible salts for accurate classification," *Applied Spectroscopy*, p. 0003702817697337, 2017.
[4] K. N. Basri, M. N. Hussain, J. Bakar, Z. Sharif, M. F. A. Khir, and A. S. Zoolfakar, "Classification and quantification of palm oil adulteration via portable NIR spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 173, pp. 335–342, 2017.
[5] R. Zhu and J.-H. Xue, "On the orthogonal distance to class subspaces for high-dimensional data classification," *Information Sciences*, vol. 417, pp. 262 – 273, 2017.
[6] R. Zhu, K. Fukui, and J.-H. Xue, "Building a discriminatively ordered subspace on the generating matrix to classify high-dimensional spectral data," *Information Sciences*, vol. 382–383, pp. 1–14, 2017.
[7] O. Yamaguchi, K. Fukui, and K.-i. Maeda, "Face recognition using temporal image sequence," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 318–323.
[8] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *The Eleventh International Symposium on Robotics Research*. Springer, 2005, pp. 192–201.
[9] M. Nishiyama, O. Yamaguchi, and K. Fukui, "Face recognition with the multiple constrained mutual subspace method," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 71–80.
[10] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2164–2177, 2015.
[11] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 684–698, 2005.
[12] Y. Chi, "Nearest subspace classification with missing data," in *Signals, Systems and Computers, 2013 Asilomar Conference on*. IEEE, 2013, pp. 1667–1671.
[13] Y. Chi and F. Porikli, "Connecting the dots in multi-class classification: From nearest subspace to collaborative representation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3602–3609.
[14] C.-P. Chen and C.-S. Chen, "Intrinsic illumination subspace for lighting insensitive face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 422–433, 2012.
[15] G. Nalbantov, P. Groenen, and C. Bioch, "Nearest convex hull classification," Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute, Tech. Rep., 2006.
[16] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2567–2573.
[17] H. Cevikalp, B. Triggs, and R. Polikar, "Nearest hyperdisk methods for high-dimensional classification," in *ICML*. ACM, 2008, pp. 120–127.
[18] X. Zhou and Y. Shi, "Nearest neighbor convex hull classification method for face recognition," in *International Conference on Computational Science*. Springer, 2009, pp. 570–577.
[19] D. Fernández-Francos, Ó. Fontenla-Romero, and A. Alonso-Betanzos, "One-class convex hull-based algorithm for classification in distributed environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.
[20] T. Kobayashi and N. Otsu, "Cone-restricted subspace methods," in *19th International Conference on Pattern Recognition (ICPR 2008)*. IEEE, 2008, pp. 1–4.
[21] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.
[22] H. Cevikalp, "Best fitting hyperplanes for classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
[23] O. L. Mangasarian and E. W. Wild, "Multisurface proximal support vector machine classification via generalized eigenvalues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 69–74, 2006.
[24] 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.
[25] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.
[26] D. Zhou, B. Xiao, H. Zhou, and R. Dai, "Global geometry of SVM classifiers," Technical Report 30-5-02, Institute of Automation, Chinese Academy of Sciences, Tech. Rep., 2002.
[27] J.-J. Moreau, "Décomposition orthogonale dun espace hilbertien selon deux cônes mutuellement polaires," *CR Acad. Sci. Paris*, vol. 255, pp. 238–240, 1962.
[28] F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media, 2006.
[29] T. Arnalds, J. McElhinney, T. Fearn, and G. Downey, "A hierarchical discriminant analysis for species identification in raw meat by visible and near infrared spectroscopy," *Journal of Near Infrared Spectroscopy*, vol. 12, no. 3, pp. 183–188, 2004.
[30] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
[31] H. Cevikalp, B. Triggs, H. S. Yavuz, Y. Küçük, M. Küçük, and A. Barkana, "Large margin classifiers based on affine hulls," *Neurocomputing*, vol. 73, no. 16, pp. 3160–3168, 2010.
[32] H. Cevikalp and B. Triggs, "Hyperdisk based large margin classifier," *Pattern Recognition*, vol. 46, no. 6, pp. 1523–1531, 2013.



**Rui Zhu** received her Ph.D. degree in statistics from University College London in 2017. She is a lecturer in the Faculty of Actuarial Science and Insurance, City, University of London. Her research interests include spectral data analysis, hyperspectral image analysis, subspace-based classification methods and image quality assessment.

**Ziyu Wang** received her Ph.D. degree in the security science and statistical science from University College London in 2017. Her research interests include hyperspectral image analysis, sparse representation and statistical classification.

**Naoya Sogi** received his B.E. and M.E. degrees from the University of Tsukuba in 2017 and 2019, respectively. He is currently a Ph.D. candidate at the University of Tsukuba. His interests include the theories of computer vision, pattern recognition, machine learning and applications of these theories.

**Kazuhiro Fukui** received his Ph.D. degree from Tokyo Institute of Technology in 2003. He is a professor in the Department of Computer Science, University of Tsukuba. His interests include the theories of computer vision, pattern recognition, and applications of these theories. He has been serving as a program committee member at many pattern recognition and computer vision conferences, including as an Area Chair of ICPR'12, 14 and 16.

**Jing-Hao Xue** received his Dr.Eng. degree from Tsinghua University in 1998 and Ph.D. degree from the University of Glasgow in 2008. He is an associate professor in the Department of Statistical Science, University College London. His research interests are statistical machine learning, high-dimensional statistics, pattern recognition and image analysis.