Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Joint Dual-Structural Constrained and Non-negative Analysis Representation Learning for Pattern Classification

Kun Jiang, Lei Zhu & Qindong Sun

APPLIED ARTIFICIAL INTELLIGENCE
2023, VOL. 37, NO. 1, e2180821 (796 pages)
https://doi.org/10.1080/08839514.2023.2180821

Joint Dual-Structural Constrained and Non-negative Analysis Representation Learning for Pattern Classification

Kun Jiang [a], Lei Zhu[a], and Qindong Sun[b]

[a]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China; [b]School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China

ABSTRACT

In recent years, analysis dictionary learning (ADL) model has attracted much attention from researchers, owing to its scalability and efficiency in representation-based classification. Despite the supervised label information embedding, the classification performance of analysis representation suffers from the redundant and noisy samples in real-world datasets. In this paper, we propose a joint Dual-Structural constrained and Non-negative Analysis Representation (DSNAR) learning model. First, the supervised latent structural transformation term is considered implicitly to generate a roughly block diagonal representation for intra-class samples. However, this discriminative structure is fragile and weak in the presence of noisy and redundant samples. To highlight both intra-class similarity and inter-class separation for class-oriented representation, we then explicitly incorporate an off-block suppressing term on the ADL model, together with a non-negative representation constraint, to achieve a well-structured and meaningful interpretation of the contributions from all class-oriented atoms. Moreover, a robust classification scheme in latent space is proposed to avoid accidental incorrect predictions with noisy information. Finally, the DSNAR model is alternatively solved by the K-SVD method, iterative re-weighted method and gradient method efficiently. Extensive classification results on five benchmark datasets validate the performance superiority of our DSNAR model compared to other state-of-the-art DL models.

ARTICLE HISTORY
Received 12 December 2022
Revised 30 January 2023
Accepted 8 February 2023

Introduction

The rapid development of AI technology and the growth of Internet big data have brought more opportunities and challenges to pattern classification. The classification task generally involves using a classifier learning model to automatically predict the label vectors for new samples based on knowledge or statistical information learned from features or patterns in a given dataset (Bishop 2006; Duda, Hart, and Stork 2001). Specifically, the classifier training process leverages a couple of supervised model learning methods, requiring

CONTACT Kun Jiang jk_365@xaut.edu.cn School of Computer Science and Engineering, Xi'an University of Technology, No. 5 South Jinhua Road, Xi'an, China

a training dataset that contains both the training samples and the corresponding label information.

The training data and its label information are then used to learn a discriminative and robust classification model (Jiang, Lin, and Davis 2013; Wang et al. 2018; Zhang and Li 2010). Then, the learned classification model can be used to predict the labels of test samples. Popular classification methods include Linear Regression (LR) (Nie et al. 2010), K-Nearest Neighbors (KNN) (Cover and Hart 1967), Support Vector Machine (SVM) (Cortes and Vapnik 1995), etc. With the arrival of the era of big data, pattern classification over large-amounts of high-dimensional data samples has become a fundamental and challenging problem in many real-world applications, such as image restoration (Mairal, Elad, and Sapiro 2008; Wright et al. 2009), image classification and recognition (Yang et al. 2009), computer vision (Zhang et al. 2013) and so on (Kong and Wang 2012; Xu et al. 2019). However, data samples collected from various modern sensing system may be composed of high-dimensional, noisy/corrupted, redundant samples. Therefore, the conventional sample-level classification models suffer from the high-dimensional noisy and redundant sample features.

To obtain robust classification models, representation-based classification (RBC) models, such as sparse representation-based classification (SRC) (Wright et al. 2009) and collaborative representation-based classification (CRC) (Zhang, Yang, and Feng 2011), have emerged as the main research aspect by exploring intrinsic property of data samples. Particularly, sparse representation has attracted much attention, due to its discriminative and compact characteristics, with which an input sample can be coded as a linear combination of a few atoms from an over-complete dictionary (Wright et al. 2009). The dictionary learning (DL) methods play a vital role in sparse representation, which can be classified into two categories by the way of encoding samples, i.e., synthesis dictionary learning (SDL) model and analysis dictionary learning (ADL) model (Aharon, Elad, and Bruckstein 2006; Rubinstein, Peleg, and Elad 2013). Given signals $Y = [y_1, y_2, \cdots, y_n] \in R^{m\times n}$, let $D = [d_1, d_2, \cdots, d_k] \in R^{m\times k}$ be a synthesis dictionary with a series of atom d_i, and $X = [x_1, x_2, \cdots, x_n] \in R^{k\times n}$ be the sparse coefficient matrix. A SDL model expects to learn an over-completed dictionary D with $k > m$ by minimizing the reconstruction errors such that it can exactly represent the samples Y with representation matrix X. The sparse optimization problem of SDL can be formulated as follows.

$$\min_{D,X} \| Y - DX \|_F^2$$

$$s.t.\ D \in \Gamma,\ \| x_i\|_0 \leq T_0,\ i = 1, 2, \ldots, n \tag{1}$$

where $\| Y - DX \|_F^2$ is the reconstruction error term, Γ is a set of constraints on over-complete D to ensure a stable solution, and T_0 is sparsity level for each coefficient x_i (Aharon, Elad, and Bruckstein 2006).

The solution of SDL model includes two basic tasks, i.e., sparse approximation and dictionary learning. On one hand, some algorithms, such as matching pursuit (M-P) (Davis, Mallat, and Avellaneda 1997; Mallat and Zhang 1993), basis pursuit (BP) (Chen, Donoho, and Saunders 2001) and shrinkage method (Hyvärinen 1999), have been well developed to find a sparse solution. On the other hand, dictionary learning is dedicated to search an optimal signal space to support the attribution of sparse vector under a certain measure. There exist a variety of numerical algorithms presented to achieve this objective, e.g., method of optimal directions (MOD) (Engan, Aase, and Hakon Husoy 1999) and K-singular value decomposition (K-SVD) (Aharon, Elad, and Bruckstein 2006). K-SVD method learns an overcomplete dictionary from training samples by updating K dictionary atoms and representation coefficients iteratively with the SVD algorithm under a predefined sparse threshold for non-zero elements in each coefficient (Aharon, Elad, and Bruckstein 2006).

However, the basic SDL model focuses on the representation ability, but lacks discrimination. Plenty of research on exploring the discrimination of SDL model has been proposed for pattern classification (Jiang, Lin, and Davis 2013; Kong and Wang 2012; Yang et al. 2011; Zhang and Li 2010). For instance, D-KSVD incorporates the classification error term into the basic SDL model to enhance the discrimination (Zhang and Li 2010). As the labels of atoms can also be used to improve the discriminative ability of the model, LC-KSVD further adds a label-consistent term into the objective function of D-KSVD (Jiang, Lin, and Davis 2013).

As a dual viewpoint of SDL model, the ADL model mainly focuses on learning a projection matrix, and constructing sparse analyzed vectors in transformation subspace (Hawe, Kleinsteuber, and Diepold 2013). The ADL model aims to learn an analysis dictionary $\Omega \in R^{k \times m}$ with $k > m$ to implement the approximately sparse representation of the signal $y \in R^m$ in transformed domain (Ravishankar and Bresler 2013; Rubinstein, Peleg, and Elad 2013). Specifically, it assumes that the product of Ω and y is sparse, i.e., $x = \Omega y$ with $\| x_i \|_0 = k - t$, where $0 \leq t \leq k$ is the number of zeros in $x \in R^k$. The sparse optimization problem can be formulated as follows:

$$\min_{\Omega, X} \| X - \Omega Y \|_F^2$$

$$s.t.\ \Omega \in \Gamma',\ \| x_i \|_0 \leq T_0,\ i = 1, 2, \ldots, n \tag{2}$$

where Γ' is a set of constraints on over-complete Ω to ensure a stable solution, and the representation error term $\| X - \Omega Y \|_F^2$ shows the disparity between representations in transformed space and the coefficients with sparsity level

T_0. Some algorithms like backward-greedy (BG), greedy analysis pursuit (GAP) and Analysis KSVD have been proposed to address this problem (Rubinstein, Peleg, and Elad 2013). However, the computational complexity of these algorithms is very high, and some recent work has relaxed the l_0-norm of sparse constraint in equation (2) to the convex l_1-norm form or adding thresholding function to each coefficient vector (Li et al. 2022; Shekhar, Patel, and Chellappa 2014).

The ADL model has aroused much attention as it has a more intuitive illustration for the role of analysis atoms and has a lower classification complexity. With ADL models, the coefficients of training samples are used as transformed space features for the jointly learned classifier. Then, the testing samples can also be linearly projected into transformed feature space by analysis dictionary efficiently, rather than by a nonlinear sparse reconstruction in SDL models. However, the classical ADL model mainly focuses on the representational ability of the dictionary without considering its discriminative capability for classification (Ravishankar and Bresler 2013). To conduct classification tasks, Shekhar et al. (Shekhar, Patel, and Chellappa 2014) performed a two-step ADL+SVM model, in which an analysis dictionary is first learned and then used to obtain projective coefficients for data samples. The coefficients of training samples and testing samples are used as transformed space features for SVM classifier. The classification process acts as a post-step for the analysis representation.

