Cross-session Emotion Recognition by Joint Label-common and Label-specific EEG Features Exploration

Yong Peng, *Member, IEEE,* Honggang Liu, *Student Member, IEEE,* Junhua Li, *Senior Member, IEEE,* Jun Huang, *Member, IEEE,* Bao-Liang Lu, *Fellow, IEEE,* and Wanzeng Kong*, *Member, IEEE*

***Abstract*—Since Electroencephalogram (EEG) is resistant to camouflage, it has been a reliable data source for objective emotion recognition. EEG is naturally multi-rhythm and multi-channel, based on which we can extract multiple features for further processing. In EEG-based emotion recognition, it is important to investigate whether there exist some common features shared by different emotional states, and the specific features associated with each emotional state. However, such fundamental problem is ignored by most of the existing studies. To this end, we propose a Joint label-Common and label-Specific Features Exploration (JCSFE) model for semi-supervised cross-session EEG emotion recognition in this paper. To be specific, JCSFE imposes the $\ell_{2,1}$-norm on the projection matrix to explore the label-common EEG features and simultaneously the ℓ_1-norm is used to explore the label-specific EEG features. Besides, a graph regularization term is introduced to enforce the data local invariance property, *i.e.*, similar EEG samples are encouraged to have the same emotional state. Results obtained from the SEED-IV and SEED-V emotional data sets experimentally demonstrate that JCSFE not only achieves superior emotion recognition performance in comparison with the state-of-the-art models but also provides us with a quantitative method to identify the label-common and label-specific EEG features in emotion recognition.**

***Index Terms*—EEG emotion recognition, graph regularization, label-common features, label-specific features, semi-supervised regression.**

I. Introduction

EMOTION refer to people's psychological reactions to external stimuli or their own stimuli accompanied by physiological reactions [1]. Emotions have an important impact on the establishment and maintenance of interpersonal relationships [2], cognition [3], decision-making [4], and other interactive activities. Many mental disorders are closely related to emotions [5]; therefore, identifying the emotional state of people with emotional expression disorders is helpful to their treatment and healthcare. In past decades, emotion recognition has been attracting increasing attention from both academia and industry [6]. Compared with the traditional data modalities such as facial expressions, text, and speech, EEG can offer us more reliable emotion recognition results because it is originated from the neural activities of our central nervous system and is not easily camouflaged [7]. With the development of weak signal acquisition equipments and processing techniques, EEG has been widely used in multiple scenarios such as drowsiness estimation [8], rehabilitation engineering [9], and disease diagnosis [10]. In the present work, we put the emphasis on EEG emotion recognition [11].

Current studies in EEG emotion recognition mainly focused on two aspects. One is the feature extraction methods to characterize the statistics, frequency, and nonlinear characteristics of EEG data [12], [13]. Generally, the popular EEG features for emotion recognition are extracted from the time-, frequency-, time-frequency and spatial domains. The other focus is the feature transformation and recognition models [14]. Roughly, we can categorize these existing models into linear, kernel-based and neural networks-based nonlinear ones. They improved emotion recognition performance by diverse motivations such as enhancing the model robustness [15], distinguishing the different discriminative abilities of features [16], and minimizing the inter-subject variabilities [17]. Instead of using the handcrafted EEG features, sometimes raw EEG data is fed into deep learning models to simultaneously obtain the data representations and emotion recognition results. That is, the feature learning and classification are unified together to achieve the end-to-end EEG decoding [18].

The multi-channel and multi-rhythm properties of EEG provide us with abundant spatial and frequency information, based on which the extracted features are used for emotional state estimation. Based on the consensus that different EEG frequency bands and channels correlate differently to mental states [8], [19], different dimensions of a certain feature type (*e.g.*, power spectra density or differential entropy) should also correlate differently to different types of emotional states. In pattern recognition, different features have different discriminative abilities in classifying the emotional states. Then, a fundamental problem in EEG emotion recognition is whether

Manuscript received September 20, 2022; revised December 18, 2022 and December 23, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61971173 and Grant U20B2074, and in part by the Natural Science Foundation of Zhejiang Province under Grant LY21F030005. *(Corresponding author: Wanzeng Kong).*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Shanghai Jiao Tong University under Protocol No. 2017060.

Yong Peng and Wanzeng Kong are with the School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence, Hangzhou 310018, China (email: kongwanzeng@hdu.edu.cn).

Honggang Liu is with the HDU-IMTO Joint Institute, Hangzhou Dianzi University, Hangzhou 310018, China.

Junhua Li is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK.

Jun Huang is with the School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243002, China.

Bao-Liang Lu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

there exist some common features that are discriminative for all the involved emotional states. Accordingly, we also want to investigate whether there exist label-specific features that are discriminative only to a specific emotional state. However, this problem has not been fully studied yet within the community of EEG emotion recognition.

