Pattern Recognition 139 (2023) 109450

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Discriminative subspace learning via optimization on Riemannian manifold

Wanguang Yin[a], Zhengming Ma[b], Quanying Liu[a,*]

[a] Shenzhen Key Laboratory of Smart Healthcare Engineering, Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[b] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 12 April 2022
Revised 2 February 2023
Accepted 16 February 2023
Available online 24 February 2023

Keywords:
Discriminative subspace learning
Riemannian manifold optimization
Dimensionality reduction
Classification

ABSTRACT

Discriminative subspace learning is an important problem in machine learning, which aims to find the maximum separable decision subspace. Traditional Euclidean-based methods usually use Fisher discriminant criterion for finding an optimal linear mapping from a high-dimensional data space to a lower-dimensional subspace, which hardly guarantee a quadratic rate of global convergence and suffers from the singularity problem. Here, we propose the manifold optimization-based discriminant analysis (MODA) which is constructed by using the latent subspace alignment and the geometry of objective function with orthogonality constraint. MODA is solved by using Riemannian version of trust-region algorithm. Experimental results on various image datasets and electroencephalogram (EEG) datasets show that MODA achieves the best separability and is significantly superior to the competing algorithms. Especially for the time series of EEG signals, the accuracy of MODA is 20–30% higher than existing algorithms. The code for MODA is available at https://github.com/ncclabsustech/MODA-algorithm.

1. Introduction

Extracting statistically significant features or patterns from high-dimensional data satisfying maximal separability of a lower-dimensional subspace is a long-standing research topic in machine learning. This is called the *discriminative subspace learning (DSL)*. DSL has been widely used in many practical applications, such as computer vision [18], signal processing [33], and biomedical engineering [12]. Previous studies on this topic are mostly based on the Euclidean space for feature space transformation.

As a consequence, a variety of supervised learning algorithms have been proposed to learn the discriminative subspace by computing the distance or similarity among the extracted features from input data, including the linear discriminant analysis (LDA) [22], general tensor discriminant analysis (GTDA) [33], and uncorrelated multilinear discriminant analysis (UMLDA) [20]. However, the performance of these algorithms largely depends on the strategy to search the discriminative subspace on Euclidean space, and most existing supervised algorithms for subspace learning are linear, which learn a linear embedding to the discriminative subspace based on Fisher discriminant criterion (i.e., minimizing of the ratio between within-class and between-class variances). These methods suffer from two major issues. The first issue is that Fisher discriminative criterion may not capture the data distribution well. In particular, when the between-class scatter matrix in Fisher discriminant criterion is singular, these methods will fail. The second issue is that these methods assume the search space in Euclidean space and cannot guarantee a *quadratic rate* of global convergence [1]. Consequently, they easily get stuck in spurious local minima and hardly obtain the holistically optimal solution [5]. To address these issues, we aim to develop a solution for DSL based on Riemannian manifold optimization.

In this study, we propose the manifold optimization-based discriminant analysis (MODA) for discriminative subspace learning. Specifically, we construct the objective function by using the latent subspace alignment and exploit the geometry of the objective function with orthogonality constraint, and then solve the learning objective by using Riemannian trust-region algorithm. We compare the performance of MODA with state-of-the-art algorithms on

Abbreviations: CMDA, Constrained multilinear discriminant analysis; DATER, Discriminant analysis with tensor representation; DSL, Discriminative subspace learning; EEG, Electroencephalogram; HODA, High-order discriminant analysis; HOSVD, Higher-order singular value decomposition; GLTD, Graph-Laplacian tucker tensor decomposition; GTDA, General tensor discriminant analysis; MODA, Manifold optimization-based discriminant analysis; NTF, Nonnegative tensor factorization; NTD, Nonnegative Tucker decomposition; UMLDA, Uncorrelated multilinear discriminant analysis.

* Corresponding author.
E-mail addresses: yinwg@sustech.edu.cn (W. Yin), issmzm@mail.sysu.edu.cn (Z. Ma), liuqy@sustech.edu.cn (Q. Liu).

https://doi.org/10.1016/j.patcog.2023.109450
0031-3203/

image and electroencephalography (EEG) classification. Experimental results demonstrate that MODA is superior in finding the best discriminative subspace. The main contributions of this study are summarized as follows.

- MODA leverages the orthogonal constraint of the objective function and solves the learning objective by using Riemannian version of trust-region algorithm (Section 3). We formulate the objective function in a subtractive form instead of the division form (Eq. (8)), which can effectively avoid the calculation of inverse Hessian matrix and reduce the computational complexity.
- The objective function of MODA is constructed by the latent subspace alignment technique. Although it is constructed as a linear model, we adopt a nonlinear optimization strategy, which can obtain the overall optimal solution by using the equivalent relation of quotient space and has a great advantage in learning small sample datasets.
- The second-order geometry of Riemannian Hessian is used to solve the learning objective. The performance of MODA is compared with various state-of-the-art (SOTA) algorithms in classification task (Section 4.1) on both image datasets and EEG signals (Section 4.2). The experimental results on multiple image datasets (e.g., COIL20, ETH80, MNIST, USPS, CMU PIE) and EEG signals (e.g., MI, SSVEP) demonstrate that MODA can robustly obtain the higher performance than traditional algorithms of discriminative subspace learning.

2. Related work

2.1. Dimensionality reduction

Dimensionality reduction can be defined as searching for a low-dimensional subspace that preserves as closely as possible the intrinsic global or local structure of the input data [13,28]. Different scenarios have different preferences of which attributes should be retained in the dimension reduction process. For example, there may be a need to reduce the complexity of a system, enhance the interpretability of high-dimensional data, or preserve the nonnegativity of the input data. Generally speaking, most dimensionality reduction methods can be grouped into three categories: *matrix/tensor decomposition, manifold learning*, and *discriminative subspace learning*.

For matrix/tensor decomposition, principle component analysis (PCA) [36], independent component analysis (ICA) [6], and nonnegative matrix factorization (NMF) [16] are three classical algorithms for linear dimensionality reduction, which can simultaneously generate a set of the nested subspace of all possible dimensions. Specifically, PCA aims to find a subspace that can maximally retain the variance of the original data. ICA assumes that the sources of input data are mutually statistically independent, thereby it is widely used in blind source separation. To learn the parts-based representation of nonnegative objects like images and spectrum signals, NMF aims to find two nonnegative matrices whose product provides a good approximation to the input data. In the context of tensor data, multilinear PCA aims to find a tensor-to-tensor projection that maximally captures the variations of the input tensorial data [19], and multilinear ICA learns the statistically independent component of multiple factors [34]. The nonnegative tensor factorization (NTF) enforces the nonnegative conditions on the factor matrices of CP factorization [11], and nonnegative Tucker decomposition (NTD) enforces the nonnegative constraints on both projection matrices and core tensor, which can provide a better interpretation of the physical meaning of nonnegativity, such as energy, spectrum, and probability distribution [15,26].

Although these linear methods can work well if only such a linear subspace exists, they always fail to discover the nonlinear structure of data. To address such a problem, manifold learning is an effective approach for learning the nonlinear structure of high-dimensional data, which assumes that a set of geometrically related points lying on or close to the surface of a smooth low-dimensional manifold embedded in the ambient space [40]. The representative algorithms for manifold learning include the Laplacian Eigenmaps (LE) [2], locally linear embedding (LLE) [27], and locality preserving projections (LPP) [21]. To preserve the local similarity of high-dimensional data, LE constructs the similarity matrix of Laplacian graph by using the k-nearest search, and there naturally emerges a lot of graph learning algorithms [2,40]. Instead, LLE attempts to preserve the local linearity of the nearest neighbors by applying the affinity approximation, where the local neighborhoods of a point on a manifold can be approximated by an affinity subspace spanned by the k-nearest neighbors of that point [27]. Likewise, to preserve the local neighborhood structure of the data, LPP constructs a certain affinity graph by the data and preserves the local geometry of the original data. The main target of these algorithms is to preserve the underlying manifold-based geometrical relationship existing in the original data during the transformation, and a number of graph learning and manifold regularization techniques for nonlinear dimensionality reduction have been developed [28,40]. However, these algorithms are inherently unsupervised and non-discriminative. As a result, feature vectors belonging to different classes may not be optimally separated in the learned subspace.

