Adaptive and Asymptotic Mean-based Subclass Discriminant Analysis

Yuzhe Feng[1], Yunlong Gao[1*] and Feiping Nie[2]

[1] Pen-Tung Sah Institute of Micro-Nano Science and Technology,Xiamen University,China
[2] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN),Northwest Polytechnical University,China
fengyuzhe1@huawei.com, gaoyl@xmu.edu.cn, feipingnie@gmail.com

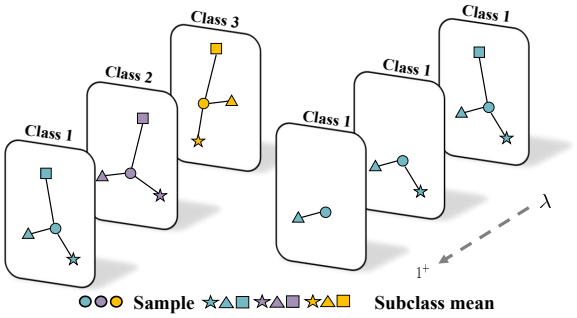
Abstract

Traditional Discriminant analysis (DA) is one of the classical supervised learning algorithms to reduce the dimensionality of data with Gaussian assumption. Since the unique class mean in traditional DA is intractable to estimate the non-Gaussian distrbution of data, some existing DA algorithms based on the clustering criterion focus on learning multiple means in each class so as to address the non-Gaussian issue. The clustering-based DA inevitably involved the constraint optimization problem to learn multiple means, which may lead to the locally optimal solution. To address these issues, inspired by the smooth approximation theory and the concept of Kolmogorov mean, this paper explores an unconstraint function with asymptotic property as an alternative proxy to clustering-based DA algorithms. Thus the derived DA algorithm, i.e., adaptive and asymptotic mean-based subclass discriminant analysis (AASDA), which not only leverages multiple means to represent different subclasses in same class but also adaptively and asymptotically learns the similar mean for each sample in the learned optimal subspace via the gradient-based optimizer. The asymptotic analysis of unconstraint function, the gradient analysis and convergence guarantee of proposed criterion verify the effectiveness of AASDA algorithm. Its merits are thoroughly assessed on a suite of synthetic and real world data experiments.

Introduction

Discriminant analysis (DA) is a fundamental task in supervised learning for minimizing the compactness of data within each class in the learned subspace based on some distance measure (Zhang et al. 2025; Yan, Zhang, and Mai 2025; Moretti, Pellizzoni, and Silvestri 2025; Kalavas, Kipouridis, and Varma 2025). Perhaps the most popular DA criterion is the Fisher criterion, which aims to minimize distances between samples and the corresponding unique class mean in the learned subspace. However, the unique class mean implicitly assumes that samples of each class obey Gaussian distribution. In real world scenarios, samples of each class may be non-Gaussian distribution, which means that a naive distribution assumption will lead to inaccurate estimation of covariance in Fisher criterion.

*Corresponding author



Each sample connects all subclass means in bipartite graph of each class.

Each sample only connects with neighbor subclass means in bipartite graph of same class asymptotically.

Figure 1: Illustration of strategies to learn neighbor subclass means for each sample in bipartite graph.

Recently, researchers have recognized this essential limitation of Fisher criterion, a series of learning techniques have been applied to deal with the non-Gaussian issue in Fisher criterion. One of the representative learning techniques is graph embedding (Yan et al. 2007; Nie, Wang, and Huang 2014; Zhou et al. 2024; Zhu et al. 2023; Peterfreund et al. 2025), many graph embedding-based DA criterions are dedicated to learning the relations among pairwise samples of each class in the optimal subspace iteratively, so as to eliminate the unique class mean in Fisher criterion and preserve the relations of data in the optimal subspace (Li et al. 2022; Chang et al. 2022a; Nie et al. 2020; Pang, Zhou, and Nie 2019). However, it is computationally demanding to learn the relations of pairwise samples. Inspired by the equivalence between unsupervised Fisher criterion and clustering criterion (Nie et al. 2023a; Ding and Li 2007; Wang et al. 2023), the clustering-based DA algorithms have been proposed (Zhao et al. 2023; Nie et al. 2023c; Wang et al. 2024b), which utilize multiple mean vectors to represent different subclasses of same class and learn the relations between each sample and multiple mean vectors in the learned subspace. Since the number of mean vectors is far less than the number of samples of each class, the clustering-based DA algorithms are suitable to deal with large-scale problems.

Specifically, the clustering-based DA algorithms share an unified learning paradigm, which involves graph learning

and graph embedding. The recent researchers (Chen et al. 2025; Zhao et al. 2024; Nie et al. 2023b; Carrasco and Sun 2025) prefer to design different constraint probabilistic weighting schemes in graph learning, such that the distance among each sample vector and dissimilar mean vectors to be larger than that of same sample vector and similar mean vectors, where the mean vectors are adaptively estimated in the optimal subspace. The drawback of constraint probabilistic weighting schemes lies in the introduction of weights and the pre-defined constraint on weights among each sample and multiple means, which reformulates the graph learning as the constraint optimization problem. Such constraint optimization problem is actually a non-convex optimization problem w.r.t. the weight for each sample (Wang et al. 2024a; Nie et al. 2022). Thus, it is easily converging to local solutions and intractable to learn the weight vector of each sample in an unconstraint and adaptive way.

In order to investigate the clustering-based DA algorithms from the fundamental principle of distance, and improve the locality of data while maintaining a tractable time complexity at the same time. Inspired by the smooth approximation theory, we propose to design an unconstraint proxy and learn the most similar means by utilizing the asymptotic property-based Kolmogorov mean. The intuitive illustrations of existing and the proposed learning strategies for subclass means of each sample are in Fig. 1. Specifically, the clustering-based DA criterion is reformulated as the non-smooth optimization criterion, and the concept of Kolmogorov mean is introduced as a proxy criterion of the non-smooth optimization criterion with theoretical approximation guarantee. Considering a special case of the proposed proxy criterion, we further prove the asymptotic property of proposed proxy criterion, which can learn the locality of mean vectors of each sample in optimal subspace in an adaptive and unconstraint way.

According to the findings, we propose a novel unconstraint clustering-based DA algorithm called adaptive and asymptotic mean-based subclass discriminant analysis (AASDA), where the optimization criterion of AASDA is equivalent to the unconstraint clustering criterion of samples of each class in the optimal subspace. The asymptotic analysis of unconstraint function, the gradient analysis and convergence guarantee of proposed criterion verify the effectiveness of AASDA algorithm. The main contributions of this paper are summarized as follows:

- In this paper, a novel unconstraint proxy criterion is proposed for clustering-based DA algorithm. In order to derive such unconstraint proxy, the concept of Kolmogorov mean is introduced as a proxy with theoretical guarantee.
- Considering a special case of proposed unconstraint proxy, we further prove the asymptotic property of the special proxy. Thus, a novel clustering-based DA algorithm AASDA is proposed, which can not only learn similar means of each sample in an unconstraint and asymptotic way but also learn an optimal subspace via the gradient-based optimizer.
- The asymptotic analysis of proposed proxy, the gradient analysis of proposed clustering-based DA criterion and the convergence of AASDA are provided.
- An extensive empirical study on real world and synthetic data, we demonstrate the efficacy and superiority of AASDA against state-of-the-art DA algorithms over multiple downstream tasks.