In this paper, we propose a new Joint label-Common and label-Specific Features Exploration (JCSFE) model for cross-session EEG emotion recognition, which is implemented based on the semi-supervised regression. Specifically, we impose the $\ell_{2,1}$-norm on the regression projection matrix to explore the label-common features by achieving row-sparsity; simultaneously, the ℓ_1-norm is used to explore the label-specific features due to its isotropic property. Moreover, a graph regularizer is incorporated into JCSFE by enforcing the local invariance property of data. As a summary, the present work consists of the following contributions.

- We propose a new emotion recognition model by joint label-common and label-specific EEG features exploration, which are respectively achieved by imposing the $\ell_{2,1}$-norm and ℓ_1-norm on the projection matrix in the semi-supervised regression.
- As the secondary contribution, JCSFE incorporates the graph regularizer to enforce the local invariance property of data. Besides, an efficient optimization algorithm is proposed to optimize the JCSFE model objective whose convergence and complexity are analyzed.
- On the emotion recognition performance, JCSFE not only obtains improved accuracy but also provides us with quantitative measurement of the EEG spatial-frequency activation patterns for emotion recognition from two perspectives, *i.e.*, each feature in terms of all emotional states and each emotional state in terms of all features.

We organize the rest of this paper as follows. Section II introduces the JCSFE model formulation and optimization. Comparative studies are conducted and the results are analyzed in section III. Discussions to clarify the connections as well as differences between JCSFE and some related models are provided in section IV. Section V concludes this paper and describes the potential future work.

II. Method

A. Problem Definition

In this paper, matrices are denoted by blodface uppercase letters and vectors are written as bolaface lowercase letters. For matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its (i,j)-th element is m_{ij}. Its i-th row, j-th column are respectively denoted as $\mathbf{m}^i$, $\mathbf{m}_j$. The boldface $\mathbf{1}_m$ represents an all-one column vector whose length is m. The ℓ_1-norm of vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$. The $\ell_{2,1}$-norm of $\mathbf{M}$ is $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^n \|\mathbf{m}^i\|_2 = \sum_{i=1}^n \sqrt{\sum_{j=1}^m m_{ij}^2}$.

Generally, in semi-supervised EEG emotion recognition, we are often given an EEG data set $\mathbf{X} = [\mathbf{X}_l, \mathbf{X}_u] \in \mathbb{R}^{d\times n}$, where $\mathbf{X}_l \in \mathbb{R}^{d\times l}$ is the labeled subset and $\mathbf{X}_u \in \mathbb{R}^{d\times u}$ is the unlabeled subset. Accordingly, $\mathbf{Y}_l \in \mathbb{R}^{l\times c}$ is the emotional state indicator matrix of these labeled samples. Here, d is the sample dimensionality, c is the number of emotional states, l and u are respectively the numbers of labeled and unlabeled samples (*i.e.*, $n = l + u$). The $i|_{i=1}^l$-th row of $\mathbf{Y}_l$, $\mathbf{y}^i \in \mathbb{R}^{1\times c}$, encodes the label information of sample $\mathbf{x}_i \in \mathbb{R}^d$ as

$$y_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ corresponds to the } j\text{-th state;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

By defining $\mathbf{Y}_u \in \mathbb{R}^{u\times c}$ as the indicator matrix of the unlabeled EEG data and $\mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u] \in \mathbb{R}^{n\times c}$, our aim is to estimate $\mathbf{Y}_u$ as accurately as possible given $\mathbf{X}$ and $\mathbf{Y}_l$.

Below we use an example to illustrate the label-common and label-specific features in pattern classification. Suppose that we have a data matrix which contains two instances $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$ with five dimensional feature space $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_5\}$. The corresponding label vectors are $\mathbf{Y} = [\mathbf{y}^1; \mathbf{y}^2]$. The two elements in each label vector represents the probability of the corresponding instance belonging to the two classes, respectively. By fitting $(\mathbf{X}, \mathbf{Y})$ by a projection matrix $\mathbf{W}$, we obtain one possible solution of $\mathbf{W}$ shown in Fig. 1. Through the non-zero values of the two columns of $\mathbf{W}$, *i.e.*, $\mathbf{w}_1$ and $\mathbf{w}_2$, we know the specific features of each class. Specifically, $\mathbf{w}_1 = [1,1,1,0,0]^T$ means that features $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ determines the first class, while $\mathbf{w}_2 = [0,0,1,1,1]^T$ indicates that features $\mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_5$ determines the second class. $\mathbf{f}_3$ is the common feature for both classes.

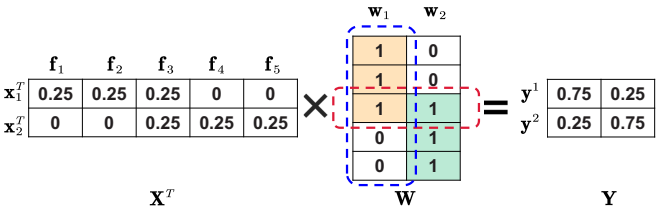


Fig. 1. An example to illustrate the two different types of features.