In the context of classification, discriminative subspace learning is generally believed to be a more effective approach for learning the discriminative features, and linear discriminant analysis (LDA) is one of the most well-known algorithms to extract statistically significant features. LDA learns to discriminate different classes by computing the distance or similarity among the extracted features from the input data, and then assigns the test data to a specific class based on the measured distance and the learned threshold. Otherwise, considering the number of training samples is often less than the dimensionality of feature space, tensor-based discriminative subspace learning methods have been developed. For instance, discriminant analysis with tensor representation (DATER) aims to find a tensor to tensor projection while maximizing the tensor-based scatter ratio [37]. However, these algorithms do not always converge during its iterations. To provide a stable solution of convergences, the general tensor discriminant analysis (GTDA) learns a discriminant subspace with a tensor-to-tensor projection while maximizing the discriminant information in a low-dimensional subspace [33]. In contrast to the DATER, GTDA has good convergence and it is the first discriminative subspace learning algorithm that converges to a local solution. Considering that independence between the extracted features is a desirable property in many real-world applications, to this end, the uncorrelated multilinear discriminant analysis (UMLDA) is developed to extract the uncorrelated discriminative features directly from tensorial data, with an assumption that each class is represented by a single cluster [20]. To ensure the convergence within a limited number of iterations, the tensor rank-one discriminant analysis (TR1DA), which learns the projection subspace by repeatedly calculating the residues of the original data with the scatter difference criterion, and eventually leads to a set of rank-one projections [32]. Based on the multilinear structure of Tucker model, the high-order discriminant analysis (HODA) directly finds discriminative bases by using the tensorial Fisher criterion [25], and the constrained multilinear discriminant analysis (CMDA) which looks for an optimal tensor-to-tensor projection for discrimination in a lower-dimensional tensor subspace [17]. Theoretically, the value of the scatter ratio criterion in CMDA approaches its extreme value, if only it exists and with a bounded error. However, the performance of these algorithms heavily depends on the optimization strat-

egy to solve the Fisher discriminant criterion [10], which has not been explored by Riemannian manifold of nonlinear optimization and still faces a lot of challenge problems, regarding the singularity and instability of the within-class scatter. When the between-class scatter matrix in Fisher discriminant criterion is singular, there is no guarantee of the monotonicity for its objective function value [17,25]. This is due to the current discriminative subspace learning algorithms usually need to calculate the discriminant score, and the discriminant score requires the calculation of inverse covariance matrix [29]. To address such issues, we adopt two strategies: the first one is to formulate the discriminative subspace learning directly on the curved manifolds (e.g., Stiefel manifold or Grassmann manifold), instead of optimization on the flat Euclidean space with specific constraints [3,9,38], and the second one is to reformulate the discriminative model into a subtractive form.

2.2. Riemannian manifold optimization

Prior to presenting the general framework of MODA, we first introduce some basic terms and notations, such as the projection operator (i.e., $\mathrm{P}^t(\cdot)$), parallel transport (i.e., $\mathcal{T}_{x^{k-1}\to x^k}(\xi^{k-1})$), retraction, and geodesic. Recall that the *projection operator* is to project the embedded space (i.e., ambient space) to its tangent space by subtracting its component in the orthogonal complement of the tangent space (i.e., normal space $\mathcal{N}_x$). If the Riemannian manifold is a quotient manifold, we can further define a *projection operator* from the tangent space to the horizontal space by removing the component in the orthogonal complement of the horizontal space (i.e., vertical space $\mathcal{V}_x$). The *parallel transport* allows movements from a tangent space to another tangent space. The *retraction* is a mapping from the tangent space back onto the manifold. In Riemannian manifold optimization, retraction ensures that each update of optimization remains on the manifold, and *exponential retraction* is the most expensive retraction, which refers to the movement along a geodesic. *geodesic* which is defined as the curve with the minimal length connecting two points on the manifold. Table 1 lists the ingredients for Riemannian manifold optimization.

Motivated by the recent progress in discriminative subspace learning and Riemannian manifold optimization (e.g., solving partial least squares regression via manifold optimization approaches [5], and high order discriminant analysis based on Riemannian optimization [39]), we propose a novel algorithm, named manifold optimization-based discriminant analysis (MODA).

Table 1
Ingredients for Riemannian manifold optimization.

Notations	Descriptions
$\mathcal{M}$	A smooth manifold
$U \in \mathrm{St}(D,d)$	The elements of Stiefel manifold
$[U] \in \mathrm{Gr}(D,d)$	The elements of Grassmann manifold
$\mathcal{O}(d)$	The orthogonal group
$T_U\mathcal{M}$	Tangent space of a manifold $\mathcal{M}$ at U
ξ, η	Vector fields on a smooth manifold $\mathcal{M}$
$g_U(\xi,\eta)$	Riemannian metric for any $\xi, \eta \in T_U\mathcal{M}$
$\mathrm{P}^t(\cdot)$	Projection of $U \in \mathbb{R}^{D\times d}$ onto the tangent space
$\mathcal{T}_{x^{k-1}\to x^k}(\xi^{k-1})$	Parallel transport
∇	Levi–Civita onnnection
$\mathrm{grad} f(U)$	Riemannian gradient of f at U
$\mathrm{hess} f(U)[\xi]$	Riemannian Hessian of f at U evaluated for ξ
R_U	Retraction mapping
$\triangle$	Trust-region radius

3. Manifold optimization based discriminant analysis (MODA)

In this section, we present the manifold optimization-based discriminant analysis (MODA). The notations for Riemannian optimization used in this paper are listed in Table 1.

3.1. MODA model

The goal of MODA model is to minimize the reconstruction error from a high-dimensional space to a lower-dimensional space by $\mathrm{y} = U^T\mathrm{x}$, while maintaining the maximized discrimination between classes. Assume $\mathrm{x} \in \mathbb{R}^D$ is the input data with high dimension D, and $\mathrm{y} \in \mathbb{R}^d$ is the output with a lower dimension d. The number of samples $N = \sum_{c=1}^{C} N_c$, and N_c is the number of samples from cth class. $\bar{\mathrm{y}} = \frac{1}{N}\sum_n \mathrm{y}_n$ is the sample mean calculated from all samples, and $\bar{\mathrm{y}}_c = \frac{1}{n_c}\sum_n [\mathrm{y}_n | n \in C_c]$ is the sample mean calculated from cth class. $S_W = \left(X - \bar{X}_C\right)\left(X - \bar{X}_C\right)^T$ is a covariance matrix relative to the within-class scatter, and $S_B = \left(\bar{X}_C - \bar{X}\right)\left(\bar{X}_C - \bar{X}\right)^T$ is a covariance matrix relative to the between-class scatter. Therefore, the objective function of MODA (i.e., $f(U)$) is formulated to minimize the within-class scatter S_W and maximize the between-class scatter S_B, that is given by:

$$\begin{aligned}
&\min_U f(U)\\
&= \sum_{c=1}^{C}\sum_{n\in C_c} \|\mathrm{y}_n - \bar{\mathrm{y}}_c\|_F^2 - \sum_{c=1}^{C} N_c \|\bar{\mathrm{y}}_c - \bar{\mathrm{y}}\|_F^2\\
&= \sum_{c=1}^{C}\sum_{n\in C_c} \left\|U^T(\mathrm{x}_n - \bar{\mathrm{x}}_c)\right\|_F^2 - \sum_{c=1}^{C} N_c \left\|U^T(\bar{\mathrm{x}}_c - \bar{\mathrm{x}})\right\|_F^2\\
&= \left\|U^T\left(X - \bar{X}_C\right)\right\|_F^2 - \left\|U^T\left(\bar{X}_C - \bar{X}\right)\right\|_F^2\\
&= tr\left(U^T S_W U\right) - tr\left(U^T S_B U\right)\\
&\text{s.t. } U^T U = \mathrm{I}_d,
\end{aligned} \tag{1}$$

where U is the projection matrix, subjecting to the orthogonal constraint (i.e., $U^TU = \mathrm{I}_d$). Due to the geometric property of objective function with orthogonal constraint, we can transform the objective function in Eq. (1) from Euclidean space to Stiefel manifold. On other words, the constrained optimization problem in Euclidean space becomes a Riemannian manifold optimization, which can be solved by using Riemannian trust-region algorithm. The objective function on Stiefel manifold $U \in \mathrm{St}(D,d)$ can be formulated in Eq. (2).

$$\min_{U\in \mathrm{St}(D,d)} f(U) = tr\left(U^T S_W U\right) - tr\left(U^T S_B U\right). \tag{2}$$

Consider that Grassmann manifold $\mathrm{Gr}(D,d)$ is the set of all Stiefel manifold elements that are invariant to the right rotations [7], we can further formulate the objective function for optimization on Grassmann manifold as shown in Eq. (3).

$$\min_{[U]\in \mathrm{Gr}(D,d)} f(U) = tr\left(U^T S_W U\right) - tr\left(U^T S_B U\right), \tag{3}$$

where $[U] \in \mathrm{Gr}(D,d)$ is the element of Grassmann manifold, indicating the equivalence class of U. Following the equivalence relation defined by the orthogonal group $\mathcal{O}(d)$, Grassmann manifold $\mathrm{Gr}(D,d)$ is the quotient space (i.e., *each element is an equivalence class*) of Stiefel manifold [1].