The proofs of all involved notations, lemmas and theorems together with supplementary experiments are relegated to **Appendix** [†].

Related Works

Linear Discriminant Analysis

Given the training data matrix $\boldsymbol{X} = \{\boldsymbol{X}^1, ..., \boldsymbol{X}^c\} \in \mathbb{R}^{d\times n}$ with c classes and $n = \sum_{i=1}^{c} n_i$, where $\boldsymbol{X}^i \in \mathbb{R}^{d\times n_i}$ denotes the data matrix in class i, $\boldsymbol{x}_j$ and n_i denote the j-th sample vector in training dataset and the number of samples in i-th class, respectively. The purpose of linear discriminant analysis (i.e., LDA) aims to learn a transformation matrix $\boldsymbol{W} \in \mathbb{R}^{d\times r} (r \leqslant d)$ so as to project samples of each class into a low-dimensional representation with discriminative objective, and thus Fisher defines the within-class covariance matrix $\boldsymbol{S_w}$ and total covariance matrix $\boldsymbol{S_t}$ as follows:

$$\boldsymbol{S_w} = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^i - \bar{\boldsymbol{x}}^i)(\boldsymbol{x}_j^i - \bar{\boldsymbol{x}}^i)^\top \tag{1}$$

$$\boldsymbol{S_t} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \tag{2}$$

where $\boldsymbol{x}_j^i$ is the j-th sample vector in class i, $\bar{\boldsymbol{x}}^i$ is the arithmetic mean vector of all samples in i-th class and $\bar{\boldsymbol{x}}$ is the arithmetic mean vector of all samples in the training set. Since the definition of arithmetic mean vector $\bar{\boldsymbol{x}}$, the total covariance matrix $\boldsymbol{S_t}$ is equivalent to the matrix form $\boldsymbol{S_t} = \boldsymbol{XHX}^\top$, where $\boldsymbol{H} = \boldsymbol{I}_n - (\boldsymbol{1}_n\boldsymbol{1}_n^\top)/n$ is the centering matrix.

In order to minimize the distances of samples of each class in the optimal subspace, the criterion of LDA can be formulated as the uncorrelated constraint-based within-class distance criterion, which is formulated as the following otimization problem:

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{S_w} \boldsymbol{W}) \\ \text{s.t.} \quad & \boldsymbol{W}^\top \boldsymbol{S_t} \boldsymbol{W} = \boldsymbol{I}_r \end{aligned} \tag{3}$$

where the uncorrelated constraint in (3) ensures uncorrelated property of subspace samples, and thus minimize distances among subspace samples of same class in (3) is equivalent to maximize distances among subspace samples from different classes (Nie et al. 2012).

Clustering-based Discriminant Analysis

It is worth noting that the trace function-based criterion in (3) is limited to consider the matrix-based criterion, according to the property of trace function, the criterion in (3) is

[†]Appendix will be uploaded in arXiv

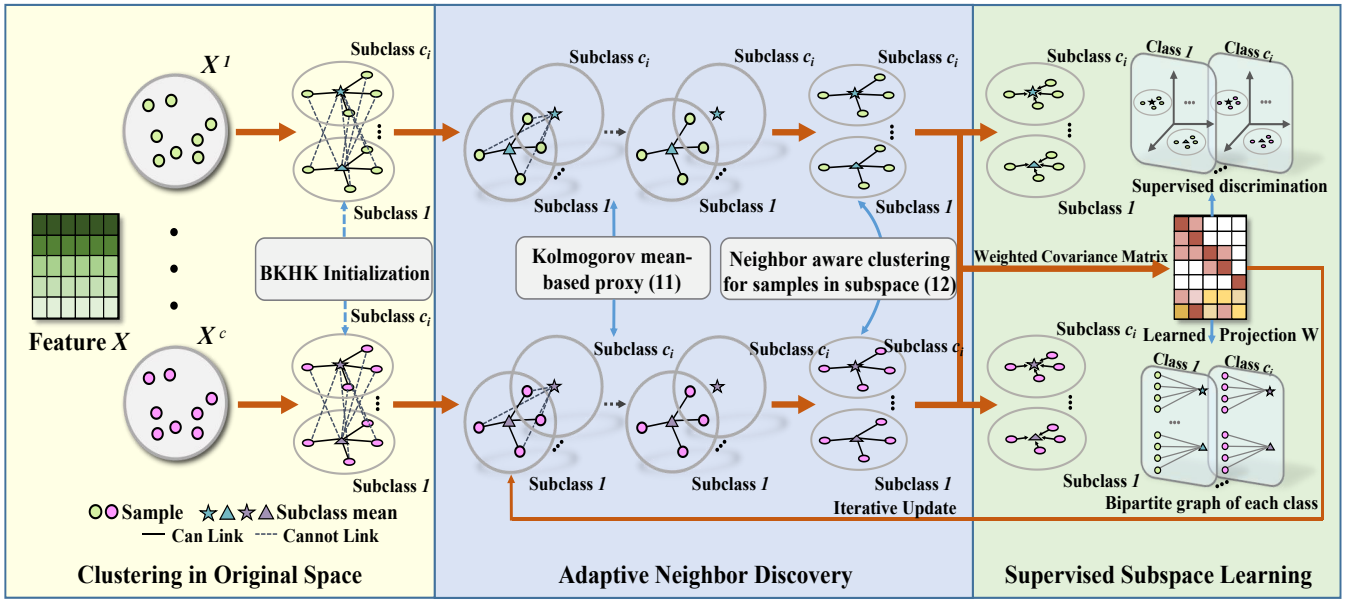


Figure 2: Illustration of our proposed algorithm AASDA.

equivalent to the following vector-based distance criterion

$$\min_{\boldsymbol{W}} \quad \sum_{i=1}^{c}\sum_{j=1}^{n_i}\left\|\boldsymbol{W}^\top \boldsymbol{x}_j^i - \boldsymbol{W}^\top \bar{\boldsymbol{x}}^i\right\|_2^2 \qquad (4)$$
$$\text{s.t.} \quad \boldsymbol{W}^\top \boldsymbol{S_t} \boldsymbol{W} = \boldsymbol{I}_r$$

However, the unique class mean in within-class covariance matrix $\boldsymbol{S_w}$ implicitly assumes that samples of each class obey Gaussian distribution. When the distribution of samples in each class is more complex than Gaussian, the pre-computed unique class mean is intractable to learn the mean and covariance statistics of samples with non-Gaussian distribution (Li et al. 2022).