B. JCSFE Model Formulation

In Fig. 2, we show the the overall framework of applying JCSFE into semi-supervised EEG emotion recognition task. The second stage is the JCSFE-based model learning, which is implemented under a semi-supervised regression framework due to its simplicity and effectiveness. The three components in JCSFE consist of the label-common and label-specific features mining, the consideration of data local invariance property by graph regularizer.

Given a centered data matrix $\mathbf{X}$ and the label indicator matrix $\mathbf{Y}_l$, semi-supervised regression can be expressed by

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{Y}_u} \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_2^2 + \mathcal{C}(\mathbf{W}), \\ &s.t.\ \mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u], \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u\mathbf{1}_c = \mathbf{1}_u, \end{aligned} \tag{2}$$

where $\mathcal{C}(\mathbf{W})$ defines some constraints on $\mathbf{W}$ to be described. The non-negative and row-normalization constraints make $\mathbf{Y}_u$ essentially define the probabilities of a certain EEG sample to different emotional states, based on which we can directly determine the emotional state of each unlabeled EEG sample. For example, if the $j|_{j=1}^u$-th row of the learned $\mathbf{Y}_u$ is [0.12, 0.78, 0.04, 0.06], we accordingly annotate the emotional state of this sample as the second one.

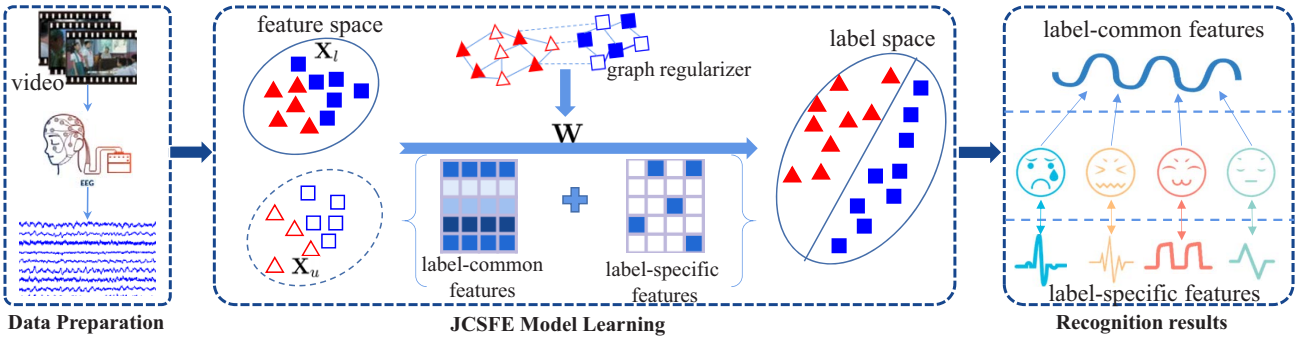


Fig. 2. The overall framework of semi-supervised EEG emotion recognition by JCSFE.

On the label-common EEG features, they have common discriminative ability for all emotional states. Based on the feature selection (ranking) theory [20], [21], for the $i|_{i=1}^{d}$-th feature, we can use the normalized ℓ_2-norm of the i-th row of $\mathbf{W}$ (*i.e.*, θ_i) to measure the extent of a feature to be a label-common one. Mathematically, a larger value of θ_i represents that the i-th feature is more discriminative in classifying the emotional states. To this end, we impose the $\ell_{2,1}$-norm on $\mathbf{W}$ to achieve the label-common features exploration, which essentially enforces it to be row-sparse. Besides the label-common features, we consider that each emotional state might be additionally determined by several specific features of its own. Therefore, we use the ℓ_1-norm regularization to select label-specific features, which enforces the projection matrix $\mathbf{W}$ to be element-wisely sparse. Currently, we achieve the following objective function

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{Y}_u} \|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_2^2+\alpha\|\mathbf{W}\|_1+\beta\|\mathbf{W}\|_{2,1}, \\ &s.t.\ \mathbf{Y}=[\mathbf{Y}_l;\mathbf{Y}_u], \mathbf{Y}_u\geq\mathbf{0}, \mathbf{Y}_u\mathbf{1}_c=\mathbf{1}_u. \end{aligned} \tag{3}$$

Denote $\mathbf{W}\triangleq[\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_c]\in\mathbb{R}^{d\times c}$ in which $\mathbf{w}_j$ is its j-th column. The coefficient vector $\mathbf{w}_j$ is expressed as $\mathbf{w}_j=[w_{1j},w_{2j},...,w_{dj}]^T$, where w_{ij} expresses the discrimination of the i-th feature in terms of the j-th emotional state. That is, $w_{ij}\neq 0$ means that the i-th feature is discriminative for recognizing the j-th emotional state. Then it is considered as a label-specific feature of the j-th emotional state. On the contrary, $w_{ij}=0$ means that it is useless for recognizing the j-th emotional state. In objective function (3), non-negative regularization parameters α and β are used to balance these impacts of the three terms, which respectively control the element-sparsity and row-sparsity of the projection matrix $\mathbf{W}$ in exploring the label-common and label-specific features.