To facilitate discrimination for analysis dictionary and representation, researchers have presented several classification task-oriented models incorporating the supervised label embedding (Wang et al. 2017, 2018), class-oriented reconstruction (Wang et al. 2017) and ideal structural information (Tang et al. 2019) during learning procedure. The performance of classification of ADL models benefits from these discriminative terms and the higher speed for testing in representation projections (Guo et al. 2016). However, the aforementioned discriminative ADL models are all performed in an ideal ambient space without noises or corruptions. Since data samples usually contain noisy or redundant samples, the coding coefficients would be contaminated and the discrimination of the ADL model may be degraded (Wang et al. 2017). Moreover, the hidden discriminative information could not be fully exploited by the supervised label information of the training samples due to the noises and redundant samples (Jiang, Lin, and Davis 2013; Li et al. 2017; Zhang and Li 2010). The supervised information in common-used label matrix could deviate a lot or become weaker during the learning procedure, which may lead to unreasonable substraction of atom's contributions in transforming the training samples.

In view of these limitations, we are dedicated to explore more discrimination on analysis representation for classifier learning under complex data environment, and propose a joint Dual-Structural constrained and Non-negative Analysis Representation (DSNAR) learning model. Specifically, we introduce a joint dual-structural term to guarantee block diagonal compactness and off-block diagonal separation for analysis representation. In particular, we first construct a latent structural transformation term with the supervised information of dictionary atoms, which is often overlooked in previous work. Under noisy and redundant data environment, the off-block entries in representation matrix indicates inter-class contributions, due to the correlations induced by extra information with the same distribution. Therefore, we propose to gradually reduce the inter-class contribution by adding an off-block diagonal suppression term into the objective function. And we also consider the non-negative constraint to make the analysis contributions physically meaningful, despite contributions from intra-class samples. With these components, the proposed DSNAR model can guarantee a clear block diagonal structure and meaningful interpretability for low-dimensional representation by reducing the adverse effect of noisy and redundant samples, and inter-class correlation of representation.

The main contribution of this paper is as follows.

- Firstly, we incorporate the latent structural transformation term and off-block suppression term into the classical ADL model with non-negative constraint for discriminative and robust representation.
- Secondly, an alternating direction solution is proposed to optimize the proposed objective function, including K-SVD method, adaptive iterative re-weighted method and gradient method for each subproblem.
- Thirdly, we present an efficient and robust classification scheme in latent label space. Empirical study on five benchmark datasets shows the efficacy and efficiency of our DSNAR model.

The rest of this paper is organized as follows. The related work is presented in Section 2. The discriminative and robust ADL model and the solution are presented in Section 3, together with computational complexity and convergence analysis. Section 4 presents a robust classification scheme. Experimental results on pattern classification are presented in Section 5. Relevant conclusions are finally given in Section 6.

Related Work

In this section, we review some discriminative ADL models and the non-negative constraint technique that are closely related to our work.

Analysis Dictionary Learning

The conventional ADL models mainly focus on the representational ability of the dictionary without considering its discriminative capability for classification (Ravishankar and Bresler 2013). To enhance the discriminative power and efficiency of ADLs, there are various classification task-oriented improvements with well-conditioned regularizers, such as SK-ADL (Wang et al. 2018), DADL (Guo et al. 2016), CADL (Wang et al. 2017) and SADL (Tang et al. 2019). To achieve a global optimum solution for discriminative ADL model, Guo et al. incorporated the structural code consistent term and topology preserving term (DADL) into the conventional ADL model to yield a discriminative representation (Guo et al. 2016). By introducing a synthesis-linear-classifier to map the label information into feature space, Wang et al. (Wang et al. 2017) presented a synthesis linear classifier-based ADL (SLC-ADL) algorithm. Note that the synthesis linear classifier term can be assumed as a label-consistent term of the dictionary atoms that leads to an ideal structural representation. At the same period, Wang et al. (Wang et al. 2018) designed a synthesis K-SVD based ADL (SK-ADL) model by jointly learning ADL and a linear classifier through K-SVD method. After that, Tang et al. (Tang et al. 2019) incorporated a class characteristic structure of independent subspaces term and classification error term into the framework of ADL (SADL). The SK-ADL (Wang et al. 2018) and SADL (Tang et al. 2019) models combine the classification error term with the basic ADL framework, in which the supervised class label information is utilized to guide the generation of representation.

Recently, researchers have focused on exploring the underlining structural information of the label information of analysis dictionary (Du et al. 2021; Li et al. 2021; Wang et al. 2017). For example, Du et al. Du et al. (Du et al. 2021) proposed a structured discriminant analysis dictionary learning (SDADL) method that exploits the partial class-oriented analysis subdictionaries and the ideal analysis sparse code error term. The uniform class-oriented atom allocation scheme is adopted to initialize the analysis representation and the ideal sparse-code matrix. Li et al. (Li et al. 2021) incorporate the Discriminative Fisher criterion constraint on the profiles and atoms pairs of the Structured Analysis Dictionary Learning (SADL-DFP) model, which enhance the class-oriented separability and compactness of the analysis dictionary. Besides, these models ignore the inter-class projective contributions of the fine-grained analysis atoms under the noises and redundant data environment (Li et al. 2021; Zhang et al. 2021).

Another research direction is to consider both the synthesis and analysis dictionary for projective dictionary pair learning (PDPL), which considers both representative and discriminative ability for discrimination representation (Gu et al. 2014; Yang et al. 2017; Zhang et al. 2018). Recently, Zhang et al. incorporated the the synthesis dictionary incoherence penalty, analysis sparse code extractor and multi-variant classifier to learn a structured and compact dictionary pair (Zhang et al. 2018). However, the class-oriented objective function always involves time-consuming multiplications of subdictionaries with their corresponding complementary samples (Chen, Wu, and Kittler

2022; Jiang et al. 2022). Chen et al. proposed a relaxed block-diagonal PDPL method by dynamically optimizing the block diagonal entries of representation, while directly setting the off-block diagonal counterparts to zero (Chen, Wu, and Kittler 2022).

Non-Negative Representation

For real-world data samples with noisy or redundant information, representation-based classification models intend to learn a robust and discriminative representation by leveraging some predetermined underlining structural characteristics of data samples. Existing researches have focused on the sparsity (Wright et al. 2009), collaboration (Zhang, Yang, and Feng 2011) or grouping effect (Lu et al. 2013) to characterize the representation contribution of data samples or dictionary atoms. Despite these characteristics, the negative entries in coefficient are neglected, which lacks a reasonable interpretation as the substraction of sample contribution should be physically prohibited. The non-negative constraint is widely applied in non-negative matrix factorization (Yi et al. 2020), subspace clustering (Chen et al. 2021; Zhuang et al. 2012), representation-based classification (Xu et al. 2019) and so on, for explanative part-based representation. Specifically, among representation-based classification methods, there still exists controversy between sparsity and collaborative mechanism (Wright et al. 2009; Zhang, Yang, and Feng 2011). Nevertheless, Xu et al. (Xu et al. 2019) further found that without non-negative constraint, a sample will be represented by both heterogeneous and homogeneous samples, which brings about a difficult physical interpretation. By restricting the values of representation vector to be non-negative, the contributions of homogeneous samples can be enlarged, meanwhile, eliminating the adverse effects caused by heterogeneous samples. However, to the best of our knowledge, there are few works to address the non-negative representations of ADL model. Viewed from this perspective, traditional ADL methods have overlooked the physical meaningful contribution of dictionary atoms from heterogeneous classes when generating representation. And even some of the intra-class atoms could have subtracted contribution to neutralize larger contributions from other nearby atoms. Motivated by the conjecture that the non-negativity can boost the selection of representative atoms, we consider the non-negative constraint to analysis representation, so that the learned analysis dictionary atoms are more high-quality and discriminative.

In view of the advantages and limitations of discriminative ADL models in classification task, we believe that the discriminative promotion can be further carried out by enhancing the discrimination of the analysis atoms and the robustness of the representation to noises and corruptions. To have a clear understanding of the existing discrimination promotion techniques, we summarize the comparative technical characteristics of the existing ADL models,

Table 1. Comparative characteristics of the related works on discriminative ADL models. The regularization term of the DSNAR model corresponding to each technical component is shown in parentheses.

ADL Models	Class-Oriented	Label-Embedded Dictionary	Label-Embedded Representation	Nonnegative Representation
ADL-SVM	✗	✗	✗	✗
DADL	✗	✓	✗	✗
CADL	✓	✗	✓	✗
SK-ADL	✗	✗	✓	✗
SADL	✗	✓	✓	✗
SDADL	✓	✓	✓	✗
SADL-DFP	✗	✓	✓	✗
DSNAR	✓ ($\|\|X \odot S\|\|_F^2$)	✓ ($\|\| L - AX \|\|_F^2$)	✓ ($\|\| H - AX \|\|_F^2$)	✓ ($X \geq 0$)

together with the proposed DSNAR model in Table 1. The regularization term of the proposed DSNAR model corresponding to each technical component described in Section 3 is shown in parentheses. As can be seen, our proposed DSNAR model exhibits superior theoretical merits compared with existing ADL models. The latent structural transformation term contains label embedding on analysis dictionary atoms and label embedding on representations, which provides more robustness to models under real-world data environment. Specifically, the label embedding on atoms guarantees a compact and discriminative analysis dictionary when handling noises and outliers, while class-oriented and label embedding on representations yields ideal block diagonal representation. The non-negative representation term mainly promotes the intra-class similarity of representations by eliminating the meaningless contributions from inter-class samples.