3.2. The Riemannian trust-region algorithm

Analogous to the trust-region algorithm in the Euclidean space, the Riemannian trust-region algorithm is an extension of the classical unconstrained trust-region algorithm to the Riemannian manifolds with guaranteed *quadratic rate* convergence. It is suitable for large-scale optimization on Riemannian manifolds [24]. Each iteration consists of two stages: (1) the first is to approximate the solution of trust-region subproblem; (2) the second is to compute a new iterate based on the retracting mapping. The trust-region subproblem (i.e., $F(U)$ at $U \in \mathrm{Gr}(D,d)$) means to approximately minimize a *quadratic model* of the objective function within a trust-

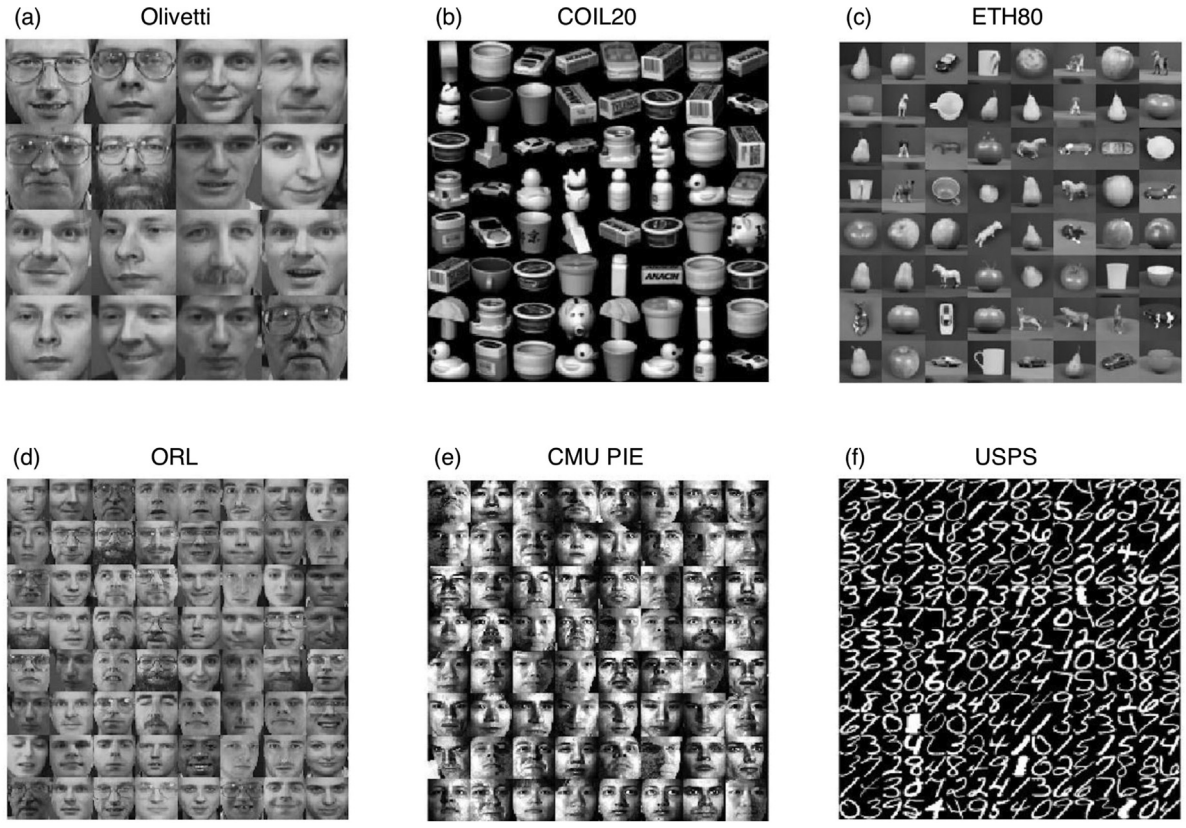

Fig. 1. Some examples from six datasets used in experiments. (a) Olivetti dataset. (b) COIL20 dataset. (c) ETH80 dataset. (d) ORL dataset. (e) CMU PIE dataset. (f) USPS dataset.

region radius, formulated as follows.

$$\begin{aligned} &\min_{\xi \in T_U \mathrm{Gr}(D,d)} F(U) \\ &= f(U) + g_U(\mathrm{grad} f(U), \xi) + \tfrac{1}{2} g_U(\mathrm{hess} f(U)[\xi], \xi) \\ &\text{s.t. } \|\xi\| \le \triangle, \end{aligned} \tag{4}$$

where $\triangle$ is the trust-region radius, and $\|\xi\| = \sqrt{g_U(\xi, \xi)}$. $\mathrm{grad} f(U)$ denotes the Riemannian gradient, and $\mathrm{hess} f(U)[\xi]$ is the Riemannian Hessian of objective function $f(U)$ along a tangent vector ξ. Then, it includes three steps: (1) the first step is to compute the new iterate by retracting mapping (i.e., $U^{k+1} := \mathrm{R}_{Uk}(\xi)$); (2) the second step is to update the new iterate by rejecting or accepting U^{k+1} depends on its quality; (3) the third step is to update the trust-region radius $\triangle$. Each of these steps is defined on the tangent space $T_U\mathcal{M}$ of a manifold $\mathcal{M}$ (e.g., Stiefel manifold or Grassmann manifold) at U, and is based on the *Riemannian gradient* and *Riemannian Hessian*. Thus, in order to apply the Riemannian trust-region algorithm, we need to derive the Riemannian gradient (i.e., $\mathrm{grad} f(U)$) and Riemannian Hessian (i.e., $\mathrm{hess} f(U)[\xi]$) in the following subsection.

3.3. Calculation of Riemannian gradient and Hessian

To compute the Riemannian Hessian of objective function, it needs to be noted that *Riemannian connection*, which is an important notion that intimately relevant to the Riemannian Hessian, and Levi–Civita connection (i.e., $\nabla_\xi \eta$) is a unique affine connection used to define the Riemannian Hessian of the objective function [1]. Given an example, the covariant derivative of $D\mathrm{Grad} f(U)[\xi]$ is the Euclidean directional derivative of Euclidean gradient $\mathrm{Grad} f(U)$ along the direction of tangent vector ξ on a manifold. We reformulate the objective function in Eq. (3) as

$$f(U) = tr\left(U^T (S_W - S_B) U\right). \tag{5}$$

Consider that Stiefel manifold (i.e., $U \in \mathrm{St}(D, d)$) is a submanifold of the matrix Euclidean space (i.e., $U \in \mathbb{R}^{D \times d}$), and Riemannian metric (i.e., $g_U : T_U\mathcal{M} \times T_U\mathcal{M} \to \mathbb{R}$) of $\mathcal{M}$ can be defined by the inner product:

$$g_U(\xi, \eta) := tr\left(\xi^T \eta\right). \tag{6}$$

Furthermore, the Euclidean gradient (i.e., $\mathrm{Grad} f$) is defined by the gradient of f on $\mathcal{M}$ with respect to the endowed metric, and we can obtain the Euclidean gradient of objective function Eq. (5) as follows,

$$\mathrm{Grad} f(U) := 2 S_W U - 2 S_B U. \tag{7}$$

For the computational space can be decomposed into two complementary spaces (i.e., the tangent space and normal space), thus the Riemannian gradient (i.e., $\mathrm{grad} f(U)$) of objective function can be obtained by using the orthogonal projection of Euclidean gradient (i.e., $\mathrm{Grad} f(U)$) onto the tangent space of Riemannian manifold, which is shown in Eq. (8),

$$\begin{aligned} \mathrm{grad} f(U) &= \mathrm{P}_U^t(\mathrm{Grad} f(U)) \\ &= \mathrm{Grad} f(U) - U \mathrm{sym}\left(U^T \mathrm{Grad} f(U)\right), \end{aligned} \tag{8}$$

where $\mathrm{sym}(X) = \left(X + X^T\right)/2$ denotes to extract the symmetric part of a square matrix X. The Euclidean Hessian (i.e., $\mathrm{Hess} f(U)$) of f at U is defined through the covariant derivative with respect to the Levi–Civita connection ∇ on $\mathcal{M}$ by:

$$\begin{aligned} \mathrm{Hess} f(U)[\xi] &:= D\nabla f(U)[\xi] \\ &= 2 S_W \xi - 2 S_B \xi. \end{aligned} \tag{9}$$