Inspired by the equivalence (Ding and Li 2007) between within-class distance criterion in (4) and clustering criterion, an unified criterion of within-class distance criterion and clustering criterion has been proposed in terms of the non-Gaussian distribution scenario (Chen et al. 2025; Wang et al. 2024b,a; Zhao et al. 2024). Specifically, the clustering-based within-class distance criterion utilized multiple learnable means in each class to represent the mean statistics of different subclasses in same class (Wang et al. 2024b), which is formulated as the following optimization criterion

$$\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{M}} \quad \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{c_i}(u_{jk}^i)^\lambda \left\|\boldsymbol{W}^\top \boldsymbol{x}_j^i - \boldsymbol{W}^\top \boldsymbol{m}_k^i\right\|_2^2 \qquad (5)$$
$$\text{s.t.} \quad {\boldsymbol{u}_j^i}^\top \mathbf{1}_{c_i} = 1, 0 \le u_{jk}^i, \boldsymbol{W}^\top \boldsymbol{S_t} \boldsymbol{W} = \boldsymbol{I}_r$$

where $\boldsymbol{M}$ is a cell array with the mean matrix of i-th class $\boldsymbol{M}^i = \{\boldsymbol{m}_1^i, ..., \boldsymbol{m}_{c_i}^i\} \in \mathbb{R}^{d\times c_i}$ being the i-th column of it, $\boldsymbol{m}_k^i$ and c_i denotes the k-th mean vector and number of means of i-th class, respectively. In addition, $\boldsymbol{U}$ is also a cell array with $\boldsymbol{U}^i \in \mathbb{R}^{n_i\times c_i}$ being the i-th column of it and λ is a hyperparameter.

In order to learn the relations between j-th sample and all mean vectors in i-th class, the weight vector $\boldsymbol{u}_j^i \in \mathbb{R}^{c_i\times 1}$ is introduced as the optimization variable and the probabilistic simplex constraint (i.e., ${\boldsymbol{u}_j^i}^\top \mathbf{1}_{c_i} = 1, 0 \le u_{jk}^i$) is imposed on the weight vector $\boldsymbol{u}_j^i$ w.r.t. each sample $\boldsymbol{x}_j^i$, such that the optimization problem (5) w.r.t. the weight vector $\boldsymbol{u}_j^i$ is formulated as the constraint optimization problem. The solution of optimization problem (5) w.r.t. the weight vector $\boldsymbol{u}_j^i$ can learn the probabilistic weight based on the distance among each sample and other means of same class.

Compared to the criterion in (4), the clustering-based DA in (5) utilize multiple learnable means in each class, which is learned based on the weighted sum of samples of each class. Thus, minimizing the objective function value of problem (5) can effectively cluster samples of each class into subclasses so as to learn a weighted covariance matrix, and then learn a discriminative subspace based on the learned weighted covariance matrix.

A shortcoming shared by existing clustering-based DA algorithms is that they all introduce the weight vector as an optimization variable and impose predefined constraints on the weight vector, so as to learn the relations among each sample and multiple means of same class. Such constraint optimization problem is a non-convex optimization problem, which is prone to obtain local optimal solution (Kalavas, Kipouridis, and Varma 2025).

Method and Discussions

Problem Formulation

Since the clustering-based DA is essentially the distance-based DA criterion, it is intuitive to rethink the clustering-based DA from a distance perspective. Given that $\boldsymbol{M}^i = \{\boldsymbol{m}_1^i, ..., \boldsymbol{m}_{c_i}^i\}$ denotes a set of mean vectors in the subset

D^i of a finite dimensional Euclidean real space $\mathbb{R}^d$, where a set of means of i-th class are initialized by specific clustering strategy. The objective of clustering criterion in DA is the assignment of a set of n_i samples of i-th class into c_i subclasses in the subspace, where each subclass is represented by its mean vector. Thus, it is intuitive to consider minimizing the distance from each sample to the nearest mean vector of same class in the subspace. Suppose the distance function on two vectors $d_{\boldsymbol{W}}(\cdot,\cdot)$ parameterized with transformation matrix $\boldsymbol{W}$, the clustering-based DA is reformulated as following optimization problem

$$\min_{\forall\, i,\boldsymbol{m}_1^i,\cdots,\boldsymbol{m}_{c_i}^i\in D^i} \sum_{i=1}^{c}\sum_{j=1}^{n_i} \min_{1\leqslant k\leqslant c_i} d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i) \tag{6}$$

However, the optimization problem (6) typically requires to invoke sorting algorithm to search the nearest neighbor mean for each sample, which is intractable to extend to an end-to-end training manner and preserve the neigbors of each sample adaptively. Inspired by the smooth approximation theory (Rockafellar 1970; Ben-Tal and Teboulle 1986; Teboulle 2007), the concept of Kolmogorov mean defined below, was introduced in 1930 by Kolmogorov (Kolmogorov 1930) as a natural generalization of all the well known concepts of mean.

Definition 1. *(Kolmogorov 1930) Let $g:\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ be a strictly increasing and convex function, and let g^{-1} be its inverse function, which is thus strictly increasing and concave. The Kolmogorov mean of c_i real numbers $d_1^i,\cdots,d_{c_i}^i$ associated to g is defined by*

$$H_g(\boldsymbol{d}^i)=g^{-1}\left(\frac{1}{c_i}\sum_{k=1}^{c_i} g(d_k^i)\right) \tag{7}$$

where the vector $\boldsymbol{d}^i=(d_1^i,\cdots,d_{c_i}^i)$.

The Lemma 1 shows that the Kolmogorov mean provides a lower bound for the non-smooth maximum function $\max\limits_{1\leqslant k\leqslant c_i} d_k^i$.

Lemma 1. *For each $\boldsymbol{d}^i\in\mathbb{R}^{c_i}$, the following inequalities hold*

$$\sum_{k=1}^{c_i}\frac{1}{c_i}d_k^i \leqslant H_g(\boldsymbol{d}^i)\leqslant \max_{1\leqslant k\leqslant c_i} d_k^i \tag{8}$$

Recall the reformulated clustering-based DA (6) consists of minimizing the value of non-smooth function f, where the non-smooth function f can be rewritten as:

$$\begin{aligned} f(\boldsymbol{m}^i)&=\sum_{j=1}^{n_i}\min_{1\leqslant k\leqslant c_i} d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)\\ &=-\sum_{j=1}^{n_i}\max_{1\leqslant k\leqslant c_i} -d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)\end{aligned} \tag{9}$$

where the notation $\boldsymbol{m}^i=(\boldsymbol{m}_1^i,\cdots,\boldsymbol{m}_{c_i}^i)$ denotes the $p^i:=c_i\times d$ dimensional vector $\boldsymbol{m}^i\in P^i$ and $P^i\subseteq\mathbb{R}^{p^i}$ denotes the p^i-fold Cartesian product of D^i.

Since the non-negativity of $d_{\boldsymbol{W}}(\cdot,\cdot)$, in order to approximate $\max\limits_{1\leqslant k\leqslant c_i} -d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)$, we need only to consider Kolmogorov mean in Definition 1 with domain belongs to $\mathbb{R}^d$. Thus one has to approximate the non-smooth function f by the smooth proxy:

$$\hat{f}(\boldsymbol{m}^i)=-\sum_{j=1}^{n_i} g^{-1}\left(\frac{1}{c_i}\sum_{k=1}^{c_i} g(-d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i))\right) \tag{10}$$

Since the smooth proxy is designed for the clustering-based DA, and the Definition 1 shows that any increasing and convex functions can be the choice of function g. In order to design the function g, it is intuitive to consider the relations between proposed proxy and existing clustering-based DA criterion. Thus, we next convert the existing clustering-based DA criterion in (5) into an equivalent function with the Lemma 2, see the Appendix for its detailed proof.