Besides the label-common and label-specific features exploration, we additionally take the data connections into consideration which is inspired by the consensus that learning performance can be greatly improved if the data manifold is explored and utilized. Specifically, a k-nearest neighbor (KNN) graph is adopted to measure the pairwise correlations between EEG samples. Correspondingly, a similarity matrix $\mathbf{S}\in\mathbb{R}^{n\times n}$ is built in which s_{ij} characterizes the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. For simplicity, the '0-1' weighting scheme is used in this paper, based on which we define

$$s_{ij}=\begin{cases}1, & \mathbf{x}_i\in\mathcal{N}(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in\mathcal{N}(\mathbf{x}_i);\\ 0, & \text{otherwise;}\end{cases} \tag{4}$$

where $\mathcal{N}(\mathbf{x}_i)$ contains the k-nearest neighbors of sample $\mathbf{x}_i$ based on the Euclidean distance metric. The data local invariance property asks that if two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar in original data space, their representations in projected space should be also similar. This can be achieved by

$$\min_{\mathbf{W}}\sum_{i,j=1}^{n}s_{ij}\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}\mathbf{x}_j\right\|_2^2=\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}), \tag{5}$$

where the Laplacian matrix $\mathbf{L}$ can be calculated by $\mathbf{D}-\mathbf{S}$. $\mathbf{D}$ is a diagonal matrix, whose i-th diagonal element d_{ii} is $\sum_{j=1}^{n}s_{ij}$. By incorporating (5) into (3) as a regularizer, we finally achieve the JCSFE objective function as

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{Y}_u} \|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_2^2+\alpha\|\mathbf{W}\|_1+\beta\|\mathbf{W}\|_{2,1}+ \\ &\gamma\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), s.t.\ \mathbf{Y}=[\mathbf{Y}_l;\mathbf{Y}_u], \mathbf{Y}_u\geq\mathbf{0}, \mathbf{Y}_u\mathbf{1}_c=\mathbf{1}_u, \end{aligned} \tag{6}$$

where γ is a newly introduced regularization parameter. $\mathbf{F}\triangleq\mathbf{X}^T\mathbf{W}$ is an intermediate variable to simplify the notations. Once the variables in objective function (6) are fitted by given EEG data, we can directly obtain the emotional state information of unlabeled samples by $\mathbf{Y}_u$. Moreover, based on the learned $\mathbf{W}$, we can explore the label-common and label-specific features by the form of analyzing the respective EEG spatial-frequency patterns in emotion recognition.

C. JCSFE Model Optimization

On the two variables $\mathbf{W}$ and $\mathbf{Y}_u$ in the JCSFE objective function, we propose to optimize them in alternating manner. That is, we update one variable by fixing the other.

■ **$\mathbf{Y}_u$-step**. When $\mathbf{W}$ is fixed, objective function (6) degenerates to

$$\min_{\mathbf{Y}_u}\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_2^2,\ s.t.\ \mathbf{Y}_u\geq\mathbf{0}, \mathbf{Y}_u\mathbf{1}_c=\mathbf{1}_u, \tag{7}$$

where $\mathbf{Y}=[\mathbf{Y}_l;\mathbf{Y}_u]$. The above problem can be decoupled for each $i\in\{l+1,l+2,\cdots,l+u\}$; therefore, we can optimize $\mathbf{Y}_u$ in the row-wise manner. That is, for the i-th subproblem, we need to solve

$$\min_{\mathbf{y}^i}\|\mathbf{x}_i^T\mathbf{W}-\mathbf{y}^i\|_2^2,\ s.t.\ \mathbf{y}^i\geq\mathbf{0}, \mathbf{y}^i\mathbf{1}_c=1, \tag{8}$$

which is an Euclidean projection with a simplex constraint [22]. It can be optimized by the Lagrange multiplier method together with the Karush-Kuhn-Tucker (KKT) condition. Detailed derivations can be found in the supplementary material.

■ **W-step**. Though objective function (6) is convex, it is not smooth due to the existence of the $\ell_{2,1}$-norm and the ℓ_1-norm regularization terms. Therefore, we first relax $\|\mathbf{W}\|_{2,1}$ as $\mathrm{Tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})$ to simplify the derivations [20], where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a diagonal matrix. The i-th diagonal value of $\mathbf{A}$ is $a_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}$, where $\mathbf{w}^i$ is the i-th row vector of $\mathbf{W}$. Then, we employ the accelerated proximal gradient (APG) method to deal with the ℓ_1-norm regularizer. The derivation to the updating rule of $\mathbf{W}$ is provided in the supplementary material.