Analogously, Riemannian Hessian (i.e., $\mathrm{hess} f(U)[\xi]$) equals to the Euclidean Hessian followed by the orthogonal projection onto the tangent space, as shown in Eq. (10),

$$\mathrm{hess} f(U)[\xi] = \mathrm{P}_U^t(\mathrm{Hess} f(U)[\xi]). \tag{10}$$

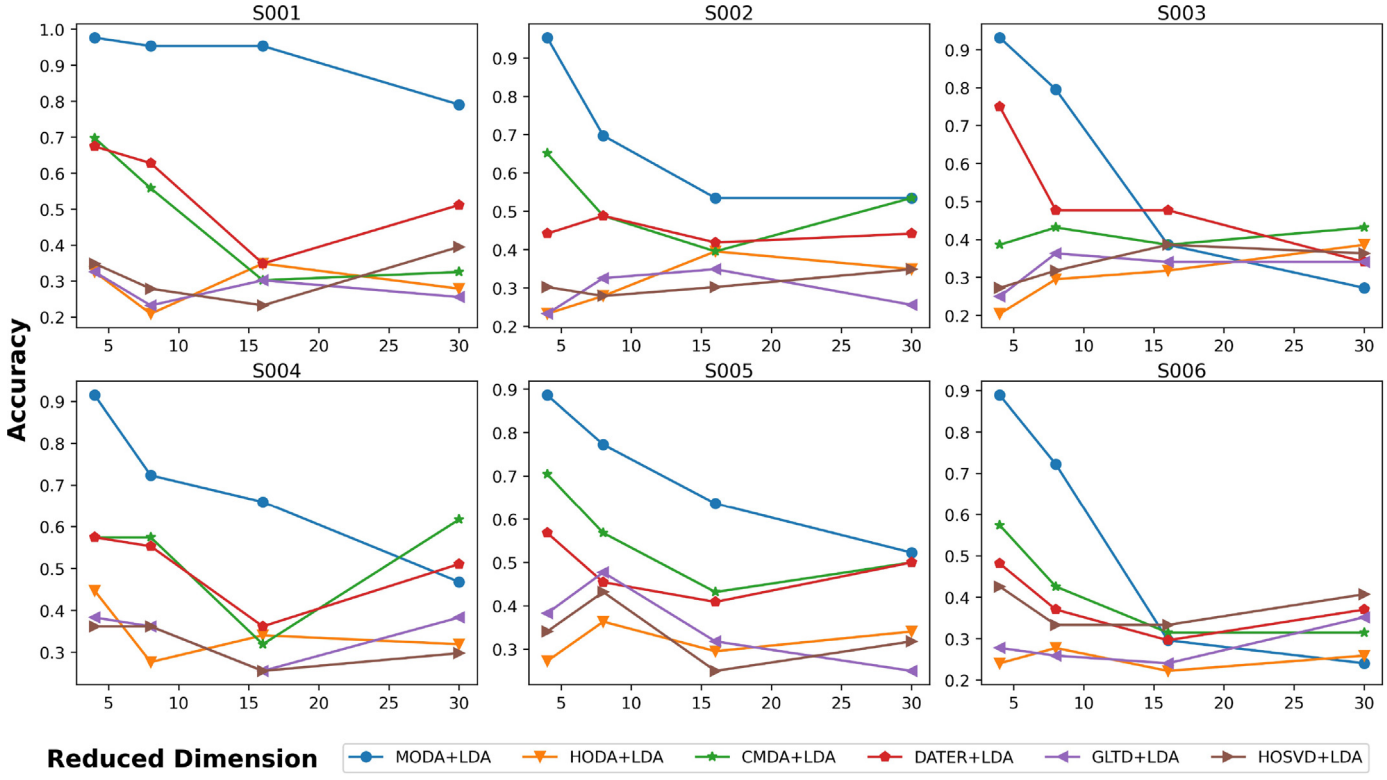

Fig. 2. The classification accuracy using LDA classifier on the EEG MI dataset.

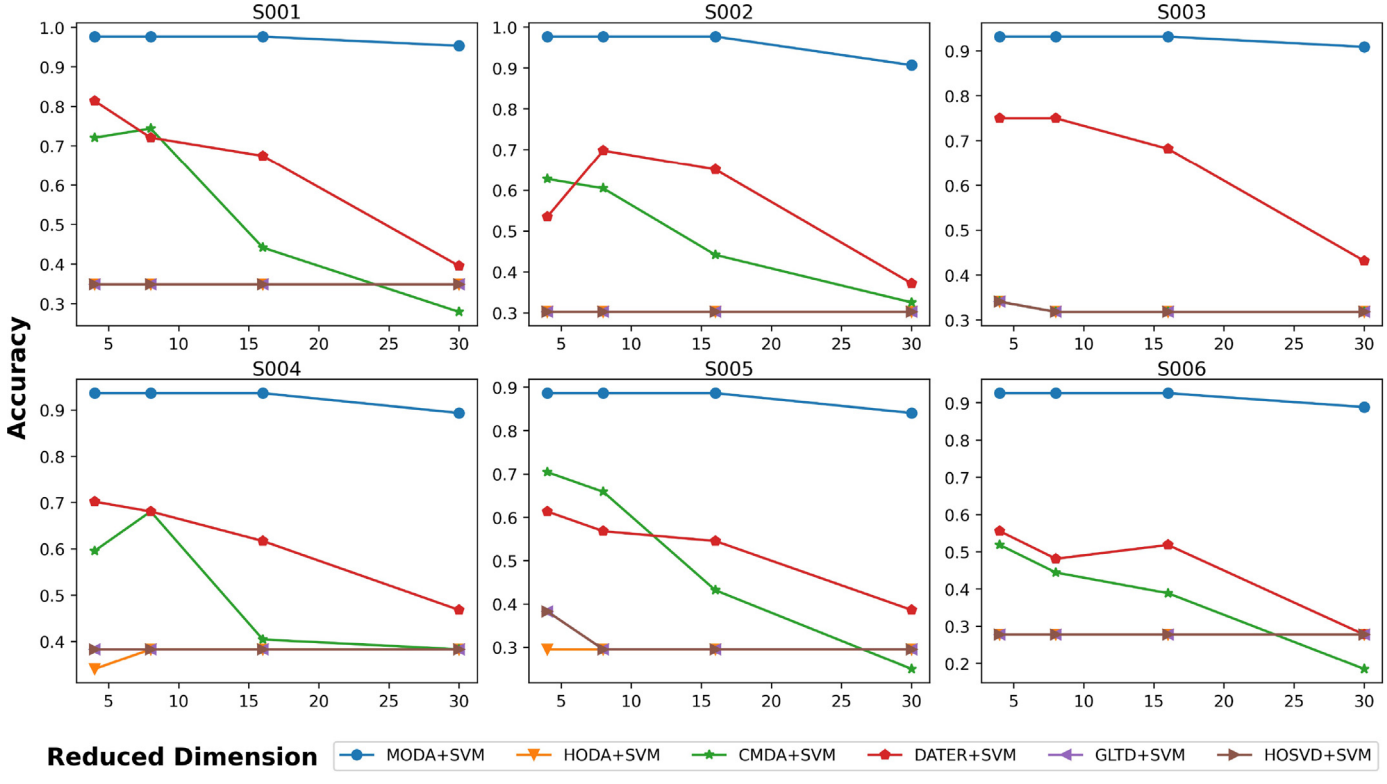

Fig. 3. The classification accuracy using SVM classifier on the EEG MI dataset.