The Proposed Model

In this section, we present a joint Dual-Structural Constrained and Non-negative Analysis Representation (DSNAR) learning model that leverages robust latent structural transformation and off-block suppressed representation for pattern classification.

Latent Structural Transformation

As mentioned above, the discrimination of the analysis dictionary cannot be fully explored only with the supervised information of training samples, due to the noisy and redundant high-dimensional features (Guo et al. 2016). Recently, the supervised label vectors of adaptive updated atoms and class-oriented concatenated subdictionaries have been utilized to improve the discrimination of dictionary (Jiang, Lin, and Davis 2013; Li et al. 2017). Inspired by Jiang, Lin, and Davis (2013); Li et al. (2017), we explore the discriminative structural properties in the analysis representation model by leveraging the class-oriented label information of analysis dictionary atoms.

Suppose $H = [h_1, h_2, \cdots, h_n] \in R^{c\times n}$ be the label matrix with each column $h_i = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ describes the label vector of the jth sample; the non-zero value of h_i occur at the index j where training sample belong to class $j(1 \leq j \leq c)$. Since there is only one nonzero element h_{ij} in each label vector h_j of the label matrix H, the class of a sample is only determined by one element's location i in a label vector h_j, which may easily drift a lot during the training procedure. Therefore, we extend the label matrix H into a latent space L by a Kronecker product to improve its robustness to some extent (Wang et al. 2017). Specifically, given an all ones vector $\mathbf{l}$ with length L, the extended class label matrix L is defined as

$$L \equiv H \otimes \mathbf{l} = \begin{bmatrix} h_{1,1}\mathbf{l} & \cdots & h_{1,n}\mathbf{l} \\ \vdots & \ddots & \vdots \\ h_{c,1}\mathbf{l} & \cdots & h_{c,n}\mathbf{l} \end{bmatrix} \in R^{lc\times n} \quad (3)$$

For example, assuming that there is a training set $Y = [y_1, y_2, \ldots, y_7]$ and an analysis dictionary $\Omega = [\omega^1; \omega^2; \ldots; \omega^6]$, in which y_1, y_2, ω^1 and ω^2 are from class 1, y_3, y_4, y_5, ω^3 and ω^4 are from class 2 and y_6, y_7, ω^5 and ω^6 are from class 3. As is recommended in previous work (Jiang, Lin, and Davis 2013), to allocate dictionary atoms to classes uniformly, we select 2 atoms for each one of the 3 classes. According to equation (3), the extended class label matrix L can be defined as

$$L \equiv \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \in R^{6\times 7} \quad (4)$$

To establish the corresponding relationship between the analysis representation X and the robust latent label vectors, we introduce $\| L - AX \|_F^2$ as the latent structural transformation term for implicit block diagonal constraint on representation, where $A \in R^{lc\times k}$ is the transformation matrix. This regularization term encourages similarity among sparse coefficient vectors belonging to the same class. Note that the latent structural transformation term can be viewed as an ideal sparse-code error term of the dictionary atoms if we set the number of dictionary atoms $k = lc$, which means that there are l atoms chosen from each class and arranged sequentially. And this uniform class-oriented allocation strategy is a commonly used composition of analysis dictionary in some existing ADL models (Du et al. 2021; Li et al. 2021). In this case, the analysis representation X and extended label matrix l have the identical matrix size. In this paper, we define the dictionary size $k = lc$ for simplicity, and the $L_{k\times n}$ is denoted as the ideal sparse-code matrix. This indicates the latent space transform term roughly leads to an ideal structural representation X.

Suppose the label matrix H is not extended, i.e., $l = 1$, then the traditional linear classifier $\| H - WX \|_F^2$ $(W \in R^{c\times k})$ is constructed and incorporated into ADL model for generating discriminative label embedded representation. That is to say, the traditional $\| H - WX \|_F^2$ term can be seen as a special case of our latent structural transformation term $\| L - AX \|_F^2$. In Section 5, we prove that adding an extra linear classifier term is redundant compared to the latent structural transformation term for ADL models. This is mainly because the latent structural transformation term can simultaneously guarantee the discrimination of analysis dictionary and representation by combining label embedded atoms and label embedded representation.

According to aforementioned analysis on ideal structural representation, we formulate the analysis representation learning model by incorporating the latent structural transformation term into the classical ADL model as follows.

$$\min_{\Omega,W,X} \| X - \Omega Y \|_F^2 + \alpha \| L - AX \|_F^2$$

$$s.t.\ \Omega \in \Gamma', \| x_i \|_0 \leq T_0, i = 1, 2, \ldots, n \tag{5}$$

Off-Block Diagonal Representation

When there are noisy and redundant samples, the fragile block diagonal structure would be destroyed with the implicit ideal structural constraint (Chen, Wu, and Kittler 2022). Thus the discrimination of representation and the performance of classification could be degraded. This is mainly because of the inter-class coefficients in the downgraded discriminative representation that represent randomly distributed noisy or redundant features across all samples. Motivated by the block diagonal representation learning method in subspace clustering (Zhang et al. 2018), we further present a discriminative off-block diagonal suppression term to the coefficients X as follows,

$$\min_{X} \|X \odot S\|_F^2 \tag{6}$$

where $\odot$ means the Hadamard product (element-wise multiplication) operator, and the off-block indicator matrix $S \in R^{k\times n}$ is predefined as

$$S_{ij} = \begin{cases} 0, & \textit{if } \omega^i \textit{ and } x_j \textit{ belong to the same class } r \\ 1, & \textit{otherwise} \end{cases} \quad (7)$$

where ω^i is the ith row vector of the dictionary Ω and belongs to the rth subdictionary $\Omega_r \subset \Omega$, and x_j is the jth column of training samples matrix X. In other words, if representation x_j corresponding to sample y_j belongs to class r, then the indicator entry S_{ij} associated with atoms $(\quad i)\ \omega^i \in \Omega_r$ are all 0s, and the others are all 1s. With the uniform class-oriented analysis dictionary allocation scheme mentioned in Section 3.1, S can be computed from the ideal sparse-code matrix $L_{k\times n}$ as $S = \mathbf{1}_k\mathbf{1}_n^{\mathrm{T}} - L_{k\times n}$. For the example in subsection 3.1, S is computed as

$$S \equiv \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \in R^{6\times 7} \quad (8)$$

The off-block diagonal suppression term encourages the model discrimination by restraining the off-block diagonal coefficient entries of inter-class projections. From this perspective, the off-block diagonal suppression term is equivalent to class-oriented constraint on representation for inter-class separation. Some existing DL models leverage the class-oriented projection form $\sum_{i=1}^{n} \| \Omega_i Y_{\hat{i}} \|_F^2$, where $Y_{\hat{i}}$ denotes the complementary subset of Y_i in the entire dataset Y (Du et al. 2021; Gu et al. 2014). Under the ideal conditions, minimizing the class-oriented projection term ensures that each subdictionary Ω_i can map all sample subsets $Y_j(j\neq i)$ from other classes into a nearly null representation space, and which is equivalent to the off-block diagonal suppressing term in Equation (6). But the advantage of the off-block diagonal suppressing term is that all the class-oriented dictionary can be optimized simultaneously instead of the time-consuming multiplications of the subdictionary with the complementary samples from other class.

The non-negative constraint on representation $X \geq 0$ is also considered in our model to prevent substraction of contributions from heterogeneous samples. Although some of the intra-class atoms could have subtracted contribution to neutralize larger contributions from other nearby atoms. In most cases, the non-negative constraint term promotes the intra-class similarity of representations by eliminating the meaningless contributions mainly from inter-class samples. Benefiting from the non-negative constraint, the off-block diagonal structure is further guaranteed and the compactness of intra-class coefficient vectors is also promoted. On the other hand, with iterative updating rules, the non-negativity will boost the selection of representative atoms; thus, the learned analysis dictionary atoms are more high-quality and discriminative.

The Objective Function

To achieve optimal solutions for pattern classification, we integrate the latent structural transformation term in Equation (5), the off-block diagonal suppression term in Equation (6) and non-negative representation constraint with the basic ADL model to formulate the proposed joint Dual-Structural Constrained and Non-negative Analysis Representation (DSNAR) learning model. The objective function of the DSNAR model is presented as follows:

$$\min_{\Omega,A,X} \| X - \Omega Y \|_F^2 + \alpha \| L - AX \|_F^2 + \beta \|X \odot S\|_F^2$$

$$s.t. \quad \Omega \in \Gamma', X \geq 0, \| x_i\|_0 \leq T_0, i = 1, 2, \ldots, n \tag{9}$$

where α and β are regularization coefficients. The first term is the representation error term of ADL model; the second term is the latent structural transformation term which combines the virtues of classification error term and ideal structural term. These dual-structural constraints ensure well-structured representation for high-dimensional data despite noisy and redundant information. And the third term yields discriminative suppression on the cross-class representation coefficients. We believe that the DSNAR model could sufficiently provide a discriminative analysis dictionary and robust well-structured