The pseudo-code of the optimization procedure to JCSFE objective function is provided in Algorithm 1. Notation $\mathcal{S}$ is a soft-shrinkage operator to solve the ℓ_1-norm regularized problem which is defined as

$$\mathcal{S}_\varepsilon[x] = \begin{cases} x-\varepsilon, & \text{if } x > \varepsilon, \\ x+\varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where ε usually has a small positive value.

Algorithm 1 The optimization of JCSFE objective function

Input: Labeled EEG samples $\mathbf{X}_l \in \mathbb{R}^{d\times l}$ and the corresponding label indicator matrix $\mathbf{Y}_l \in \mathbb{R}^{l\times c}$, unlabeled EEG samples $\mathbf{X}_u \in \mathbb{R}^{d\times u}$, model parameters α, β and γ;

Output: The estimated label indicator matrix $\mathbf{Y}_u \in \mathbb{R}^{u\times c}$.

Initialize $t = 1$, $\mathbf{Y}_u = \frac{1}{c}\mathbf{1}_c\mathbf{1}_c^T$, $b^{(0)} = b^{(1)} = 1$, and $\mathbf{W}^{(0)} = \mathbf{W}^{(1)} = (\mathbf{X}\mathbf{X}^T + 0.1 * \mathbf{I})^{-1}\mathbf{X}\mathbf{Y}$;
Calculate the diagonal matrix $\mathbf{A}$;
Calculate the similarity matrix $\mathbf{S}$ via (4);
Calculate $L_f = \sqrt{3(\|\mathbf{X}\mathbf{X}^T\|_2^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|_2^2 + \|\beta\mathbf{A}\|_2^2)}$;
while not converged **do**
  $\mathbf{W}^{(t)} = \mathbf{W}^{(t)} + \frac{b^{(t-1)}-1}{b^{(t)}}(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)})$;
  $\mathbf{G}^{(t)} = \mathbf{W}^{(t)} - \frac{1}{L_f}\nabla f(\mathbf{W}^{(t)})$, where $f(\mathbf{W}) = \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_2^2 + \gamma\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \beta\mathrm{Tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})$;
  $\mathbf{W}^{(t+1)} = \mathcal{S}_\varepsilon[\mathbf{G}^{(t)}]$, where $\varepsilon = \frac{\alpha}{L_f}$;
  $b^{(t+1)} = \frac{1+\sqrt{4(b^t)^2+1}}{2}$;
  Update the diagonal matrix $\mathbf{A}^{(t+1)}$ by $a_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}$;
  Update the Lipschitz constant L_f;
  Update $\mathbf{Y}_u$ by solving (8) for each $i|_{l+1}^{l+u}$;
end while

D. Complexity and Convergence Analysis

We analyze the time complexity of Algorithm 1 below. In the initialization step, the complexity of initializing $\mathbf{W}$ is $\mathcal{O}(nd^2 + d^3 + ndc + d^2c)$. The complexity of calculating the sample similarity matrix by k-nearest neighbor is $\mathcal{O}(n^2d)$. Furthermore, the complexity of initializing L_f is $\mathcal{O}(d^3)$. In the main loop, the time cost is primarily dominated by calculating the gradient of $f(\mathbf{W})$, which can be measured by $\mathcal{O}(nd^2 + d^2c + ndc + n^2d)$. When updating $\mathbf{Y}_u$, it occupies the complexity of $\mathcal{O}(uc)$. Considering the usual case of semi-supervised EEG emotion recognition is $n \approx u > d >> c$, we conclude that the overall complexity of optimizing JCSFE model objective function by Algorithm 1 is $\mathcal{O}(tn^2d)$, where t is the number of iterations.

On the convergence property of JCSFE, we provide the analysis below. When row-wisely updating the label indicator matrix $\mathbf{Y}_u$ by the Lagrange multiplier method, the involved multipliers are analytically determined, leading to its analytical solution. When updating the projection matrix $\mathbf{W}$, the APG method is used whose convergence property has been extensively studied [23]. Therefore, we declare that the convergence of Algorithm 1 can be guaranteed.

E. Label-common and Label-specific Features Exploration

This section illustrates how to quantitatively measure a certain feature to be a common one in terms of all the emotional states, and a specific one to each of the emotional states, by the learned JCSFE model.

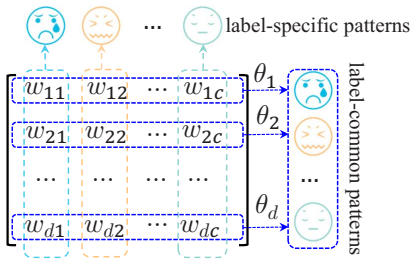


Fig. 3. Illustration of label-common and label-specific patterns.