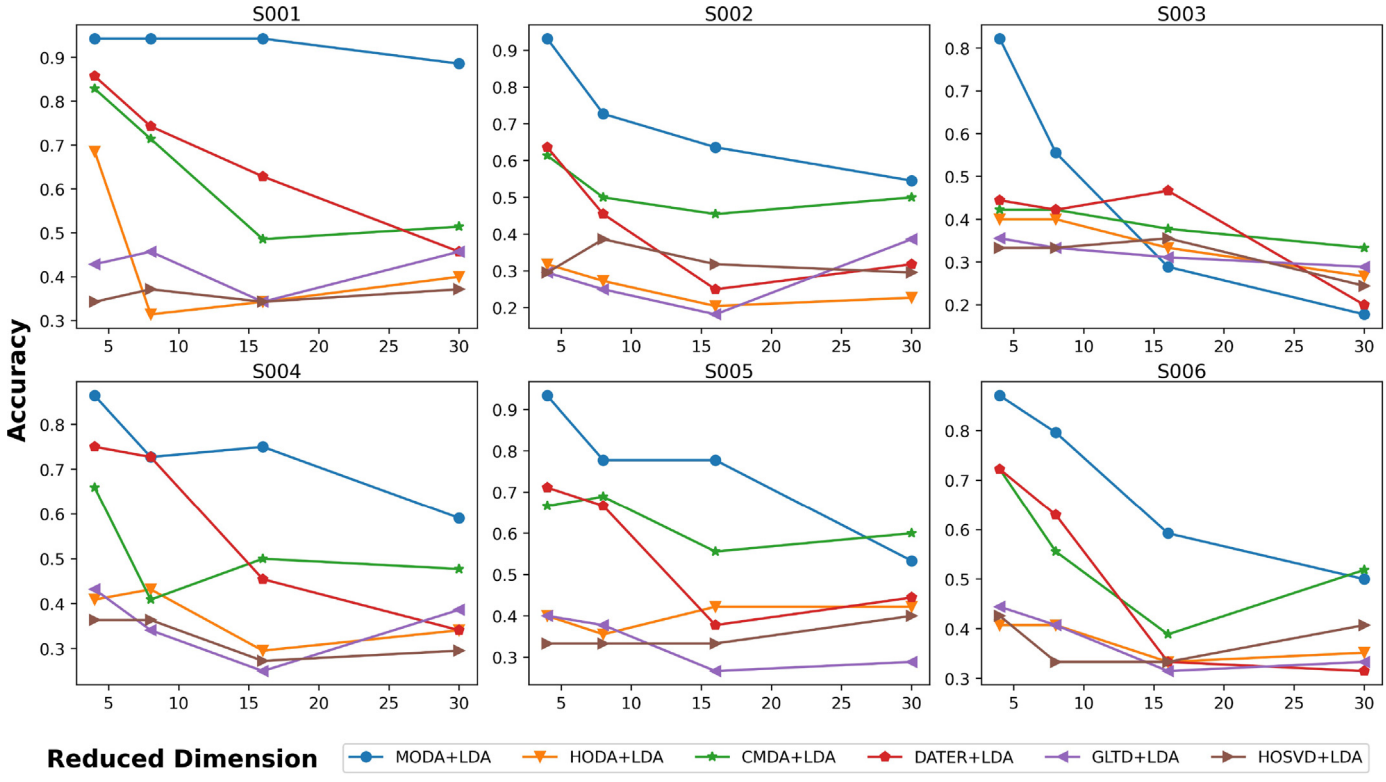

Fig. 4. The classification accuracy using LDA classifier on the EEG SSVEP dataset.

Once the Riemannian gradient in Eq. (8) and Riemannian Hessian in Eq. (10) are calculated, we can apply the Riemannian version of trust-region algorithm (implemented in the Manopt toolbox) to solve MODA, which is the recent retraction-based framework of Riemannian manifold optimization [4]. More details about this retraction-based methods can refer to Absil et al. [1], Mishra and Sepulchre [23] as the reference therein. Algorithm 1 shows the pseudo-code of MODA. The code for MODA is available at https://github.com/ncclabsustech/MODA-algorithm.

Algorithm 1 Manifold optimization based discriminant analysis (MODA).

Require: input data $X \in \mathbb{R}^{D\times N}$, sample labels $L \in \mathbb{R}^{N\times 1}$

initial matrix U, gradient norm tolerance $\varepsilon^1 = 10^{-5}$, and max iteration number maxit = 200. Let $0 < c < 1$, $\beta^1 = 0$, $\xi^0 = 0$.
for $k \leq$ maxit **do**
  Compute Hessian in the Euclidean space, Hess$f(U)[\xi]$, by Eq.(9)
  Compute the Riemannian Hessian, hess$f(U)[\xi]$, by Eq.(10)
  Compute the weighted value $\beta^k = tr\left(\eta^{kT}\eta^k\right)/tr\left(\eta^{(k-1)T}\eta^{k-1}\right)$.
  Compute a transport direction $\mathcal{T}_{U^{k-1}\to U^k(\xi^{k-1})} = \mathcal{P}_{U^k}\left(\xi^{k-1}\right)$.
  Compute the search direction $\xi^k = -\text{grad}_{\text{R}} f\left(U^k\right) + \beta^k \mathcal{T}_{U^{k-1}\to U^k(\xi^{k-1})}$.
  Compute the step size $t^k \geq 0$ using backtracking $f\left(\text{R}_{U^k}\left(t^k\xi^k\right)\right) \geq f\left(U^k\right) + ct^k tr\left(\eta^{kT}\xi^k\right)$.
  Compute the next iterate by retraction *i.e.* R_U $U^{k+1} := \text{R}_{U^k}\left(t^k\xi^k\right)$
  Terminate and output U^{k+1} if one of the stopping conditions, $\left\|\eta^{k+1}\right\|_F^2 \leq \varepsilon^1$, or iteration number $k \geq$ maxit is met.
end for
OUTPUT Discriminative Subspace U.

3.4. Calculation of reduced data

After the discriminative subspace U is learned, we can obtain the low-dimensional representation of the high-dimensional input data as follows,

$$Y = U^T X, \tag{11}$$

where Y is the low-dimensional representation of the input data X. In comparison to the original data, which can give an enhancement of separating the different classes.

4. Numerical experiments and results

In this section, we validate the effectiveness of MODA on feature extraction for classification tasks. MODA is compared with three discriminative subspace learning methods (i.e., HODA [25], DATER [37], CMDA [17], and two tensor decomposition methods (i.e., GLTD [35] and HOSVD [8]). All subsequent numerical experiments are carried out on a desktop (Intel Core i5-5200U CPU with a frequency of 2.20 GHz and a RAM of 8.00 GB). Each experiment is repeated 10 times, each time using different random sampling data.

4.1. Image classification

Our experiment involves seven benchmark image datasets, namely the COIL20 Object, ETH80 Object, MNIST Digits, USPS Digits, ORL Faces, Olivetti Faces, and CMU PIE Faces. Figure 1 shows some examples of sampling from these datasets.

The COIL20 dataset contains 1420 grayscale images of 20 objects (72 images per object). Objects in COIL20 have a variety of complex geometric and reflective properties. In our experiments,

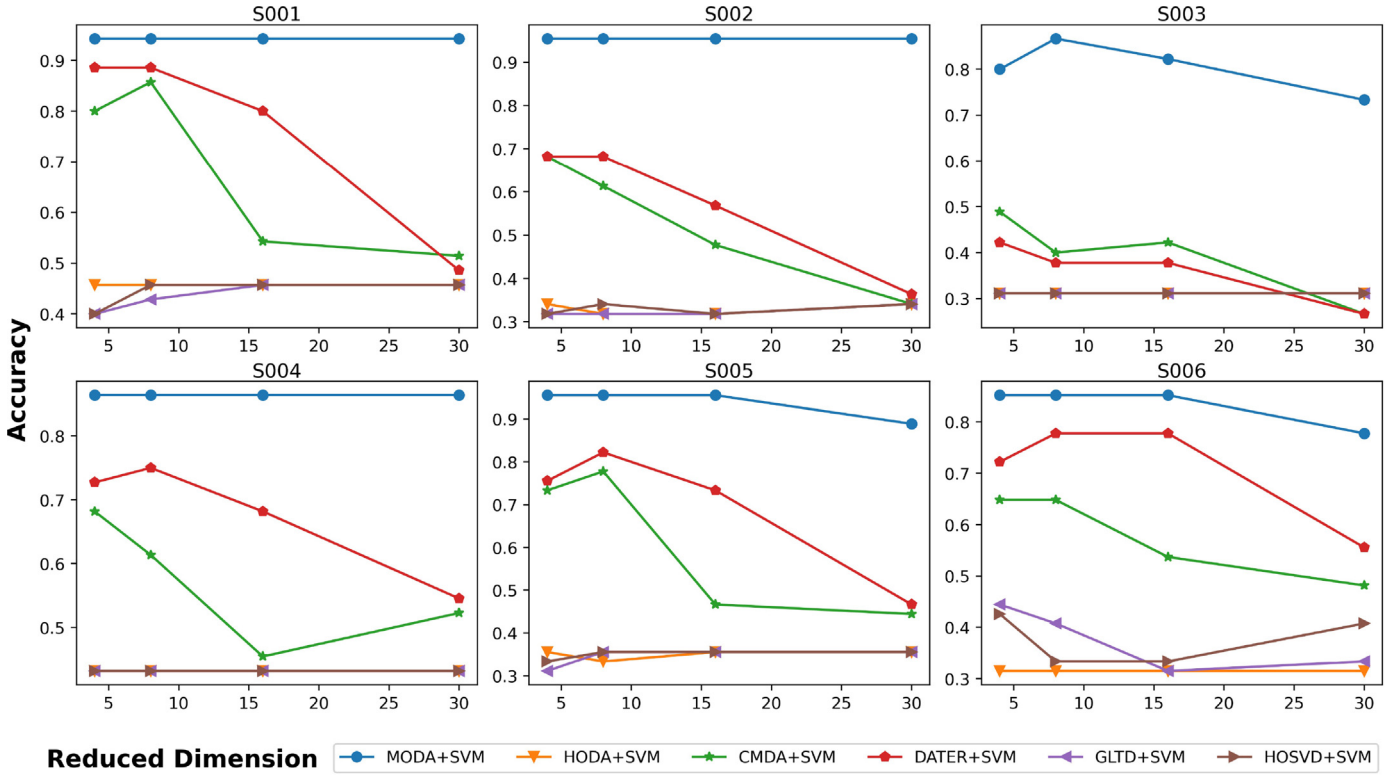

Fig. 5. The classification accuracy using SVM classifier on the EEG SSVEP dataset.

the image from COIL20 is downsampled to a size of 32×32 with 0–255 grayscale.