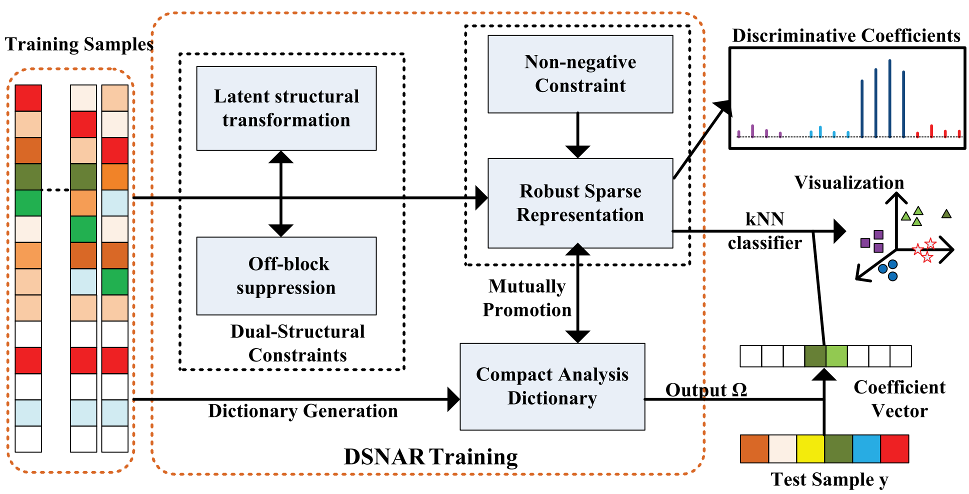

Figure 1. Framework of the proposed DSNAR model.

representations for classification under complex real-world data environment. The framework of the proposed DSNAR model is illustrated in Figure 1.

Solution to DSNAR Model

In this section, we present the optimization algorithm of the objective function, including the alternative updating procedure and the initialization details, together with computational and convergence analysis.

Optimization Algorithm

The optimization problem (9) is a multiple variables optimization problem, and it is non-convex w.r.t. Ω, X and A jointly. We first add one variable $Z = X$ to make the problem separable, and then problem (9) can be rewritten as

$$\min_{\Omega,A,X,Z} \| X - \Omega Y \|_F^2 + \alpha \| L - AX \|_F^2 + \beta \|Z \odot S\|_F^2$$

$$s.t.\ \Omega \in \Gamma', Z = X, Z \geq 0, \| x_i\|_0 \leq T_0, i = 1, 2, \ldots, n \tag{10}$$

Then, we can get the following objective function of the problem by the ALM method. Here, the augmented Lagrangian function of problem (10) is

$$\mathcal{L}(\Omega, A, X, Z, C_1) = \| X - \Omega Y \|_F^2 + \alpha \| L - AX \|_F^2 + \beta \|Z \odot S\|_F^2$$

$$+ < C_1, Z - X > + \frac{\mu}{2} \| Z - X \|_F^2 \tag{11}$$

where $<P, Q> = tr(P^T Q)$, C_1 is the Lagrangian multipliers, and $\mu > 0$ is a penalty parameter. The problem can be divided into three convex sub-problems.

(1) Fix {Ω,Z} and update {A,X}. The sub-problem for updating {A,X} is

$$\min_{A,X} \| \Omega Y - X \|_F^2 + \alpha \| L - AX \|_F^2 + \frac{\mu}{2} \| Z - \frac{C_1}{\mu} - X \|_F^2$$

$$s.t.\quad \| x_i\|_0 \leq T_0,\ i = 1, 2, \ldots, n \tag{12}$$

which is equivalent to

$$\min_{A,X} \left\| \begin{pmatrix} \Omega Y \\ \sqrt{\alpha} L \\ \sqrt{\mu/2}(Z - C_1/\mu) \end{pmatrix} - \begin{pmatrix} R_1 \\ \sqrt{\alpha} A \\ \sqrt{\mu/2} R_2 \end{pmatrix} X \right\|_F^2 \tag{13}$$

$$s.t.\quad \| x_i\|_0 \leq T_0,\ i = 1, 2, \ldots, n$$

where R_1 and R_2 are initialized as an identity matrix I. The optimization problem in (13) can be efficiently solved by K-SVD method (Aharon, Elad, and Bruckstein 2006).

For convenience of description, let $Y_{new} = \left(Y^T\Omega^T, \sqrt{\alpha}L^T, \sqrt{\mu/2}(Z^T - C_1^T/\mu)\right)^T$, $D_{new} = \left({R_1}^T, \sqrt{\alpha}A^T, \sqrt{\mu/2}R_2^T\right)$, then equation (13) is equivalent to the following problem:

$$\begin{aligned} &\min_{D_{new},X} \parallel Y_{new} - D_{new}X \parallel_F^2 \\ s.t. \quad & D_{new} \in \Gamma, \parallel x_i\|_0 \leq T_0, i = 1, 2, \ldots, n \end{aligned} \tag{14}$$

where Γ is constrained to be column-wise l_2 norm. After convergence, we could take out A from D_{new} with a column-wise l_2 normalization respectively. Then, let x_R^k, which is the k-th row of X, be the corresponding coefficients of the k-th column of D_{new}, and denoted as d_k. Following K-SVD method, the atom $d_k \in D_{new}$, and its corresponding coefficients, which is the k-th row of X, denoted as x_R^k are updated simultaneously. Let $E_k = (Y_{new} - \sum_{j\neq k} d_j x_R^j)$, then discard the zero entries in x_R^k and E_k to form two matrices as $\widetilde{x}_R^k$ and $\widetilde{E}_k$, respectively. Finally, d_k and $\widetilde{x}_R^k$ can be obtained by

$$< d_k, \widetilde{x}_R^k > = \arg\min_{d_k, \widetilde{x}_R^k} \left\| \widetilde{E}_k - d_k \widetilde{x}_R^k \right\| \tag{15}$$

Specifically, decomposing $\widetilde{E}_k$ by an SVD operation, we have $\widetilde{E}_k = U\Sigma V^T$. Let $d_k = U(:,1)$ and $\widetilde{x}_R^k = \Sigma(1,1)V(:,1)$. The nonzero values of x_R^k are replaced by $\widetilde{x}_R^k$.

After one iteration, D_{new}^{t+1} for next iteration's K-SVD process is directly assigned by $D_{new}^{t+1} = D_{new}^t$, instead of renewing it with identity matrix and A^t. This is due to the independent variables A, R_1 and R_2 in D_{new}^{t+1}. In addition, R_1 and R_2 are empirically very similar to identity matrix until convergence. Consequently, it is tolerable to take out A from D_{new} at the end of the training stage without extra operations, such as strictly constraining R to be identity matrix or further column-wise l_2 normalization.

(2) Fix $\{\Omega, A, X\}$ and update Z. The sub-problem for updating Z is

$$\mathcal{L} = \min_Z \beta \parallel Z \odot S \parallel_F^2 + < C_1, Z - X > + \tfrac{\mu}{2} \parallel Z - X \parallel_F^2 \quad s.t. \; Z \geq 0 \tag{16}$$

The above equation is difficult to be solved due to the Hadamard product. We propose an iterative re-weighted method to capture the block diagonal elements of Z; thus, the subproblem is equivalent to

$$\mathcal{L} = \min_Z \beta \parallel Z - Q_1 \parallel_F^2 + \tfrac{\mu}{2} \parallel Z - X - \tfrac{C_1}{\mu} \parallel_F^2 \quad s.t. \; Z \geq 0 \tag{17}$$

where $Q_1 = Q \odot Z^{(t)}$ represents the block diagonal elements of Z in last iteration, and $Q = [q_1, q_2, \cdots, q_n] \in R^{k\times n}$ is a predefined ideal representation matrix and $q_i \in R^k$ has a form of $[0,\cdots,0,1,\cdots,1,0,\cdots,0]^T$, which can be viewed as a coding vector for sample $y_i \in R^m$. If sample y_i belongs to class k, the coefficients in q_i associated with the row vectors in Ω_k are all 1s, whereas others are all 0s. Equation (17) can be reformulated into trace form as

$$\mathcal{L} = \min_{Z} \ \beta tr(ZZ^T - 2Q_1Z^T + Q_1Q_1^T) + \frac{\mu}{2} tr(ZZ^T - 2X_1Z^T + X_1X_1^T) \quad (18)$$

By setting the derivation $(\partial\mathcal{L}/\partial Z) = 0$, the closed-form solution of Z is

$$Z = (2\beta Q_1 + \mu X + C_1)/(2\beta + \mu) \quad s.t. \ \ Z \geq 0 \quad (19)$$

In each iteration, all the negative elements and the lowest values in Z are set to 0, thus generating non-negative representations.

(3) Fix $\{A, X, Z\}$ and update Ω. For simplicity, we constrain the set Γ' to be matrices with relatively small Frobenius norm and unity row-wise norm. Thus, we add a regularization term into the sub-problem for updating Ω as

$$\min_{\Omega} \parallel X - \Omega Y \parallel_F^2 + \gamma \parallel \Omega \parallel_F^2 \quad (20)$$

where $\gamma > 0$ is a parameter which weighs the penalty term to avoid singularity and overfitting issues as well as ensuring a stable solution. After omitting the independent terms w.r.t. Ω in trace forms, an equivalent problem is obtained as

$$\min_{\Omega} Tr(Y^T\Omega^T\Omega Y - 2X^T\Omega Y + \gamma\Omega^T\Omega) \quad (21)$$

By setting the derivative of the objective function w.r.t. Ω be zero, we can finally obtain the analytical solution of Ω as follows.

$$\Omega = XY^T\left(YY^T + \gamma I\right)^{-1} \quad (22)$$

where I is an identity matrix with appropriate size and γ is a small positive scalar for regularizing the solution. Finally, each row vector of Ω is normalized to unit norm to avoid trivial solution.