As shown in Fig. 3, each row of the projection matrix characterizes the discriminative ability of the corresponding feature in classifying all the involved emotional states. We use θ_i as the quantitative importance measure of the i-th feature to be a label-common one. However, $\theta_i|_{i=1}^d$ is not explicitly learned by JCSFE, and only the $\ell_{2,1}$-norm was used to enforce the row-sparsity of the projection matrix. Inspired by the underlying rationality of the $\ell_{2,1}$-norm based feature auto-weighting [24], for each feature dimension, we propose to use the normalized ℓ_2-norm of the corresponding row of the projection matrix to serve as its quantitative importance. Specifically, the importance of the i-th EEG feature (*i.e.*, $\theta_i|_{i=1}^d$) can be calculated by

$$\theta_i = \frac{\|\mathbf{w}^i\|_2}{\sum_{j=1}^d \|\mathbf{w}^j\|_2}, \tag{10}$$

where $\mathbf{w}^i$ is the i-th row of $\mathbf{W}$ and $\|\mathbf{w}^i\|_2$ is the ℓ_2-norm of $\mathbf{w}^i$. Obviously, θ_is satisfy the non-negative and normalization constraints, *i.e.*, $\theta_i|_{i=1}^d \geq 0$ and $\sum_{i=1}^d \theta_i = 1$. The larger value of θ_i, the i-th EEG feature is considered to be more discriminative in distinguishing the emotional states. In other words, it should be regarded more as a label-common feature.

Intuitively, the $\ell_{2,1}$-norm based label-common feature exploration process is completed by investigating the elements of the projection matrix along the horizontal direction. Somewhat differently, the ℓ_1-norm pursues the isotropic sparsity of the projection matrix by shrinking its elements in order to identify features which might be specific to a certain emotional state.

Since the emotional label indicator matrix is arranged by one-hot encoding, the quantitative importance measure of a feature to be a label-specific one can be obtained by investigating the elements in each column of the projection matrix, which is along the vertical direction, as depicted by Fig. 3. For example, the feature importance descriptor $\theta_i|_{i=1}^d$ to identify the j-th emotional state can be calculated as the normalized ℓ_1-norm of each element in the j-th column; namely,

$$\theta_i = \frac{|w_{ij}|}{\|\mathbf{w}_j\|_1} = \frac{|w_{ij}|}{\sum_{i=1}^d |w_{ij}|}, \tag{11}$$

where $|\cdot|$ is the absolute value operator. Essentially, equations (10) and (11) are equivalent because each row has only one element in this label-specific case.

III. Experiments

A. Data Description

Two benchmark emotional EEG data set, SEED-IV and SEED-V, are used in the following experiments. We first describe the main properties of the SEED-IV data set and then point out the differences in SEED-V.

In SEED-IV, EEG data was collected from 15 subjects when they were watching the movie clips. 72 movie clips were carefully selected to evoke the four discrete emotional states, *i.e.*, *sad*, *fear*, *happy* and *neutral*. In each of the three sessions, each subject was asked to watch 24 movie clips, among which six clips correspond to one emotional state. The EEG acquisition devices include the ESI Neuroscan system and a 62-electrode cap in compliance with the international 10-20 placement. When raw EEG data was recorded with a sampling frequency of 1000 Hz, it was first down-sampled to 200 Hz and then band-pass filtered to 1-50 Hz. In the following experiments, we use the differential entropy features which were extracted from five frequency bands, *i.e.*, *Delta* (1-3 Hz), *Theta* (4-7 Hz), *Alpha* (8-13 Hz), *Beta* (14-30 Hz) and *Gamma* (31-50 Hz). The sample vector was formed by concatenating the 62 values corresponding to each of the five frequency bands, leading to its dimensionality 310. There respectively have 851, 832 and 822 EEG samples in the three sessions.

SEED-V is also a video-evoked emotional EEG data set, which consists of five different types of emotional states. Specifically, SEED-V data set has one more state, *disgust*, in comparison with SEED-IV. 20 subjects participated the data collection experiments and the EEG data from 16 subjects was made public. In each session, three of the total 15 trials correspond to one emotional state. There are 681, 541 and 601 samples in the three sessions, respectively.

B. Experimental Setup

In the following experiments, we compare JCSFE with the several semi-supervised learning models including

- Semi-supervised Support Vector Machine (ssSVM) with linear kernel.
- Rescaled Linear Square Regression (RLSR) [21], which explicitly defines a feature importance descriptor in semi-supervised regression to characterize the different contributions of features in classification.
- Semi-supervised Linear Square Regression (ssLSR) and graph regularized ssLSR (LSRG). ssLSR is modified from RLSR, which has no feature auto-weighting ability. LSRG introduces a graph regularization into ssLSR.
- Semi-supervised Feature Selection with Redundancy Minimization (SFSRM) [25], which penalizes the redundancy in feature selection by enforcing the strongly correlated features to be far apart in feature ranking.
- Robust Discriminative Sparse Regression (RDSR) [26], in which the $\ell_{2,1}$-norm based sparse regression is used to enhance the robustness and the projection matrix is enforced to be row sparse for feature selection.
- Semi-supervised Structured Manifold Learning (SSML) [27], which proposes to learn a structured graph to exploit the submanifold of both labeled and unlabeled data to solve the multimodality problem that samples in some classes lie in several separated clusters.
- Sparse Discriminative Semi-Supervised Feature Selection (SDSSFS) [28], which improves RLSR by introducing the label dragging technique to maximize the margin between different classes.