The ETH80 dataset is a multi-view image dataset used for object classification. It includes 8 categories: apple, car, cow, cup, dog, horse, pear, and tomato. Each category contains 10 objects, and each object has 41 images from different views, resulting in a total of 3280 images. The resolution of original images is 128×128, and we adjust the size of each image to 32×32 pixel.

Both USPS and MNIST datasets are 0–9 handwritten digits. The USPS dataset has 11,000 images, with a size of 16×16 pixels, while the MNIST dataset has 60,000 images belonging to the training set, with a size of 28×28 pixels. In our experiment, we randomly selected 2000 images (200 images per category) from the USPS dataset, and 3000 images (300 images per category) from the MNIST dataset.

The ORL dataset contains 400 images from 40 different people, each with 10 different images. These images were taken multiple times under different lighting conditions and facial expressions (eyes open/closed; with/without smile) and facial details (with/without glasses). All images were taken against a dark uniform background, with the subject in an upright frontal position (tolerance to certain lateral movements). We adjust the size of each image to 32×32 pixels.

The Olivetti dataset consists of 400 faces from 40 people (10 per person). The viewing angle of those images changes very little, but people's expressions change a lot, and occasionally they wear glasses. The size of the image is $64 \times 64 = 4096$ pixels, and the data is labeled according to the identity.

The CMU PIE dataset is a gray-scale face dataset, including 68 people, and each person has 141 face images. The images were taken under different lighting conditions. We extracted a subset of 50 individuals and the corresponding 50 facial images of each person, resulting in a total of 2500 images.

Table 2 shows the general description of seven datasets, where the attributes of each dataset are the total number of samples, dimension of the original data, reduced feature dimension, and the number of classes. Note that each sample has a real category label (e.g., object, identity, or digit). We preprocess the dataset with the following two steps: (i) randomly shuffle all the data, (ii) normalize the gray value of pixels to the unit.

Table 2
Illustrations of the datasets.

Dataset	#samples	$\text{size}_{original}$	size_{final}	#classes
ETH80	3280	32×32	8×8	8
MNIST	3000	28×28	8×8	10
USPS	2000	16×16	8×8	10
COIL20	1440	32×32	8×8	20
ORL	400	32×32	8×8	40
Olivetti	400	64×64	8×8	40
CMU PIE	2500	32×32	8×8	50

In the following numerical experiments, the image data is represented by a third-order tensor, where the first two modes are associated with the spatial information of image pixels, and the last mode represents the number of samples. It is worth noting that MODA and its implementation are very general, and there is no such restriction on the data format. In the test stage, we perform the subspace learning and reduce the dimension of the input data (from $\text{size}_{original}$ to size_{final} in Table 2). We select four supervised algorithms (e.g., MODA, HODA, CMDA, and DATER) to conduct the subspace learning, and then execute classification tasks by using the learned subspace. To fair comparison, we randomly initialize 10 times and average the results across 10 times, and we quantify the results by using LDA and SVM classifier [31].

As shown in Table 3, MODA achieves the best performance in ETH80, ORL, Olivetti and CMU PIE datasets, and CMDA achieves the best performance in COIL20 dataset. However, when the between-class matrix is singular, both CMDA and DATER fail to deal with ETH80 and MNIST datasets. In general, our proposed method

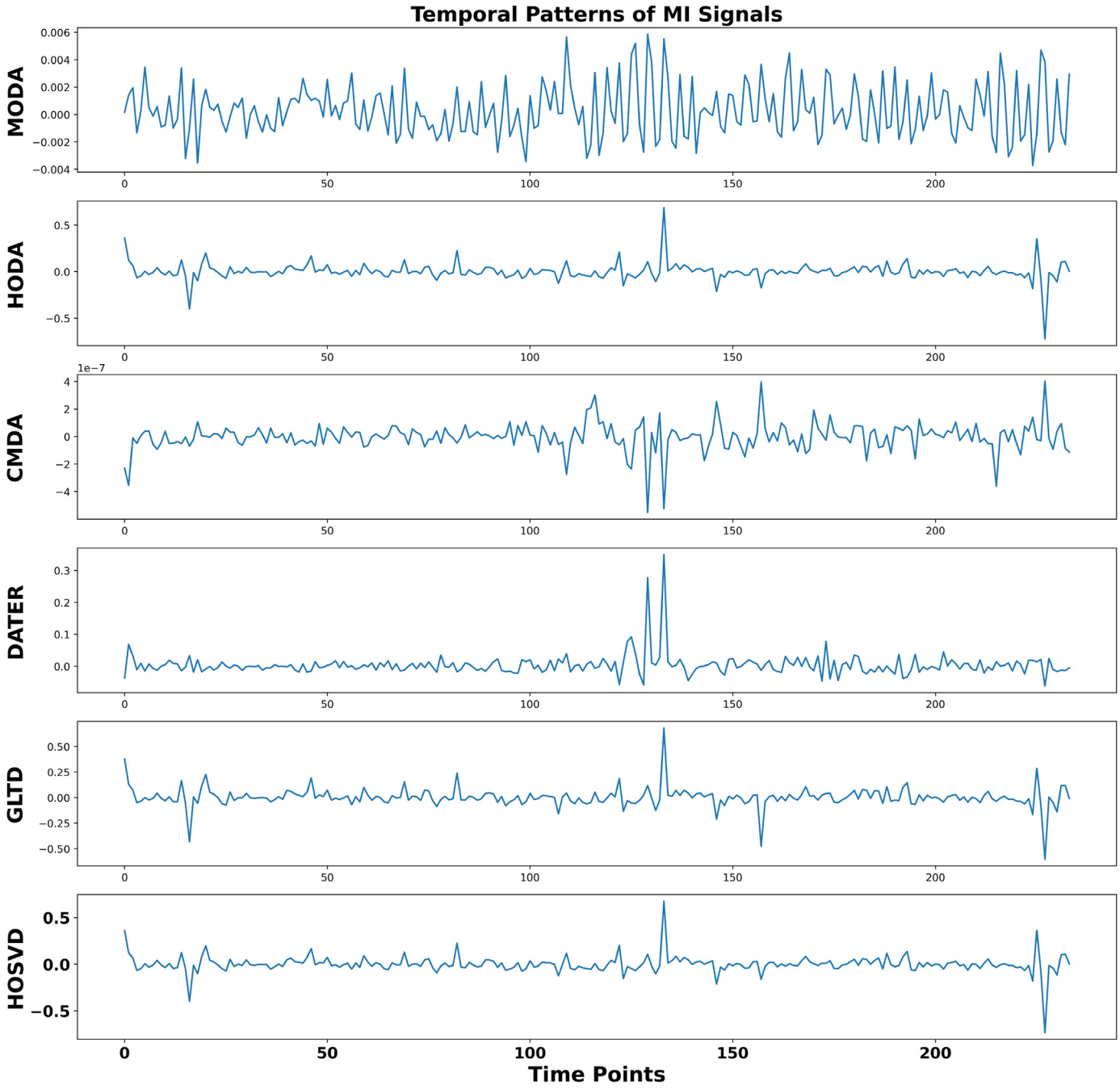

Fig. 6. Temporal patterns comparison using MODA, HODA, CMDA, DATER, GLTD and HOSVD methods on the EEG MI dataset. The sample frequency is 200 Hz.

provides better classification accuracy than Euclidean-based algorithms, demonstrating that MODA has a better capacity to extract complex features.

4.2. EEG classification

EEG classification is an important application for discriminative subspace learning. Here, we test the efficiency and accuracy of our proposed MODA on multiple EEG datasets to classify motor imagery (MI) and steady-state visual evoked potential (SSVEP) signals.