Lemma 2. *Denote $\{d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)|d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)=\|\boldsymbol{W}^\top\boldsymbol{x}_j^i-\boldsymbol{W}^\top\boldsymbol{m}_k^i\|_2^2, k\in\{1,\cdots,c_i\}\}$ is a given set of c_i real numbers. For the samples in i-th class, the optimization criterion in (5) is equivalent to*

$$\sum_{j=1}^{n_i}\left(\sum_{k=1}^{c_i} d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)^{\frac{1}{1-\lambda}}\right)^{1-\lambda} \tag{11}$$

According to the Lemma 2, it is trivial to verify that $g(t)=(-t)^{\frac{1}{1-\lambda}}$ is a strictly increasing and convex function when $t>0$ and $\lambda>1$. Thus, with the choice $g(t)=(-t)^{\frac{1}{1-\lambda}}$ in the smooth proxy $\hat{f}$ as given in (10), one obtains a special case of the smooth proxy $\hat{f}$. According to the previous analyses, we approximate the non-smooth function in (6) by the special case of the smooth proxy $\hat{f}$ and eliminate the constant $\frac{1}{c_i}$:

$$\begin{aligned}\min_{\boldsymbol{M}}\quad & \sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(\sum_{k=1}^{c_i} d_{\boldsymbol{W}}(\boldsymbol{x}_j^i,\boldsymbol{m}_k^i)^{\frac{1}{1-\lambda}}\right)^{1-\lambda}\\ \text{s.t.}\quad & \forall\, i,\boldsymbol{m}_1^i,\cdots,\boldsymbol{m}_{c_i}^i\in D^i\end{aligned} \tag{12}$$

where $\boldsymbol{M}$ is a cell array with the same definition in related work and λ is an asymptotic parameter that satisfies $\lambda>1$. Compared with the clustering-based DA criterion in (5), it is worth noting that the optimization problem (12) eliminates the optimization variable w.r.t. weight vector, such that the learning task for the set of mean vectors is an unconstraint optimization problem. Thus, it avoids the local optimal solution to some extent, which will be revealed in experiment section.

In order to further achieve the purpose of clustering-based DA, which aims to learn a discriminative subspace that is determined by the projection matrix $\boldsymbol{W}\in\mathbb{R}^{d\times r}$, and the distance function $d_{\boldsymbol{W}}(\cdot,\cdot)$ denotes the usual squared Euclidean distance. The optimization problem (12) can be further formulated as follows

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{M}}\quad & \sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(\sum_{k=1}^{c_i}(\|\boldsymbol{W}^\top\boldsymbol{x}_j^i-\boldsymbol{W}^\top\boldsymbol{m}_k^i\|_2^2)^{\frac{1}{1-\lambda}}\right)^{1-\lambda}\\ \text{s.t.}\quad & \forall\, i,\boldsymbol{m}_1^i,\cdots,\boldsymbol{m}_{c_i}^i\in D^i,\boldsymbol{W}^\top\boldsymbol{S_t}\boldsymbol{W}=\boldsymbol{I}_r\end{aligned} \tag{13}$$

Different from the clustering-based DA (13), the proposed unconstraint proxy-based criterion can not only learn multiple means of each class via the gradient-based optimizer, so as to estimate the mean statistics of different subclasses of same class in subspace in an unconstraint way, but also learn the similar means of each sample in subspace in an asymptotic way, which will be revealed in Property 1. According to the learned multiple mean vectors of each class, the weighted covariance matrix can be estimated so as to learn the optimal subspace, such that samples of each class can be clustered into the corresponding subclasses in the learned optimal subspace and preserve the neighbor relations of each sample simultaneously. The corresponding optimization algorithm for optimizing criterion (13) is relegated to **Appendix**.

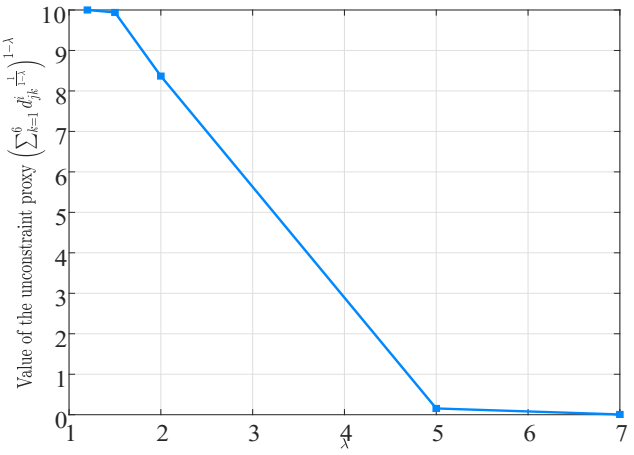


Figure 3: The illustration of the relation between the unconstraint proxy $\left(\sum_{k=1}^{6} {d_{jk}^i}^{\frac{1}{1-\lambda}}\right)^{1-\lambda}$ and λ, where $d_{jk}^i \in \{10, 110, 210, 310, 410, 500\}$ and $\lambda \to 1^+$. It is worth to note that the minimum value of elements in $\boldsymbol{d}_j^i$ is 10.

Theoretical Properties of Proxy

In order to further reveal the asymptotic property in proposed unconstraint proxy (13), the following Property 1 is presented.

Property 1. (Asymptotic Property) Denote $(i, j) \in \{1, \cdots, c\} \times \{1, \cdots, n_i\}$. Without loss of generality, given a sequence of nonnegative real numbers $\{d_{\boldsymbol{W}}(\boldsymbol{x}_j^i, \boldsymbol{m}_k^i)\}_{k=1}^{c_i}$, there exists the following relationship:

$$\min_k d_{\boldsymbol{W}}(\boldsymbol{x}_j^i, \boldsymbol{m}_k^i) = \lim_{\lambda \to 1^+} \left(\sum_{k=1}^{c_i} d_{\boldsymbol{W}}(\boldsymbol{x}_j^i, \boldsymbol{m}_k^i)^{\frac{1}{1-\lambda}} \right)^{1-\lambda} \tag{14}$$

In order to further analyze the convexity of proposed unconstraint proxy (13), we first consider a simplified form of unconstraint proxy as the Kolmogorov mean when $g(\cdot) = (\cdot)^{\frac{1}{1-\lambda}} (\lambda > 1)$, where it is trivial to verify $g(\cdot)$ is an increasing and convex function on set Ψ.

Property 2. (Convexity) The unconstraint proxy (11) is convex on Ψ^{c_i}, where Ψ^{c_i} denotes the c_i-fold Cartesian product of set Ψ and Ψ denotes the feasible set of function $g(\cdot) = (\cdot)^{\frac{1}{1-\lambda}} (\lambda > 1)$.

Connection to k-means Clustering

Proposition 1. *When $\lambda \to 1^+$, optimizing the criterion described in (13) is equivalent to perform k-means clustering for samples of each class in projection space.*

Convergence Analysis

Proposition 2. *The value of objective function of optimization problem (13) will monotonically decrease with respect to each optimization variable in each iteration until convergence.*

Time Complexity Analysis

Proposition 3. The time complexity of Alg. 1 in Appendix is $\mathcal{O}(nT)$ (see **Appendix** for more details), where T is the number of iterations in Alg. 1 in Appendix.