In terms of parameter setting, the relevant parameters in each model are uniformly tuned from the candidate values $\{2^{-10}, 2^{-9}, \cdots, 2^{10}\}$. The initialization of $\mathbf{Y}_u$ in Algorithm 1 means that each sample has the same probability to all the emotional states. On the experimental paradigm, the subject-dependent cross-session EEG emotion recognition is employed. Since each subject has three different sessions in both SEED-IV and SEED-V, for each subject we consider only the three cross-session emotion recognition tasks in chronological order, *i.e.*, session1-session2, session1-session3, and session2-session3. Taking the 'session1-session2' task as an example, EEG samples from the first session serve as the labeled ones but those from the second session are unlabeled. Accordingly, we should estimate the emotional states of these unlabeled EEG samples as accurately as possible.

C. Results and Analysis

In Tables I and II, we present the recognition accuracies of these compared models, where the bold number indicates the best result of that case. s1, s2, $\cdots$, are the indices of subjects. These results provide us with the following insights.

• Obviously, JCSFE obtained the best performance among the nine compared models on average. The average accuracies of JCSFE in the three cross-session recognition tasks of SEED-IV are 80.78%, 78.55%, and 83.89%, which respectively outperform the runner-up model by 6.57%, 5.97% and 5.78%. Similarly, the average accuracies of JCSFE on the SEED-V data set are 81.90%, 81.65% and 81.33%, which also have 2%-5% improvements in comparison with the second-best one. According to the obtained results, we generally conclude that jointly exploring the label-common and label-specific EEG features is beneficial for improving the emotion recognition accuracy. Additionally, the local invariance property of data is also useful in JCSFE.

• The performance of ssSVM is generally worse than that of the remaining models. To be specific, its average accuracies

TABLE I
EMOTION RECOGNITION RESULTS (%) ON THE SEED-IV DATA SET.