Initialization

In this subsection, we present the initialization of our DSNAR model. As the analysis dictionary learning cannot estimate the initial dictionary by K-SVD method. Fortunately, the analysis representation coefficients and the dictionary matrix have the identical row number. The latent space label matrix of the analysis dictionary atoms can be viewed as the ideal structural representation, indicating an equal contribution for each corresponding class-oriented atom. Following the initialization metric in Wang et al. (2017), the latent space label matrix is leveraged to initialize sparse analysis representation X_{ini} for training data, i.e., $X_{ini} = L$.

With initialized sparse representation X_{ini}, the analysis dictionary Ω_{ini} can be initialized according to equation (22). The transformation matrix A is initialized to by solving a multivariant ridge regression model.

$$A_{ini} = \arg\min_{A} \|L - AX_{ini}\|_F^2 + \delta\|A\|_F^2 \tag{23}$$

where δ is a scalar with a small empirical value to avoid overfitting by restricting the energy of A. Let the derivative of the objective function w.r.t. A be zero, the solution can be obtained by

$$A_{ini} = \alpha L X_{ini}{}^T \left(\alpha X_{ini} X_{ini}{}^T + \delta I\right)^{-1} \tag{24}$$

The above training procedure is summarized in Algorithm 1.

Algorithm 1 Algorithm for solving our proposed model (4)

Input: Training data Y, latent space matrix L, model parameter α,β, sparsity level T_0 and convergence tolerance $\varepsilon = 1e - 6$, maximum iteration T_{max}.
Output: The analysis dictionary Ω and transformation matrix A.
1: Initialize Ω by (22), A by (23), $Z = X = L$, $C_1 = \mathbf{0}^{k\times n}$, $\rho = 1.01$, $\mu = 10^{-8}$, $\mu_{max} = 10^8$, $t = 1$.
2: **while** not converge and $t < T_{max}$ **do**
3: fixing Ω, Z, update A, X by solving model (13) using K-SVD method;
4: fixing Ω, A, X, update Z by Eq. (19);
5: fixing A, X, Z, update Ω by Eq. (22);
6: update $C_1 = C_1 + \mu(Z - X)$, $\mu = min(\rho\mu, \mu_{max})$
7: check the convergence condition: $\|Z^t - X^t\|_\infty < \varepsilon$, and $t = t + 1$
8: **end while**

Algorithm Analysis

Structural Constraint Equivalence

Some existing work has dedicated to explore the linear classifier error term or ideal sparse-code error term to enforce label consistency of ADL model for discriminative representation (Du et al. 2021; Tang et al. 2019; Wang et al. 2018). According to existing DL work, we can prove that these two terms are the special cases of our latent structural transformation term (Kviatkovsky et al. 2017).

Suppose that we further add a linear classifier term $\| H - WX \|_F^2$ to the objective function (9) to form a comprehensive ADL model. The main difference of the two considered models lies in updating the first subproblem (13). First, we define the reshuffled matrix L' as $L' = P_\pi L$, and P_π is the permutation matrix corresponding to permutation function $\pi : \{1, \ldots, lc\} \rightarrow \{1, \ldots, lc\}$. The function π is defined over L's ith row index as

$$\pi(i) = ((i-1) \bmod c) * l + \left\lfloor \frac{i-1}{c} \right\rfloor + 1 \tag{25}$$

By reshuffling the rows of L with permutation matrix P_π, we reformulate L into the comprehensible form as

$$L' = \Big(\underbrace{H^T, \ldots, H^T}_{\times l}\Big)^T \tag{26}$$

Considering the orthonormality of P_π and some necessary algebraic derivation, the equation (13) can be deduced to its equivalent reshuffled K-SVD-like form as

$$\begin{aligned} \min_{\tilde{D}, A', X} & \left\| \begin{pmatrix} \tilde{Y} \\ \sqrt{\alpha} L' \end{pmatrix} - \begin{pmatrix} \tilde{D} \\ \sqrt{\alpha} A' \end{pmatrix} X \right\|_F^2 \\ s.t. \quad & \| x_i \|_0 \le T_0, \ i = 1, 2, \ldots, n \end{aligned} \tag{27}$$

where $\tilde{Y} = \left(Y^T \Omega^T, \sqrt{\mu/2}(Z^T - C_1^T/\mu)\right)^T$, $\tilde{D} = \left(R_1{}^T, \sqrt{\mu/2} R_2^T\right)$.

Adding an extra linear classifier term $\| H - WX \|_F^2$ into equation (27), the objective function of the joint linear classifier learning and DSNAR (JLC-DSNAR) model is

$$\begin{aligned} \min_{D', A', W, X} & \left\| \begin{pmatrix} \tilde{Y} \\ \sqrt{\gamma} L' \\ \sqrt{\theta} H \end{pmatrix} - \begin{pmatrix} \tilde{D} \\ \sqrt{\gamma} A' \\ \sqrt{\theta} W \end{pmatrix} X \right\|_F^2 \\ s.t. \quad & \| x_i \|_0 \le T_0, \ i = 1, 2, \ldots, n \end{aligned} \tag{28}$$

where θ is a regularization coefficient for the linear classifier term.

According to Theorem 3.1 in reference (Kviatkovsky et al. 2017), the solution of equation (28), i.e., $< \tilde{D}^*, A'^*, W^*, X^* >$, is exactly the solution of equation (27), when we provide a parameter setting of $\alpha = \gamma + \theta/l$ and an identical initialization of associated variables. Especially, we can conclude a reshuffled form of A'^* as

$$A'^* = \Big(\underbrace{W^{*T}, \ldots, W^{*T}}_{\times l}\Big)^T \tag{29}$$

In this way, we prove that the extra linear classifier term is redundant compared to the latent structural transformation term for our proposed DSNAR model. Moreover, if we learn $l = k/c$ atoms for each one of the c classes, the latent structural transformation term is exactly the sparse-code error term in (Du et al. 2021). This indicates that the latent structural

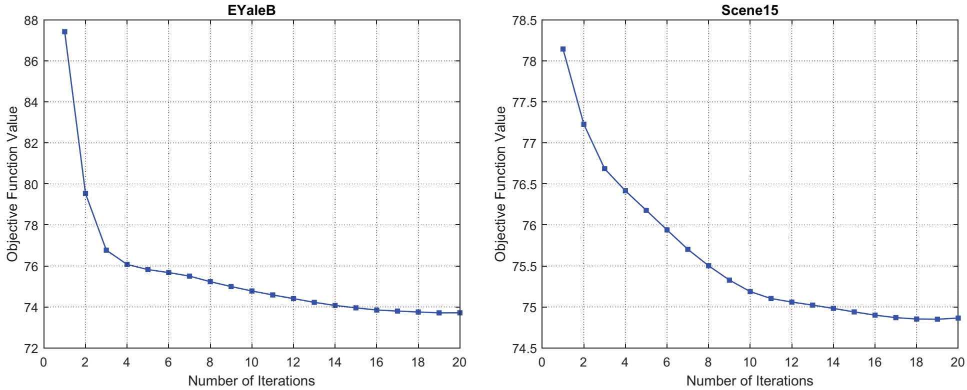

Figure 2. Convergence curves of the DSNAR model on the EYaleB and Scene15 datasets.

transformation term provides enough discrimination for supervised class label information and guarantees robustness by avoiding accidental error predictions. Additional supervised label embedding terms under identical initialization conditions will complicate parameter tuning with more regularization coefficients and deteriorate training performance due to a larger input matrix of the K-SVD method.

Complexity and Convergence Analysis

It is noteworthy that our DSNAR model in Algorithm 1 consists of three updating rules, which are computationally efficient. For the updating of variable X, the main cost lies in calculating the rank-one approximation SVD of $\tilde{E}_f \in R^{2k*n}$, with complexity $O(kn(2k)^2)$. For the updating of variable Ω, the inverse operation of $(YY^T + \beta I)^{-1}$ in the closed-form solution scales as $O(m^3)$. For the updating of variable S, the computational cost of sample-wise multiplication can be accomplished in $O(n^2)$. As a whole, the overall complexity of training costs $O(nk^3 + m^3 + n^2)$ in one iteration.

In addition, although our DSNAR model is not jointly convex due to the l_0 norm constraint, it is a convex minimization problem for solving each variable in Algorithm 1. Therefore, the value of objective function is non-increasing until Algorithm 1 converges. The convergence curves of our DSNAR model on the Extended YaleB (EYaleB) and Scene15 datasets with the best parameter settings are plotted in Figure 2. As can be seen, a satisfactory solution can be obtained by the decreasing of the objective function within 20 iterations.

Robust Classification Approach

To avoid accidental incorrect predictions caused by noisy information, we propose a robust classification scheme in latent label space. First, we predefine a desired latent space label matrix $L' \in R^{k \times k}$, in which the extended label vector for each class as $\mathbf{l}_j \in L'(1 \leq j \leq c)$. For a new test sample y_i, preliminary representation coefficient x_i is efficiently obtained via the operation of multiplying analysis Ω by the testing sample y_i. Keeping in view that the proposed method imposes sparsity and non-negativity on coefficients, we apply a hard thresholding operator $x_i = HT(\Omega y_i, T_0)$ to maintain the sparse and non-negative characteristic of x_i, where T_0 is an adjustable sparsity parameter for test samples in classification step. The operator reserves elements with T_0 biggest non-negative values and sets the others to be zero. We can easily obtain the latent space label vector for each testing sample by $\mathbf{lp}_i = Ax_i$.