Experiments

To further validate the effectiveness of AASDA, we have designed the experimental section intended to answer the following evaluation questions (EQs):

1. **EQ1** Does AASDA achieve superior performance compared to its competitors for different supervised downstream tasks?
2. **RQ2** Does the proposed unconstraint proxy in AASDA learn the neighbor mean for each sample in subspace, and does it promote learning the locality of data and finding a better solution?

AASDA has performed the complete parameter sensitivity analysis, convergence curves, visualization and initialization in the **Appendix**.

Datasets, Compared Methods, and Evaluation Metric

Datasets We adopt 9 real-world benchmark datasets widely used in different downstream tasks to evaluate the performance of AASDA compared to state-of-the-art baselines. Table 1 in **Appendix** illustrates the details of all 9 datasets. In case of supervised learning, we randomly select 70% samples from each class in same dataset for training set, and the rest of samples in same class for testing set.

Compared Methods We compare the proposed method with classical and state-of-the-art baselines, including classic LDA (Belhumeur, Hespanha, and Kriegman 1997), graph embedding-based DA LADA (Li et al. 2022) and DMEG (Wang et al. 2021), distance metric-based DA SALDA (Chang et al. 2022b) and $\ell_{2,1}$-LDA (Nie et al. 2021), clustering-based DA LDC (Nie et al. 2023b), LAFLDA (Wang et al. 2024a), ELCS (Wang et al. 2024b) and FLDA-W (Chen et al. 2025). The parameter settings of all compared methods tune best parameters according to corresponding literatures.

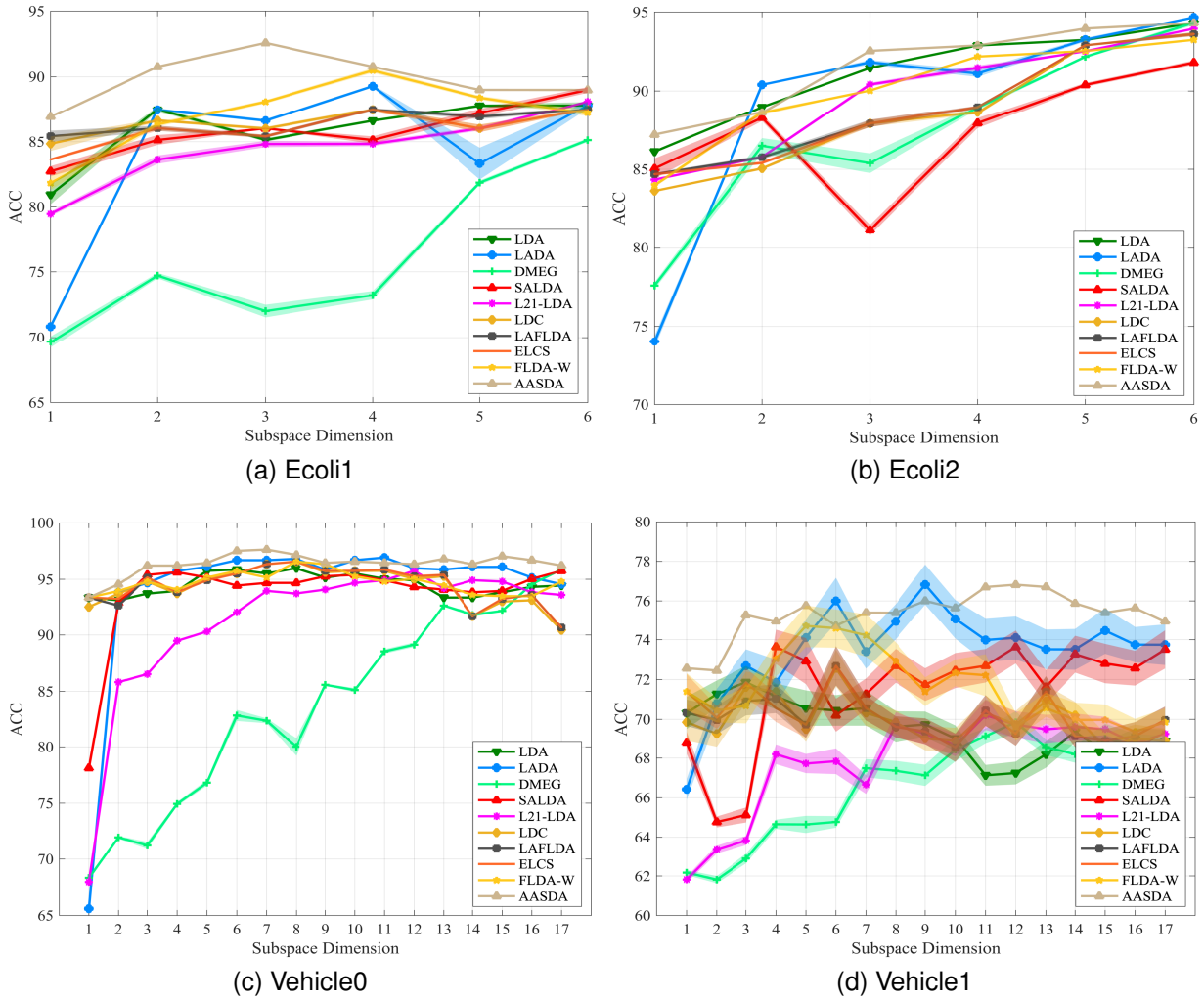


Figure 4: ACC curves with error over subspace dimensions on all test imbalanced datasets for all compared methods.

Evaluation Metrics To estimate discriminative ability effectively, four well-known classification evaluation metrics are applied to our experiments, including Accuracy (ACC), Precision (PRE), Recall (REC), F1-Score and G-Mean (He and Garcia 2009). A higher value of each metric indicates better performance of DA methods for the classification task. Average minimum pairwise trace ratio (AMPTR) metric indicates the separability of samples in pairwise classes.

Experimental Setups

Experimental rounds of the proposed method is under same dataset and the average value is taken by repeating the run ten times. The reduced dimension of all DA methods is set as 70% of original dimension of dataset in default. The parameter setting of proposed method tunes the number of subclass means c_i in each class and the asymptotic parameter λ within the value of $\{1, \cdots, n_i/3c\}$ and $\{1, 10, \cdots, 10^4\}$, respectively.