First, the Macau MI dataset is recorded by our collaborators at the University of Macau with ethical approval. It contains 128-channel EEG recordings from 6 subjects sampled at 1000 Hz. The MI task consists of 255 trials. In each trial, the subject is instructed by a visual cue to imagine one of four movements (*i.e.*, forwards, backwards, turn left and turn right). The 2-second epochs of EEG signals are extracted and then downsampled to 200 Hz, resulting in 400 time samples. The EEG data from one subject can be represented by $trial \times channel \times time$ ($255 \times 128 \times 400$). We set the embedding feature space ranging from 4×4 to 30×30 in the training, and test the model with feature space 4×4, 8×8, 16×16, and 30×30. Figures 2 and 3 show the classification accuracy by using LDA and SVM classifier, respectively.

As shown in Figs. 2 and 3 on the MI dataset, our proposed MODA algorithm has higher performance for each subject compared with the competing algorithms, and the superior of MODA is more obvious with an intensive dimensionality reduction when

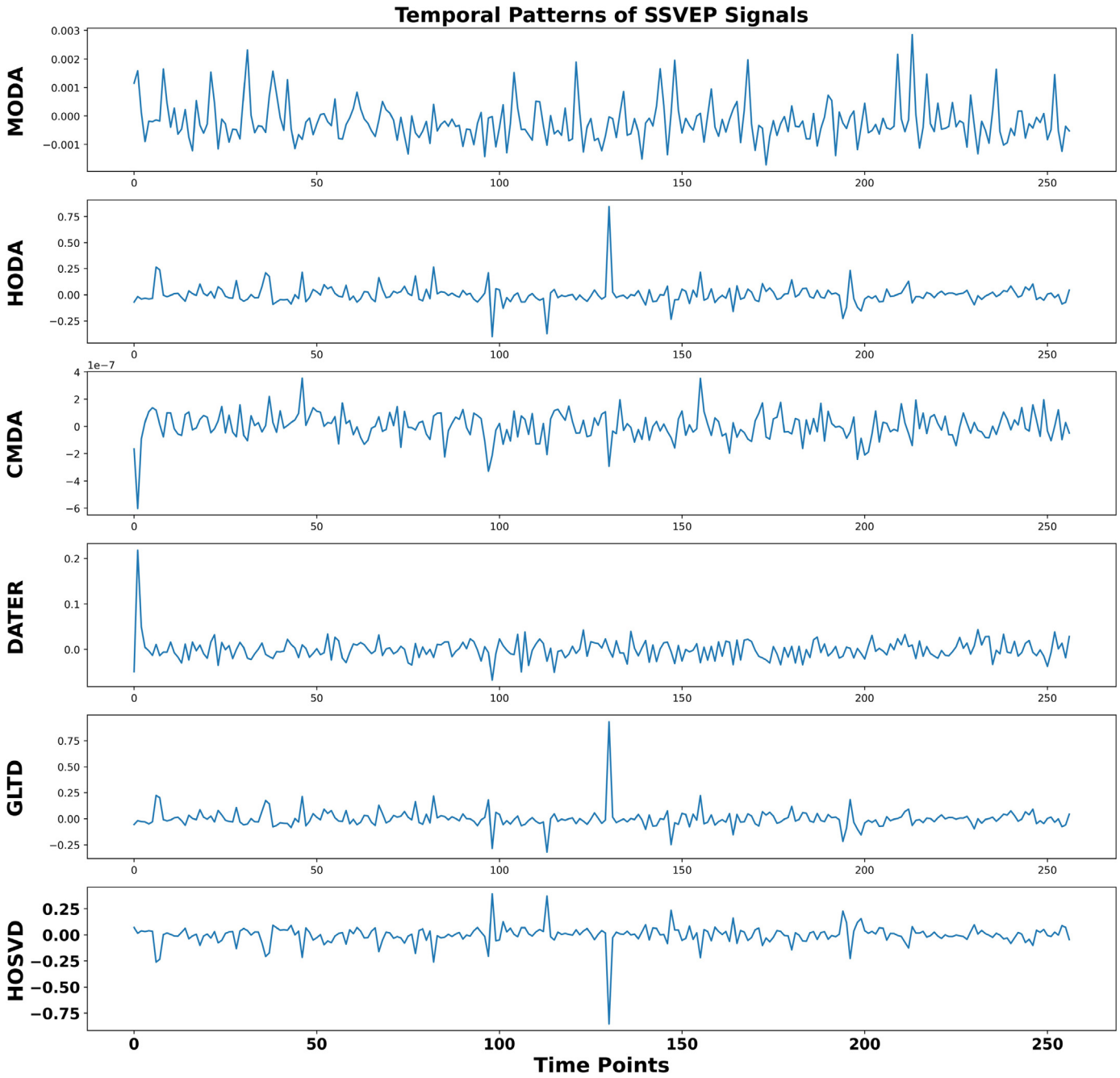

Fig. 7. Temporal patterns comparison using MODA, HODA, CMDA, DATER, GLTD and HOSVD methods on the EEG SSVEP dataset. The sample frequency is 200 Hz.

the feature dimension less than the original dimension (i.e., $d \ll D$, such as 4×4 and 8×8). Otherwise, the classification accuracy of MODA drops with an increasing feature dimension from 4×4 to 30×30.

Second, the Macau SSVEP dataset contains 128-channel EEG recordings from 6 subjects sampled at 1000 Hz, which was recorded at the University of Macau with their ethical approval. There are four types of visual stimulus with flashing frequencies at 10 Hz, 12 Hz, 15 Hz, and 20 Hz. The EEG signals are down-sampled to 200Hz and then epoched to 2-second SSVEP segments. Eventually, the SSVEP data is represented by $trial \times channel \times time$ ($255 \times 128 \times 400$). Figures 4 and 5 shows the classification accuracy by using LDA and SVM respectively, where the feature space is reduced from 4×4 to 30×30, and we select 4×4, 8×8, 16×16, and 30×30 as the test dimension.

As shown in Figs. 4 and 5 on the SSVEP dataset, when the feature dimension is reduced from 4×4 to 30×30, MODA achieves classification accuracy has a sharply down, especially when the reduced dimension is large (such as 4×4 and 8×8), our proposed MODA algorithm has a higher performance than the comparison algorithms on all of the subjects, demonstrating that the search space ($d \ll D$) benefits from finding the optimal solution. Through the EEG experimental analysis, it is shown that the performance degradation can be traced back to the selection of feature dimension used for searching for the optimal solution.

In addition, to verify that our proposed MODA can extract the significant features or patterns, we display the temporal patterns of EEG SSVEP signals in Fig. 6, and temporal patterns of EEG MI signals in Fig. 7, respectively. Figure 6 indicates that MODA can extract the effective EEG features to classify motor

Table 3
Classification performance comparisons between MODA, HODA, CMDA, DATER, GLTD and HOSVD methods on seven different datasets.

		Method					
Dataset	Classifier	MODA	HODA	CMDA	DATER	GLTD	HOSVD
ETH80	LDA	**0.8140**	0.7179	*	0.2469	0.7286	0.7439
	SVM	**0.7286**	0.8064	*	0.2454	0.8185	0.8170
MNIST	LDA	**0.8900**	0.8533	*	*	0.8750	0.8750
	SVM	0.8866	0.9433	*	*	**0.9550**	0.9533
USPS	LDA	**0.9175**	0.9125	0.9000	0.8925	0.9050	0.9000
	SVM	0.9375	0.9625	**0.9800**	0.9525	0.9775	0.9775
COIL20	LDA	0.9861	0.9583	**0.9895**	0.9548	0.9722	0.9652
	SVM	0.9548	0.9583	**0.9826**	0.9583	0.9618	0.9618
ORL	LDA	**1.0000**	0.9500	0.9750	0.9875	0.9875	0.9875
	SVM	0.0375	0.0000	**0.1000**	0.0000	0.0125	0.0125
Olivetti	LDA	**1.0000**	0.9375	1.0000	1.0000	0.9625	0.9750
	SVM	**0.0375**	0.0250	0.2125	0.6500	0.0250	0.0250
PIE	LDA	**1.0000**	0.9800	0.9880	0.9820	0.9700	0.9740
	SVM	**1.0000**	0.5680	0.9920	**1.0000**	0.5460	0.5540

* The algorithm fails, as the between-class matrix is singular.

imagery, while the comparison algorithms are difficult to obtain the interpretable features. Analogously, Fig. 7 demonstrate that MODA can extract the significant and complex features for SSVEP classification.