Some previous work adopts k Nearest Neighbor (kNN) classifier in label space (Du et al. 2021) or structural characteristic of representation (Ling, Chen, and Wu 2020), as it does not require training effort and mainly depends on the distance metrics among coefficients. To ensure a robust classification, we utilize the kNN method in the latent label space. The classification of a test sample is determined by the category labels of its $k = 1$ nearest training samples. The category of each testing sample is determined robustly by contrasting each obtained label vector $\mathbf{lp}_i$ with its nearest column vector $\mathbf{l}_j$ in the latent space label matrix L'. As the predicted vector is determined by neighborhood assignment of an extended label space transformation vector, the classification scheme prevents the incorrect by one element's location in label vector.

Experiments

Experimental Settings

In this section, we extensively evaluate our DSNAR model on five benchmark datasets, including Extended Yale B (EYaleB), AR, LFW, Scene15 and UCF50. The above datasets are widely used in evaluating the performance of sparse representation-based classification methods. Some of these hand-crafted features are publicly provided by (Jiang, Lin, and Davis 2013) and (Sadanand and Corso 2012). We randomly select the training and testing sets with fixed proportions that is consistent with previous paper (Li et al. 2017). The EYaleB face dataset contains in total 2414 frontal face images of 38 persons under various illumination and expression conditions. Each image is manually cropped and resized to 192×168 pixels. The AR face dataset contains more illumination, expression, and occlusions variations. We choose a subset consisting of 2600 face images from 50 males and 50 females. The original images were cropped to 165×120 pixels. On EYaleB and AR face datasets, random face features are extracted by the projection with a randomly generated matrix. The LFW face image dataset contains more than 13,000 images of faces collected from the Web. We selected a subset of the LFW face dataset consisting of 1251 images of 86 persons. Each image was converted into gray image and was manually cropped to 32 × 32 pixels. On Scene15 datasets, there are 200–400 images with approximate average size of 250 × 300 pixels in each category. The features are achieved by extracting SIFT descriptors, max pooling in spatial pyramid and reducing dimensions via PCA for convenient processing. UCF50 is a large-scale and challenging action recognition dataset. It has 50 action categories and 6680 realistic human action videos collected from YouTube. The features of the benchmark datasets are provided by previous study (Jiang, Lin, and Davis 2013; Li et al. 2017), and the statistics are summarized in Table 2.

We compare the proposed DSNAR model with some state-of-the-art DL approaches by classification on the hand-crafted features. The comparison models include SRC (Wright et al. 2009), D-KSVD (Zhang and Li 2010), LC-KSVD (Jiang, Lin, and Davis 2013), ADL+SVM (Shekhar, Patel, and Chellappa 2014), SK-ADL (Wang et al. 2018), SLC-ADL (Wang et al. 2017) and SDADL (Du et al. 2021). We use the K-SVD box to train the synthesis dictionary (Aharon, Elad, and Bruckstein 2006), and the ideal structural representation to initialize the analysis representation (Wang et al. 2017). In addition, the D-KSVD, LC-KSVD, SK-ADL models adopt the linear classification method, the SLC-ADL and SDADL models use the kNN classification method in label space, and our DSNAR model conducts robust classification with kNN method in latent space. For fair comparison, we use the released codes of all these models and finely tune the parameters, or directly adopt the results reported in the literatures with the same parameter settings. There are four parameters in the proposed DSNAR model, i.e., α, β, γ and δ, where γ and δ are set to $3e-3$ and $1e-6$ empirically to avoid overfitting and singular values. The dominant parameters α and β, which weigh the two regularizers in objective function, are obtained by cross validation. The iteration number of K-SVD algorithm is set as 50 for all associated models. The sparsity is set as 45 in all the methods, and the dictionary atom is set around 600 which is the integral multiple of the number of classes in different datasets. All experiments are run with MATLAB R2019a under Windows10 on a PC with Intel Core i5–8400 2.80 GHz 2.81 GHz CPU and 8 GB RAM. We repeat the experiments 5 times on randomly selected training and testing samples, and the mean accuracies and testing time are reported.

Table 2. The statistical information of the benchmark datasets and features.

Datasets	Feature/dimension	Instances(total/use)	Class(total/use)	Train/test
EYaleB	Random face/504	64/64	38/38	20/44
AR	Random face/540	26/26	126/100	10/16
LFW	Cropped face/1024	11–20/all	86/86	8/rest
Scene15	SIFT/3,000	200–400/all	15/15	100/rest
UCF50	SIFT/3,000	31–800/all	102/102	30/rest

Results and Analysis

Table 3 shows the mean classification accuracy results on different datasets. As can be seen, our method achieves notably higher accuracy than SRC, D-KSVD and LC-KSVD on all five datasets. This is mainly due to the structured non-negative representation ability and discrimination ability achieved in our method for ADL model, which is a further improvement on the D-KSVD and LC-KSVD that only consider the classification error term and label-consistent term on SDL model. The SRC model that directly uses all training samples as the dictionary will introduce noisy and redundant samples for sparse representation. The performance of D-KSVD is not stable compared to LC-KSVD on different datasets due to the noisy and redundant information. The non-negative constraint in the proposed DSNAR model can boost the selection of representative analysis atoms, which enhances the representational ability of homogeneous samples while weakening the negative effects caused by heterogeneous samples, and this may help to overcome the above noise disadvantage.

Our proposed DSNAR method also achieves favorable results compared with all four ADL-based methods. The ADL+SVM model neglects the discrimination ability during training procedure, and the learned representation shows poor classification performance. The SK-ADL model integrates linear classification error term into the basic ADL framework, which can be seen as a special case of the latent structural transformation term in our DSNAR model. The performance improvements of the DSNAR model compared to SK-ADL method verifies the effectiveness of the off-block diagonal suppression term, especially for complex noisy and redundant data samples. The better performance improvement of SLC-ADL benefits from a robust synthesis linear classifier with class label information, and robust KNN classification scheme that is also adopted in our DSNAR model. The SDADL and DSNAR models both consider the class label information and structural constraint representation, which lead to better performance compared to other ADL-based models. But the differences of the two models lie in the easy-tuned parameters and the flexible well-structured representation, which leads to obvious classification accuracy improvement.

Table 3. Classification accuracy (%) comparison on different datasets. The best results are in bold.

DL Models	EYaleB	AR	LFW	Scene15	UCF50
SRC	96.5	97.5	38.1	91.8	68.4
D-KSVD	94.1	88.8	35.9	89.1	57.8
LC-KSVD	96.7	97.8	36.7	92.9	70.1
ADL+SVM	95.4	96.1	32.6	90.1	72.3
SK-ADL	96.7	97.7	38.7	97.4	74.6
SLC-ADL	97.0	97.2	38.9	97.5	78.1
SDADL	97.0	97.8	40.1	98.1	77.8
DSNAR	**97.3**	**98.1**	**40.4**	**98.7**	**78.5**

Tables 4 and 5 show the average time for training procedure and classification procedure of different models by computing the processing of one data sample. Since RBC models use the full set of training samples as the dictionary, we do not report their training time in efficiency comparison experiment. We choose some necessary DL models to validate the efficiency superiority of our proposed DSNAR model. For training comparison, we select D-KSVD and LC-KSVD to verify if the increased dimensionality adds computational complexity for K-SVD input. The SDADL and SK-ADL are selected to evaluate the efficiency of the multiplication of class-oriented subdictionary with training samples from other classes. For classification procedure, the ADL-based models are efficient through theoretical analysis, which only projects test sample to its sparse representation and obtains the label vector by some simple classification schemes.

As is depicted in Table 4, The DSNAR model achieves as efficient training performance as those shared dictionary models, and obviously better performance than SDADL model, due to the avoiding of the class-oriented multiplication of the analysis subdictionary with the corresponding complementary sample matrix. As can be seen in Table 5, the ADL-based methods is approximately one order of magnitude faster than LC-KSVD method. This mainly owns to the low classification complexity of ADL which uses efficient feature transformation and the robust structural constraints for compact representations. Also, our method performs slightly better performance than SK-ADL and comparable to SDADL, due to the robust and efficient kNN classifier in extended label space and the non-negative constraint on sparse coding, which limits the number of dictionary atoms in projection of heterogeneous samples.

Figure 3 shows the confusion matrix for the DSNAR method on Scene15 dataset. It presents proportion of images in each category classified to all categories. We can observe that most images can be classified into the right category, with some class even getting all right classification. From the tables

Table 4. The average time (ms) for training procedure of different models.

DL Models	EYaleB	AR	LFW	Scene15	UCF50
D-KSVD	48.9	39.2	43.9	61.3	45.8
LC-KSVD	76.5	52.5	85.0	93.1	91.6
SK-ADL	38.4	40.0	48.3	44.4	43.3
SDADL	62.2	65.6	99.5	65.5	74.4
DSNAR	35.5	32.9	44.5	37.1	43.2

Table 5. The average time (ms) for classification procedure of different models.