Comparison to SOTA (EQ1)

Imbalanced Classification Since many real-world scenarios are often confronted with class imbalance issue, the distribution of each class is significantly different from other classes. Thus, it raises the difficulty to learn the distribution of each class with Gaussian distribution assumption. It is necessary to learn multiple subclass means of each class that are more favorable to non-Gaussian distribution of each class. To evaluate the performance of proposed AASDA on this kind of data, we present the overall ACC curve results of all DA methods in classification, as shown in Fig. 4, from which we observe that AASDA nearly outperforms other methods across all dimensions on class imbalanced dataset. To further evaluate classification performance, four classification metrics via the same nearest neighbor classifier were visualized using radar charts in Fig. 5. Although some DA methods, such as FLDA-W and ELCS, occasionally outperform AASDA in specific metrics, it demonstrates more stable performance accross all metrics. Finally, AASDA's

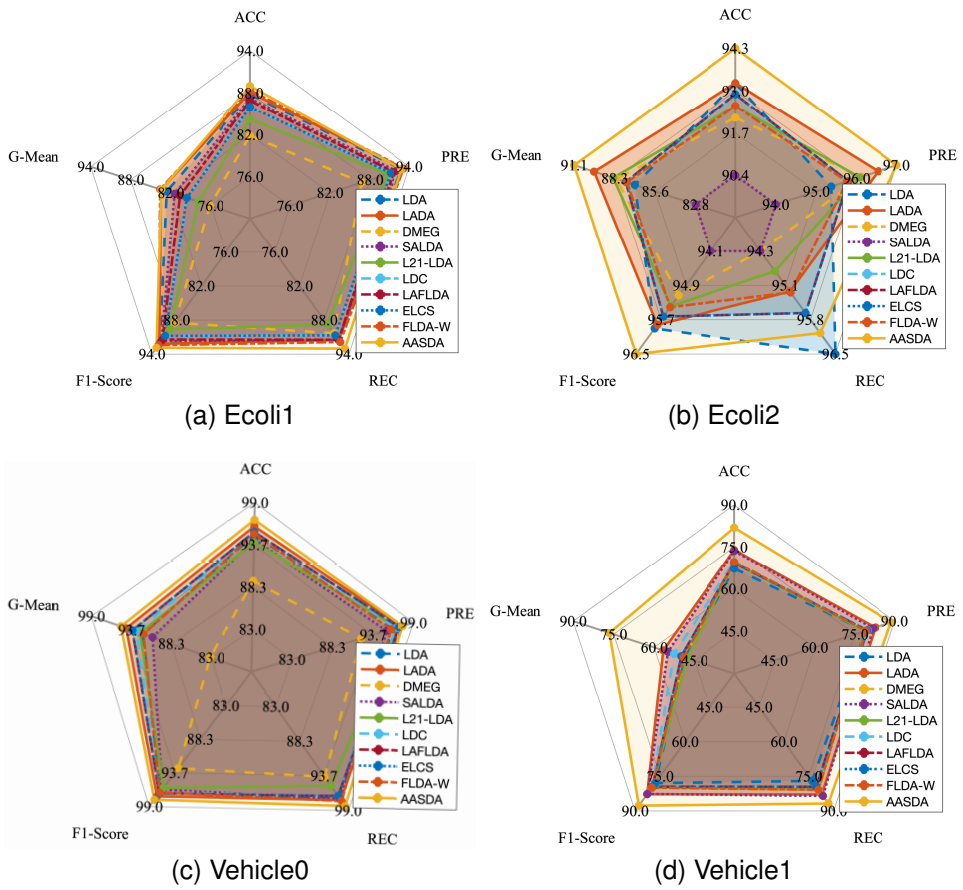


Figure 5: Comparison of ACC, PRE, REC, F1-Score and G-Mean using nearest neighbor classifier.

effective balance across different metrics with outstanding performance. This may be attributed to the effective enhancement of covariance matrix estimation through clustering criterion in each class, such as the class imbalance improvement by 1.5% seen with Vehicle1.

Large-Scale Classification Meanwhile, we further conduct classification on the real-world large-scale dataset and illsutrate the ACC, Time (in minutes of per-iteration) and SUR (the ratio of time of each comparative method and time of LDA) in Table 1 and followed with statistical test. It is evident that AASDA surpasses other DA methods in most cases, particularly under large-scale scenarios.

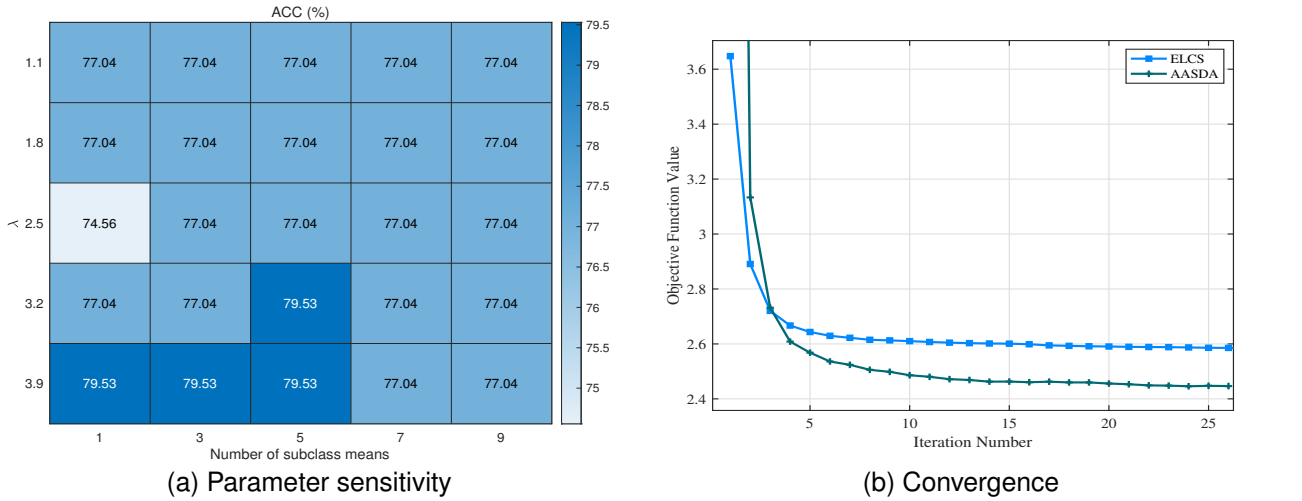
Ablation Study (EQ2)

To verify that the clustering criterion in each class and the Kolmogorov mean-based loss contributes to preserve the locality of data and improve local optimal solution for the clustering-based DA, we conduct ablation experiments. First, when we comprehensively examine the proposed AASDA w.r.t. the number of subclass means in each class on the classification task and the AMPTR metric under same parameter λ in Fig. 7 (a) and Fig. 7 (b), respectively (all in **Appendix** Fig. 5 and Fig. 6). We can excitedly discover that the utilization of multiple learnable subclass means in each class promotes performance and effective separation of samples in pairwise classes. For example, after increasing the number of subclass means in each class (more than one), the ACC of Vehicle1 improved from 70 % to 74 % and maintained steady, with increased inter-class separability. This improvement can be attributed to the learning of distribution of each subclass via multiple learnable subclass means. Meanwhile, the initialization has little effect on the ACC of embedding learned by AASDA in **Appendix** Fig. 7.

Furthermore, Fig. 8 in **Appendix** reveals the value of gradient of criterion (13) among each sample and its subclass means (denoted by sample index and subclass index) learned by AASDA on Vehicle1 when tuning the value of parameter λ to 1^+ asymptotically. It is evident that the number of **nonzero** value of gradients for each sample will asymptotically decrease to one, which empirically verifies the Asymptotic Property, as depicted in **Theoretical Properties**.