5. Discussion and conclusion

In this study, we proposed the manifold optimization-based discriminant analysis (MODA) for dimensionality reduction. In MODA, we construct a linear model (in Eq. (1)) by alignment of latent subspace [30] and display a nonlinear optimization strategy (the second-order geometry of Riemannian trust-region algorithm) for solving the learning objective. By a variety of numerical experiments, we verified that (Table 3 and Figs. 2–7) MODA outperforms the comparison algorithms defined in the Euclidean space, and robustly obtain the holistically optimal solution.

For the equivalence class of quotient space defined in the Grassmann manifold, the search space for the extremum value of objective function can be significantly decreased in MODA, which can shown that the performance of algorithm is significantly affected by the feature dimension of optimal solution in the search space. Since the retraction-based framework of Riemannian trust-region algorithm for solving objective function can guarantee a quadratic rate of global convergence [1,14], MODA algorithm is efficient. Due to the limited number of samples are available for training in biomedical engineering, Riemannian manifold optimization instead of deep learning methods can provide a new approach to solve the traditional Euclidean-based algorithms, which can effectively resolve the small sample datasets learning.

Although we have shown that MODA benefits from Riemannian manifold optimization for discriminative subspace learning, there are also many other aspects that need to be further investigated in our future work. For example, how to construct the Riemannian preconditioning properly, and what is the best Riemannian metric for algorithm. Previous studies have been shown that employing the underlying structure of constraint conditions in objective function can effectively speed up the convergence speed of algorithms [14]. In addition, manifold optimization-based discriminative subspace learning has its own limitations as well, such as suffering from an expensive optimization process to find an optimal subspace and involving a lot of concepts to learn.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank Dr. Haiyan Wu and Mr. Ruien Wang for sharing the EEG data, and Mr. Youzhi Qu for editing the draft. This work was funded in part by the National Natural Science Foundation of China (62001205), National Key R&D Program of China (2021YFF1200804), Shenzhen Science and Technology Innovation Committee (20200925155957004, KCXFZ2020122117340001, JCYJ20220818100213029), Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (SGDX2020110309280100), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003).

References

[1] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Optimization Algorithms on Matrix Manifolds, Princeton University Press, 2009.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[3] F. Bouchard, J. Malick, M. Congedo, Riemannian optimization and approximate joint diagonalization for blind source separation, IEEE Trans. Signal Process. 66 (8) (2018) 2041–2054.
[4] N. Boumal, B. Mishra, P.-A. Absil, R. Sepulchre, Manopt, a matlab toolbox for optimization on manifolds, J. Mach. Learn. Res. 15 (1) (2014) 1455–1459.
[5] H. Chen, Y. Sun, J. Gao, Y. Hu, B. Yin, Solving partial least squares regression via manifold optimization approaches, IEEE Trans. Neural Netw. Learn. Syst. 30 (2) (2018) 588–600.
[6] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (3) (1994) 287–314.
[7] C. Cruceru, G. Bécigneul, O.-E. Ganea, Computationally tractable Riemannian manifolds for graph embeddings, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 7133–7141.
[8] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.
[9] A. Douik, B. Hassibi, Manifold optimization over the set of doubly stochastic matrices: a second-order geometry, IEEE Trans. Signal Process. 67 (22) (2019) 5761–5774.
[10] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[11] T. Hazan, S. Polak, A. Shashua, Sparse image coding using a 3D non-negative tensor factorization, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 1, IEEE, 2005, pp. 50–57.
[12] R. Henriques, S.C. Madeira, Flebic: learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns, Pattern Recognit. 115 (2021) 107900.
[13] X. Jiang, Linear subspace learning-based dimensionality reduction, IEEE Signal Process. Mag. 28 (2) (2011) 16–26.
[14] H. Kasai, B. Mishra, Low-rank tensor completion: a Riemannian manifold preconditioning approach, in: International Conference on Machine Learning, 2016, pp. 1012–1021.
[15] Y.-D. Kim, S. Choi, Nonnegative tucker decomposition, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[16] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2001, pp. 556–562.
[17] Q. Li, D. Schonfeld, Multilinear discriminant analysis for higher-order tensor data classification, IEEE Trans Pattern Anal Mach Intell 36 (12) (2014) 2524–2537.
[18] M. Liao, X. Gu, Face recognition approach by subspace extended sparse representation and discriminative feature learning, Neurocomputing 373 (2020) 35–49.
[19] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, MPCA: multilinear principal component analysis of tensor objects, IEEE Trans. Neural Networks 19 (1) (2008) 18–39.
[20] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition, IEEE Trans. Neural Networks 20 (1) (2008) 103–123.

[21] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Regularized common spatial patterns with generic learning for eeg signal classification, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2009, pp. 6599–6602.
[22] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, A survey of multilinear subspace learning for tensor data, Pattern Recognit. 44 (7) (2011) 1540–1551.
[23] B. Mishra, R. Sepulchre, Riemannian preconditioning, SIAM J. Optim. 26 (1) (2016) 635–660.
[24] U. Mor, H. Avron, Solving trust region subproblems using Riemannianoptimization, arXiv preprint arXiv:2010.07547(2020).
[25] A.H. Phan, A. Cichocki, Tensor decompositions for feature extraction and classification of high dimensional datasets, Nonlinear Theory Appl., IEICE 1 (1) (2010) 37–68.
[26] Y. Qiu, G. Zhou, Y. Wang, Y. Zhang, S. Xie, A generalized graph regularized non-negative tucker decomposition framework for tensor data representation, IEEE Trans. Cybern. 52 (1) (2020) 594–607.
[27] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[28] X.-J. Shen, S.-X. Liu, B.-K. Bao, C.-H. Pan, Z.-J. Zha, J. Fan, A generalized least-squares approach regularized with graph embedding for dimensionality reduction, Pattern Recognit. 98 (2020) 107023.
[29] H. Sifaou, A. Kammoun, M.-S. Alouini, High-dimensional linear discriminant analysis classifier for spiked covariance model, J. Mach. Learn. Res. 21 (1) (2020) 4508–4531.
[30] B. Su, Y. Wu, Learning low-dimensional temporal representations with latent alignments, IEEE Trans. Pattern Anal. Mach. Intell. 42 (11) (2019) 2842–2857.
[31] Y. Sun, J. Gao, X. Hong, B. Mishra, B. Yin, Heterogeneous tensor decomposition for clustering via manifold optimization, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2015) 476–489.
[32] D. Tao, X. Li, X. Wu, S. Maybank, Tensor rank one discriminant analysis—A convergent method for discriminative multilinear subspace selection, Neurocomputing 71 (10–12) (2008) 1866–1882.
[33] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and Gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700–1715.
[34] M.A.O. Vasilescu, D. Terzopoulos, Multilinear independent components analysis, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 547–553.
[35] B. Jiang, C. Ding, J. Tang, B. Luo, Image Representation and Learning With Graph-Laplacian Tucker Tensor Decomposition, IEEE Trans Cybern. 49 (4) (2019) 1417–1426.
[36] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[37] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, Discriminant analysis with tensor representation, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 526–532.
[38] W. Yin, Z. Liang, J. Zhang, Q. Liu, Partial least square regression via three-factor SVD-type manifold optimization for eeg decoding, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2022, pp. 778–787.
[39] W. Yin, Z. Ma, High order discriminant analysis based on Riemannian optimization, Knowl Based Syst 195 (2020) 105630.
[40] W. Yin, Y. Qu, Z. Ma, Q. Liu, HyperNTF: A hypergraph regularized nonnegative tensor factorization for dimensionality reduction, Neurocomputing 512 (2022) 190–202.

Wanguang Yin received the B.Sc. degree from Guizhou University, Guiyang, China, in 2012, and the M.Sc. degree in GuiZhou University of Finance and Economics, Guiyang, China, in 2015. He obtained his Ph.D. degree with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China in 2020. Currently, he is a postdoc in Southern University of Science and Technology, his research interest is tensor decomposition, Riemannian manifold optimization and its applications in EEG and fMRI analysis.

Zhengming Ma received the B.Sc. and M.Sc. degrees from the South China University of Technology, Guangzhou, China, in 1982 and 1985, respectively, and the Ph.D. degree in pattern recognition and intelligence control from Tsinghua University, Beijing, China, in 1989. He is currently a Professor with the Nanfang College, Sun Yat-sen University, Guangzhou. His current research interests include machine learning and signal processing.

Quanying Liu is an assistant professor at Department of Biomedical Engineering (BME), Southern University of Science and Technology (SUSTech), China. She achieved the master degree in Computer Science at Lanzhou University in 2013, and the Ph.D. degree in BME at ETH Zurich in 2017. She had a postdoctoral training at CalTech before joining SUSTech. Her main research interests include machine learning, control theory and signal processing for neural data.