DL Models	EYaleB	AR	LFW	Scene15	UCF50
SRC	39.93	41.24	38.12	70.66	92.37
LC-KSVD	0.426	0.442	0.351	0.675	0.794
SK-ADL	0.029	0.078	0.052	0.113	0.169
SDADL	0.023	0.063	0.051	0.100	0.158
DSNAR	0.024	0.063	0.050	0.105	0.157

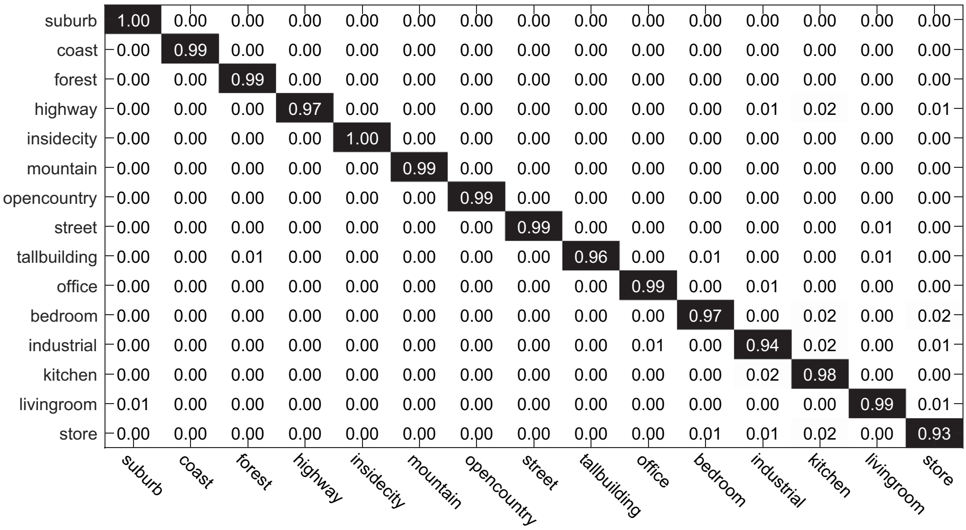

Figure 3. Confusion matrix of the ground truth on Scene 15 dataset.

and figures, we can conclude that the desired effect of our proposed DSNAR method for robust classification is reached.

7.3. Parameter Sensitivity

The dominant regularization coefficient α and β in the objective function are tuned by 5-fold cross validation and optimized by using grid search. We firstly search in the larger range of $[10^{-3}, 10^{-2}, \cdots, 10^{2}, 10^{3}]$ for each parameter and then search a smaller grid with proper interval size determined by preliminary classification results. In Figure 4, two 3D histograms show that the classification accuracy vary as parameter α and β change on AR and UCF50 datasets respectively. We can observe that the best performance is achieved at $\alpha = 60$ and $\beta = 20$ on AR dataset and $\alpha = 20$ and $\beta = 14$ on UCF50 dataset. The results are consistent with the intuitive view that both regularization terms are crucial for a discriminative and robust ADL model. The parameters we set in each dataset are listed in Table 6.

Table 7 shows the average classification accuracy of different DL algorithms with different numbers of atoms on the LFW dataset. The other parameters such as dataset segmentation, regularization coefficients and sparsity threshold remain the same as the best parameter settings in the above model comparison experiments. As can be seen, the average accuracy of the DSNAR model significantly outperformed other approaches as the number of atoms increased. When the atom number becomes larger, its performance is roughly stable for ADL models. For SDL models, such as LC-KSVD model, the larger the number of dictionary atoms is, the more negative effect on discriminative dictionary learning it has, as dictionary atoms may be contaminated by more and more noisy or redundant

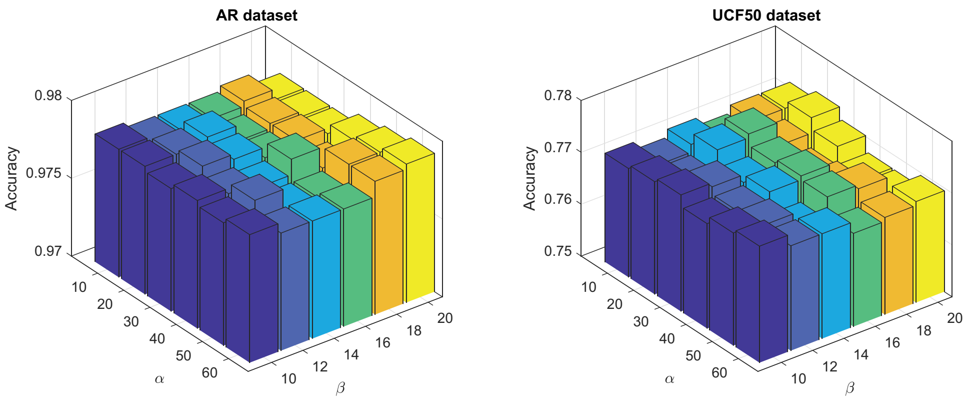

Figure 4. Results of parameter selection of α and β on AR and UCF50 datasets.

Table 6. Best parameter settings for DSNAR model by cross validation.

Parameters	EYaleB	AR	LFW	Scene15	UCF50
α	50	60	4	20	20
β	12	20	1	14	11

Table 7. Classification accuracy (%) with different numbers of atoms on the LFW dataset.

Atoms#	LC-KSVD	ADL+SVM	SK-ADL	SLC-ADL	DSNAR
86	29.3	27.9	29.6	31.4	32.3
172	34.1	29.7	34.3	34.6	35.7
258	34.8	30.3	36.1	36.8	37.5
344	35.4	31.1	37.0	37.9	38.9
430	36.9	31.7	38.2	38.6	39.9
516	36.7	32.6	38.7	38.9	40.4

samples. The analysis dictionary transforms the raw samples into representation space, and is somewhat more robust to noises or corruptions with the increase of atom number. The accuracies and time costs in tables demonstrate that our robust DSNAR model has a huge potential in classification tasks.

Conclusion

In this paper, we proposed a joint Dual-Structural constrained and Non-negative Analysis Representation (DSNAR) learning model for pattern classification under complex real-world data environment. To enhance the discrimination of analysis representation, a latent structural transformation term and off-block suppression term are jointly incorporated in ADL model for robust well-structured representation, which encourages the samples from the same class to share similar coefficients and those from different classes to have dissimilar coefficients. The non-negative constraint is also considered to prevent substraction of contributions from heterogeneous samples, which can further relieve the negative effect of noisy and redundant samples. Moreover, we designed a robust classification scheme in latent space to achieve better classification performance. The objective variables in our optimization problem are updated alternatively by K-SVD method, iterative re-weighted method and gradient method. Extensive experiments on five benchmark datasets demonstrate that the proposed DSNAR model clearly outperforms the existing state-of-the-art ADL models.

The existing ADL methods ignore the diversity of samples when evaluating representation reconstruction. That is to say, all samples are equally considered and assigned the same weight, though some are contaminated by noisy information. In the future, we will study the adaptive weighted feature learning (Yang et al. 2013; Zheng et al. 2017) to measure the importance and relevance of features, and ordinal locality topology preserving (Li et al. 2019; Luo et al. 2011) to characterize the exact ordinal similarity scores for data samples in a local manifold area.

Acknowledgements

This work is supported by the Natural Science Basic Research Program of Shaanxi, China (Program No. 2021JM-339).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the Natural Science Basic Research Program of Shaanxi Province [2021JM-339].

ORCID

Kun Jiang http://orcid.org/0000-0003-1316-5237

Data availability statement

All data generated or analyzed during this study are included in this published article.

References

Aharon, M., M. Elad, and A. Bruckstein. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54 (11):4311–22. doi:10.1109/TSP.2006.881199.

Bishop, C. 2006. *Pattern recognition and machine learning*. Springer. Retrieved from https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

Chen, S. S., D. L. Donoho, and M. A. Saunders. 2001. Atomic decomposition by basis pursuit. *SIAM Review* 43 (1):129–59. doi:10.1137/S003614450037906X.

Chen, Z., X. -J. Wu, Y. -H. Cai, and J. Kittler. 2021. Sparse non-negative transition subspace learning for image classification. *Signal Processing* 183:107988. doi:10.1016/j.sigpro.2021.107988.

Chen, Z., X. -J. Wu, and J. Kittler. 2022. Relaxed block-diagonal dictionary pair learning with locality constraint for image recognition. *IEEE Transactions on Neural Networks and Learning Systems* 33 (8): 3645–59. doi:10.1109/TNNLS.2021.3053941.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3):273–97. doi:10.1007/BF00994018.

Cover, T., and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13 (1):21–27. doi:10.1109/TIT.1967.1053964.

Davis, G., S. Mallat, and M. Avellaneda. 1997. Adaptive greedy approximations. *Constructive Approximation* 13 (1):57–98. doi:10.1007/BF02678430.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern classification*, 2nd ed., Hoboken, NJ, USA: John Wiley & Sons.

Du, H., Y. Zhang, L. Ma, and F. Zhang. 2021. Structured discriminant analysis dictionary learning for pattern classification. *Knowledge-Based Systems* 216:106794. doi:10.1016/j.knosys.2021.106794.

Engan, K., S. Aase, and J. Hakon Husoy. 1999. Method of optimal directions for frame design. In *1999 ieee international conference on acoustics, speech, and signal processing*, Phoenix, Arizona, USA, (p. 2443–46).