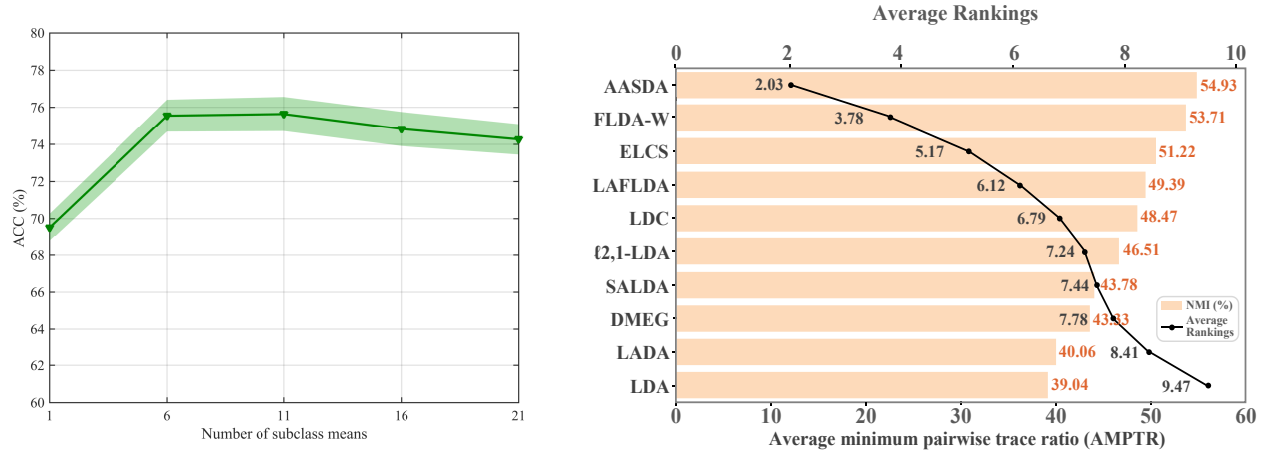
Sensitivity Analysis (EQ2)

First, Fig. 6 (a) (all in **Appendix** Fig. 3) illustrates the parameter sensitivity of AASDA on Vehicle1 in terms of λ and c_i of criterion (13). The ACC of AASDA is generally robust in most cases. Second, Fig. 6 (b) (all in **Appendix**



(a) Parameter sensitivity (b) Convergence

Figure 6: Sensitivity analysis of AASDA on Vehicle1.



(a) ACC (b) AMPTR

Figure 7: Ablation study of learnable subclass means of each class on imbalanced datasets.

Fig. 4) progressively highlights the convergence curves of Kolmogorov mean-based loss (13) and ELCS criterion (5) on Vehicle1. From this, the Kolmogorov mean-based unconstraint proxy can promote the clustering-based DA for finding a better solution under the same initialization, since it eliminates the constraint optimization problem and reduces the number of optimization variales.

Conclusion and Future Work

Traditional clustering-based DA learning paradigms are limited by the constraint optimization technique, impeding their ability to not only effectively learn the subclass means of each class in an unconstraint way but also find a better solution. In this study, drawing inspiration from Kolmogorov mean, we proposed an unconstraint proxy for the clustering-based DA criterion and derived a novel algorithm named AASDA. Therein, the underlying clustering formulation of each class in DA remains the same, but the unconstraint proxy learns neighbor subclass means of each sample asymptotically via the gradient-based optimizer, so as to learn the subclass means in an unconstraint way. Theoretical analyses and experimental evaluations demonstrate the effectiveness and superiority of AASDA. In high-dimensional scenario, the proposed algorithm involved eigen value decomposition with high time complexity. In future work, we will explore the low-rank matrix approximation algorithm to reduce its time complexity w.r.t. the dimension of each sample.

Method	10X PBMC								
	@5000			@8000			@10000		
	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓
LDA	72.03⋆	5.32	†	73.04⋆	6.14	†	73.98⋆	6.31	†
LADA	71.39⋆	5.97	1.12	72.82⋆	6.09	0.99	74.53⋆	6.67	1.06
DMEG	71.94⋆	6.00	1.13	73.10⋆	6.28	1.02	74.84⋆	6.87	1.09
$\ell_{2,1}$-LDA	72.30⋆	4.52	0.85	73.42⋆	5.59	0.92	75.03⋆	5.17	0.82
SALDA	72.56⋆	4.41	0.83	73.50⋆	5.52	0.90	75.20⋆	5.12	0.81
LDC	74.28⋆	3.79	0.71	75.31⋆	3.83	0.62	75.80⋆	3.94	0.62
ELCS	74.08⋆	3.73	0.70	75.01⋆	3.77	0.61	75.65⋆	3.89	0.62
LAFLDA	74.04⋆	3.82	0.72	74.96⋆	3.85	0.63	75.72⋆	3.96	0.63
FLDA-W	74.21⋆	3.75	0.70	75.29	3.80	0.62	75.54⋆	3.87	0.61
AASDA	**74.82**	**3.68**	**0.69**	**75.58**	**3.73**	**0.61**	**76.59**	**3.80**	**0.60**
Method	**Worm neuron cells**								
	@5000			@8000			@10000		
	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓
LDA	60.41⋆	7.31	†	61.54⋆	8.20	†	62.47⋆	9.07	†
LADA	61.44⋆	7.87	1.08	61.93⋆	8.91	1.09	63.73⋆	9.78	1.08
DMEG	62.17⋆	7.95	1.09	62.52⋆	9.02	1.10	63.89⋆	9.83	1.08
$\ell_{2,1}$-LDA	62.84⋆	6.14	0.84	63.90⋆	6.72	0.82	64.04⋆	7.17	0.79
SALDA	62.80⋆	6.18	0.85	63.78⋆	6.63	0.81	64.01⋆	7.09	0.78
LDC	64.73⋆	5.16	0.71	64.92⋆	5.58	0.68	65.02⋆	5.97	0.66
ELCS	64.65⋆	4.93	0.67	64.80⋆	5.20	0.63	64.93⋆	5.37	0.59
LAFLDA	64.49	4.89	0.67	64.85	5.17	0.63	64.90	5.42	0.60
FLDA-W	64.72	4.95	0.68	65.07	5.13	0.63	65.02	5.29	0.58
AASDA	**64.98**	**4.83**	**0.66**	**65.21**	**5.09**	**0.62**	**65.27**	**5.18**	**0.57**
Method	**Mouse ES cells**								
	@500			@800			@1000		
	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓	ACC ↑	Time ↓	SUR ↓
LDA	37.37⋆	7.15	†	37.42⋆	7.04	†	37.76⋆	7.89	†
LADA	37.73⋆	7.26	1.02	37.88⋆	7.60	1.08	37.82⋆	8.70	1.10
DMEG	37.92⋆	7.31	1.02	37.42⋆	7.69	1.09	38.14⋆	8.57	1.09
$\ell_{2,1}$-LDA	38.20⋆	5.36	0.75	38.51⋆	5.70	0.81	38.22⋆	5.92	0.75
SALDA	38.31⋆	5.30	0.74	38.51⋆	5.70	0.81	38.14⋆	5.88	0.75
LDC	39.75⋆	3.78	0.53	39.97⋆	3.70	0.53	39.75⋆	3.80	0.48
ELCS	39.42⋆	3.84	0.54	39.60⋆	3.87	0.55	39.48⋆	3.98	0.50
LAFLDA	39.07⋆	3.92	0.55	39.48⋆	3.81	0.54	39.56⋆	3.90	0.49
FLDA-W	39.26⋆	3.86	0.54	39.80⋆	3.93	0.56	40.11⋆	4.00	0.51
AASDA	**40.32**	**3.28**	**0.46**	**40.62**	**3.42**	**0.49**	**40.40**	**3.73**	**0.47**

Table 1: ACC, running time and speed up ratio of all compared methods on real-world large-scale datasets with different subspace dimensions, where @ denotes the subspace dimension and the notation † denotes the baseline. The p-value of T-Test at the significance level 5% is obtained, and $\star$ means that ACC difference is significant.