Guo, J., Y. Guo, X. Kong, M. Zhang, and R. He. 2016. Discriminative analysis dictionary learning. In D. Schuurmans and M. P. Wellman (ed.), *Proceedings of the thirtieth AAAI conference on artificial intelligence*, Phoenix, Arizona, USA, (p. 1617–23). AAAI Press.

Gu, S., L. Zhang, W. Zuo, and X. Feng. 2014. Projective dictionary pair learning for pattern classification. In *2014 International Conference on Neural Information Processing Systems*, Montreal, Canada, ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger.

Hawe, S., M. Kleinsteuber, and K. Diepold. 2013. Analysis operator learning and its application to image reconstruction. *IEEE Transactions on Image Processing* 22 (6):2138–50. doi:10.1109/TIP.2013.2246175.

Hyvärinen, A. 1999. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation* 11 (7):1739–68. doi:10.1162/089976699300016214.

Jiang, Z., Z. Lin, and L. S. Davis. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (11):2651–64. doi:10.1109/TPAMI.2013.88.

Jiang, K., C. Zhao, L. Zhu, and Q. Sun. 2022. Class-oriented and label embedding analysis dictionary learning for pattern classification. *Multimedia Tools and Applications* Retrieved from. doi: 10.1007/s11042-022-14295-9.

Kong, S., and D. Wang. 2012. A dictionary learning approach for classification: Separating the particularity and the commonality. In *Computer vision – eccv 2012*, ed. A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, 186–99. Berlin, Heidelberg: Springer.

Kviatkovsky, I., M. Gabel, E. Rivlin, and I. Shimshoni. 2017. On the equivalence of the lc-ksvd and the d-ksvd algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2):411–16. doi:10.1109/TPAMI.2016.2545661.

Li, Z., Z. Lai, Y. Xu, J. Yang, and D. Zhang. 2017. A locality-constrained and label embedding dictionary learning algorithm for image classification. *IEEE Transactions on Neural Networks and Learning Systems* 28 (2):278–93. doi:10.1109/TNNLS.2015.2508025.

Ling, J., Z. Chen, and F. Wu. 2020. Class-oriented discriminative dictionary learning for image classification. *IEEE Transactions on Circuits and Systems for Video Technology* 30 (7):2155–66. doi:10.1109/TCSVT.2019.2918852.

Li, Q., W. Zhang, Y. Bai, and G. Wang. 2022. A non-convex piecewise quadratic approximation of ℓ0 regularization: theory and accelerated algorithm. *Journal of Global Optimization*. doi:10.1007/s10898-022-01257-6.

Li, Z., Z. Zhang, J. Qin, S. Li, and H. Cai. 2019. Low-rank analysis-synthesis dictionary learning with adaptively ordinal locality. *Neural Networks* 119:93–112. doi:10.1016/j.neunet.2019.07.013.

Li, Z., Z. Zhang, S. Wang, R. Ma, F. Lei, and D. Xiang. 2021. Structured analysis dictionary learning based on discriminative Fisher pair. *Journal of Ambient Intelligence and Humanized Computing*. doi:10.1007/s12652-021-03262-1.

Lu, C., J. Feng, Z. Lin, and S. Yan (2013). Correlation adaptive subspace segmentation by trace lasso. In *2013 ieee international conference on computer vision*, Sydney, NSW, Australia, (p. 1345–52).

Luo, D., C. Ding, F. Nie, and H. Huang. 2011. Cauchy graph embedding. In *2011 International Conference on International Conference on Machine Learning*, Bellevue, Washington, USA, (p. 553–60).

Mairal, J., M. Elad, and G. Sapiro. 2008. Sparse representation for color image restoration. *IEEE Transactions on Image Processing* 17 (1):53–69. doi:10.1109/TIP.2007.911828.

Mallat, S., and Z. Zhang. 1993. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41 (12):3397–415. doi:10.1109/78.258082.

Nie, F., D. Xu, I.W. -H. Tsang, and C. Zhang. 2010. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing* 19 (7):1921–32. doi:10.1109/TIP.2010.2044958.

Ravishankar, S., and Y. Bresler. 2013. Learning sparsifying transforms. *IEEE Transactions on Signal Processing* 61 (5):1072–86. doi:10.1109/TSP.2012.2226449.

Rubinstein, R., T. Peleg, and M. Elad. 2013. Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model. *IEEE Transactions on Signal Processing* 61 (3):661–77. doi:10.1109/TSP.2012.2226445.

Sadanand, S., and J. J. Corso (2012). Action bank: A high-level representation of activity in video. In *2012 ieee conference on computer vision and pattern recognition*, Providence, RI, USA, (p. 1234–41).

Shekhar, S., V. M. Patel, and R. Chellappa (2014). Analysis sparse coding models for image-based classification. In *2014 ieee international conference on image processing (icip)*, Paris, France, (p. 5207–11).

Tang, W., A. Panahi, H. Krim, and L. Dai. 2019. Analysis dictionary learning based classification: Structure for robustness. *IEEE Transactions on Image Processing* 28 (12):6035–46. doi:10.1109/TIP.2019.2919409.

Wang, Q., Y. Guo, J. Guo, and X. Kong. 2018. Synthesis k-svd based analysis dictionary learning for pattern classification. *Multimedia Tools and Applications* 77:17023–41. doi:10.1007/s11042-017-5269-6.

Wang, J., Y. Guo, J. Guo, M. Li, and X. Kong. 2017. Synthesis linear classifier based analysis dictionary learning for pattern classification. *Neurocomputing* 238:103–13. doi:10.1016/j.neucom.2017.01.041.

Wang, J., Y. Guo, J. Guo, X. Luo, and X. Kong. 2017. Class-aware analysis dictionary learning for pattern classification. *IEEE Signal Processing Letters* 24 (12):1822–26. doi:10.1109/LSP.2017.2734860.

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (2):210–27. doi:10.1109/TPAMI.2008.79.

Xu, J., W. An, L. Zhang, and D. Zhang. 2019. Sparse, collaborative, or nonnegative representation: Which helps pattern classification? *Pattern Recognition* 88:679–88. doi:10.1016/j.patcog.2018.12.023.

Yang, M., H. Chang, W. Luo, and J. Yang. 2017. Fisher discrimination dictionary pair learning for image classification. *Neurocomputing* 269:13–20. doi:10.1016/j.neucom.2016.08.146.

Yang, J., K. Yu, Y. Gong, and T. Huang (2009). Linear spatial pyramid matching using sparse coding for image classification. In *2009 ieee conference on computer vision and pattern recognition*, Miami, FL, USA, (p. 1794–801).

Yang, M., L. Zhang, X. Feng, and D. Zhang (2011). Fisher discrimination dictionary learning for sparse representation. In *2011 international conference on computer vision*, Barcelona, Spain, (p. 543–50).

Yang, M., L. Zhang, J. Yang, and D. Zhang. 2013. Regularized robust coding for face recognition. *IEEE Transactions on Image Processing* 22 (5):1753–66. doi:10.1109/TIP.2012.2235849.

Yi, Y., J. Wang, W. Zhou, C. Zheng, J. Kong, and S. Qiao. 2020. Non-negative matrix factorization with locality constrained adaptive graph. *IEEE Transactions on Circuits and Systems for Video Technology* 30 (2):427–41. doi:10.1109/TCSVT.2019.2892971.

Zhang, Z., W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, and S. Yan. 2018. Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier. *IEEE Transactions on Neural Networks and Learning Systems* 29 (8):3798–814. doi:10.1109/TNNLS.2017.2740224.

Zhang, Q., and B. Li (2010). Discriminative k-svd for dictionary learning in face recognition. In *2010 ieee computer society conference on computer vision and pattern recognition*, San Francisco, CA, USA, (p. 2691–98).

Zhang, Z., Y. Sun, Y. Wang, Z. Zhang, H. Zhang, G. Liu, and M. Wang. 2021. Twin-incoherent self-expressive locality-adaptive latent dictionary pair learning for classification. *IEEE Transactions on Neural Networks and Learning Systems* 32 (3):947–61. doi:10.1109/TNNLS.2020.2979748.

Zhang, Z., Y. Xu, L. Shao, and J. Yang. 2018. Discriminative block-diagonal representation learning for image recognition. *IEEE Transactions on Neural Networks and Learning Systems* 29 (7):3111–25. doi:10.1109/TNNLS.2017.2712801.

Zhang, L., M. Yang, and X. Feng (2011). Sparse representation or collaborative representation: Which helps face recognition? In *Proceedings of the 2011 international conference on computer vision* (p. 471–78). USA: IEEE Computer Society.

Zhang, S., H. Yao, H. Zhou, X. Sun, and S. Liu. 2013. Robust visual tracking based on online learning sparse representation. *Neurocomputing* 100:31–40. Special issue: Behaviours in video. doi:10.1016/j.neucom.2011.11.031.

Zheng, J., P. Yang, S. Chen, G. Shen, and W. Wang. 2017. Iterative re-constrained group sparse face recognition with adaptive weights learning. *IEEE Transactions on Image Processing* 26 (5):2408–23. doi:10.1109/TIP.2017.2681841.

Zhuang, L., H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu (2012). Non-negative low rank and sparse graph for semi-supervised learning. In *2012 ieee conference on computer vision and pattern recognition*, Providence, RI, USA, (p. 2328–35).