References

Belhumeur, P.; Hespanha, J.; and Kriegman, D. 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE TPAMI*, 19(7): 711–720.

Ben-Tal, A.; and Teboulle, M. 1986. Expected Utility, Penalty Functions, and Duality in Stochastic Nonlinear Programming. *Management Science*, 32(11): 1445–1466.

Carrasco, J. S. F.; and Sun, H. 2025. Signed Laplacians for Constrained Graph Clustering. In *ICML*.

Chang, W.; Nie, F.; Wang, Z.; Wang, R.; and Li, X. 2022a. Self-weighted learning framework for adaptive locality discriminant analysis. *PR*, 129.

Chang, W.; Nie, F.; Wang, Z.; Wang, R.; and Li, X. 2022b. Self-weighted learning framework for adaptive locality discriminant analysis. *PR*, 129.

Chen, H.; Nie, F.; Ma, Y.; and Wang, R. 2025. Adaptive fast local discriminant analysis with whitening transform. *NN*, 189(107551).

Ding, C.; and Li, T. 2007. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *ICML*, 521–528.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE TKDE*, 21(9): 1263–1284.

Kalavas, A.; Kipouridis, E.; and Varma, N. 2025. Towards Better-than-2 Approximation for Constrained Correlation Clustering. In *ICML*.

Kolmogorov, A. 1930. Sur la Notion de la Moyenne. *Atti della Academia Nazionale dei Lincei*, 12: 323–343.

Li, X.; Wang, Q.; Nie, F.; and Chen, M. 2022. Locality Adaptive Discriminant Analysis Framework. *IEEE TCYB*, 52(8): 7291–7302.

Moretti, S.; Pellizzoni, P.; and Silvestri, F. 2025. Dimensionality Reduction on Complex Vector Spaces for Euclidean Distance with Dynamic Weights. In *ICML*.

Nie, F.; Dong, X.; Hu, Z.; Wang, R.; and Li, X. 2023a. Discriminative Projected Clustering via Unsupervised LDA. *IEEE TNNLS*, 34(11): 9466–9480.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *KDD*, 977–986.

Nie, F.; Wang, Z.; Wang, R.; Wang, Z.; and Li, X. 2020. Adaptive Local Linear Discriminant Analysis. *ACM TKDD*, 14(1): 1–19.

Nie, F.; Wang, Z.; Wang, R.; Wang, Z.; and Li, X. 2021. Towards Robust Discriminative Projections Learning via Non-Greedy $\ell_{2,1}$-norm MinMax. *IEEE TPAMI*, 43(6): 2086–2100.

Nie, F.; Xiang, S.; Liu, Y.; Hou, C.; and Zhang, C. 2012. Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction. *PRL*, 33(5): 485–491.

Nie, F.; Zhang, C.; Wang, Z.; Wang, R.; and Li, X. 2023b. Local Embedding Learning via Landmark-Based Dynamic Connections. *IEEE TNNLS*, 34(11): 9481–9492.

Nie, F.; Zhao, X.; Wang, R.; and Li, X. 2022. Fast Locality Discriminant Analysis With Adaptive Manifold Embedding. *IEEE TPAMI*, 44(12): 9315–9330.

Nie, F.; Zhao, X.; Wang, R.; and Li, X. 2023c. Adaptive Maximum Entropy Graph-Guided Fast Locality Discriminant Analysis. *IEEE TCYB*, 53(6): 3574–3587.

Pang, Y.; Zhou, B.; and Nie, F. 2019. Simultaneously Learning Neighborship and Projection Matrix for Supervised Dimensionality Reduction. *IEEE TNNLS*, 30(9): 2779–2793.

Peterfreund, E.; Lindenbaum, O.; Kluger, Y.; and Landa, B. 2025. Partition First, Embed Later: Laplacian-Based Feature Partitioning for Refined Embedding and Visualization of High-Dimensional Data. In *ICML*.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

Teboulle, M. 2007. A Unified Continuous Optimization Framework for Center-Based Clustering Methods. *JMLR*, 8: 65–102.

Wang, J.; Yin, H.; Nie, F.; and Li, X. 2024a. Adaptive and fuzzy locality discriminant analysis for dimensionality reduction. *PR*, 151(110382).

Wang, Q.; Wang, F.; Ren, F.; Li, Z.; and Nie, F. 2023. An Effective Clustering Optimization Method for Unsupervised Linear Discriminant Analysis. *IEEE TKDE*, 35(4): 3444–3457.

Wang, Z.; Li, Q.; Nie, F.; Wang, R.; Wang, F.; and Li, X. 2024b. Efficient Local Coherent Structure Learning via Self-Evolution Bipartite Graph. *IEEE TCYB*, 54(8): 4527–4538.

Wang, Z.; Nie, F.; Wang, R.; Yang, H.; and Li, X. 2021. Local structured feature learning with dynamic maximum entropy graph. *PR*, 111(107673).

Yan, L.; Zhang, X.; and Mai, Q. 2025. Heterogeneous Sufficient Dimension Reduction and Subspace Clustering. In *ICML*.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H. J.; Yang, Q.; and Lin, S. 2007. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE TPAMI*, 29(1): 40–51.

Zhang, H.; Chen, S.; Luo, L.; and Yang, J. 2025. Towards Better Spherical Sliced-Wasserstein Distance Learning with Data-Adaptive Discriminative Projection Direction Authors. In *AAAI*, volume 39, 22425–22433.

Zhao, X.; Nie, F.; Wang, R.; and Li, X. 2023. Joint Dynamic Manifold and Discriminant Information Learning for Feature Extraction. *IEEE TNNLS*, 34(6): 2753–2766.

Zhao, X.; Nie, F.; Wang, R.; and Li, X. 2024. Fast Discriminant Analysis With Adaptive Reconstruction Structure Preserving. *IEEE TNNLS*, 35(8): 11106–11115.

Zhou, J.; Gao, C.; Wang, X.; Lai, Z.; Wan, J.; and Yue, X. 2024. Typicality-Aware Adaptive Similarity Matrix for Unsupervised Learning. *IEEE TNNLS*, 35(8): 10776–10790.

Zhu, J.; Tao, L.; Yang, H.; and Nie, F. 2023. Unsupervised Optimized Bipartite Graph Embedding. *IEEE TKDE*, 35(3): 3224–3238.