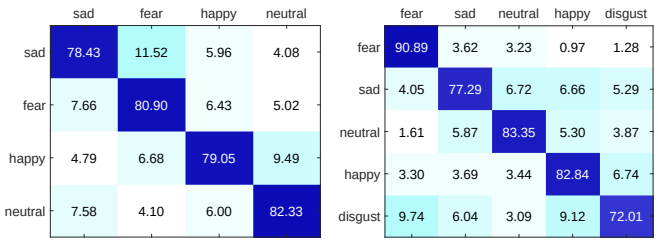
	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	Avg.
ssSVM	39.42	78.97	53.37	31.13	47.24	48.32	68.99	70.91	60.82	58.17	53.37	40.26	54.81	68.75	81.37	57.06
ssLSR	57.09	91.23	60.10	63.22	59.50	69.83	82.93	68.87	67.67	46.75	50.00	60.34	58.05	79.33	88.58	66.90
RLSR	55.65	89.18	69.71	68.39	67.67	71.03	80.77	69.95	78.73	53.85	52.04	53.13	68.63	76.92	87.14	69.52
LSRG	52.88	88.46	58.65	59.38	62.50	71.88	80.65	68.15	74.40	56.25	60.82	60.70	63.58	78.49	88.46	68.35
SFSRM	60.34	72.48	64.78	73.68	57.45	58.89	79.57	58.29	57.57	57.21	58.29	64.18	66.35	78.25	90.63	66.53
RDSR	60.22	87.62	63.46	64.30	67.91	67.19	75.00	71.39	**81.13**	68.27	60.58	56.73	67.67	75.00	97.12	70.91
SDSSFS	53.61	89.66	67.79	70.55	68.39	70.55	80.77	70.43	79.09	55.89	52.04	**70.67**	69.59	74.76	97.36	71.41
SSML	66.35	85.70	67.79	74.52	**72.84**	78.13	**89.18**	78.37	74.28	68.87	**69.95**	62.98	55.17	82.81	86.18	74.21
JCSFE	**86.30**	**96.03**	**79.45**	**78.97**	72.60	**79.21**	80.53	**84.62**	77.64	**81.49**	68.87	70.31	**74.04**	**83.53**	**98.08**	**80.78**
ssSVM	50.12	71.05	57.66	57.66	43.07	60.46	57.91	71.41	48.30	61.31	60.71	33.09	61.80	55.84	80.78	58.08
ssLSR	64.72	84.55	48.42	71.05	52.55	80.17	86.98	78.22	61.68	45.26	76.28	65.82	54.87	77.49	85.52	68.91
RLSR	70.68	89.29	48.78	71.17	58.39	83.45	**88.44**	80.78	62.77	49.64	71.17	65.45	62.41	82.85	85.40	71.38
LSRG	67.64	86.13	48.78	70.07	70.07	77.37	84.91	79.44	58.88	46.84	**81.02**	63.87	55.72	79.08	84.67	70.30
SFSRM	66.67	62.53	51.34	58.52	59.98	75.43	86.50	81.27	58.39	63.02	64.48	56.33	52.07	65.21	73.97	65.05
RDSR	69.95	86.37	54.87	68.73	60.34	83.45	82.60	78.10	**72.02**	63.99	76.76	66.18	61.31	74.94	89.05	72.58
SDSSFS	70.19	85.64	45.50	**79.32**	**74.57**	83.58	84.79	81.87	62.77	49.64	70.68	66.79	61.31	82.73	85.40	72.32
SSML	69.59	79.56	67.15	73.24	62.29	86.25	82.00	67.27	65.33	70.80	66.42	55.35	54.99	79.93	73.48	70.24
JCSFE	**76.76**	**95.99**	**67.52**	71.65	71.78	**88.08**	85.89	**82.73**	70.44	**72.51**	74.33	**70.07**	**70.19**	**90.02**	**90.27**	**78.55**
ssSVM	56.69	79.20	67.88	67.64	64.23	65.57	85.04	64.48	59.25	53.16	52.55	42.09	41.85	71.53	85.40	63.77
ssLSR	61.44	85.40	63.75	82.48	76.64	76.64	84.31	76.52	47.45	72.87	52.68	67.88	56.45	88.44	88.08	72.07
RLSR	62.65	83.21	67.27	80.17	72.87	83.70	88.56	**82.97**	61.80	78.35	59.49	63.63	64.48	87.10	89.90	75.08
LSRG	60.46	87.23	67.27	83.09	80.66	85.52	88.20	76.52	65.69	81.39	57.42	75.55	50.85	88.44	88.08	75.76
SFSRM	57.54	57.06	69.71	80.05	60.83	88.56	81.63	76.16	63.38	61.44	68.13	55.60	55.47	81.75	85.64	69.53
RDSR	65.09	82.36	72.26	75.18	75.55	79.56	86.62	72.14	**75.18**	78.10	59.61	64.36	74.94	87.83	90.27	75.94
SDSSFS	61.68	86.74	68.98	**89.54**	77.98	**93.55**	89.78	80.41	58.39	78.35	**73.60**	70.32	62.17	87.10	**93.07**	78.11
SSML	65.21	82.36	**76.40**	65.82	66.55	79.32	**95.62**	76.52	63.14	72.63	61.56	71.53	62.04	86.25	87.10	74.14
JCSFE	**72.51**	**93.67**	72.38	89.29	**85.52**	84.79	94.04	82.48	74.82	**87.83**	71.41	**82.97**	**79.08**	**97.20**	90.39	**83.89**

on the SEED-IV data set are 57.06%, 58.08%, 63.77% and they are 62.04%, 58.94% and 62.29% on the SEED-V data set. From our point of view, the linear kernel in ssSVM is not effective enough in capturing the essence of emotional information in EEG. Similarly, the performance of SFSRM is also not satisfactory. First, SFSRM performs semi-supervised feature selection by considering the $\ell_{2,1}$-norm based label-common features only. Second, the label indicator matrix in SFSRM is real-valued which cannot explicitly characterize the label information and therefore cannot effectively guide the feature selection process.

• As stated in the experimental setting, RLSR takes the adaptive feature weighting into account while ssLSR does not. Such only difference made RLSR obtain superior performance to ssLSR. Taking SEED-IV as an example, we believe that the improvements of 2.62%, 2.47%, and 3.01% achieved by RLSR are brought from the adaptive learning of the different contributions of different EEG feature dimensions to emotion recognition. Therefore, RLSR is endowed the ability to automatically identify the discriminative features while suppress the redundant and noisy features. Besides, due to the introduction of graph regularization, LSRG generally outperforms ssLSR in terms of the average performance.

• For the three recently proposed models, RDSR, SDSSFS and SSML, they have generally shown good performance in emotion recognition. For example, RDSR improves the performance by 1.39%, 1.20% and 0.86% in the three tasks of SEED-IV, in comparison with RLSR. In RLSR, direct mapping between the data matrix and the label indicator matrix is built by a row-sparse projection matrix. While in RDSR, it additionally takes the local label consistency into consideration to constrain the projection matrix. Similarly, by taking RLSR as a baseline method, SSDSFS additionally includes the label-dragging strategy to maximize the margin between classes, leading to superior performance. As for SSML, the graph learning technique is used to more effectively characterize the underlying connections of samples.

In addition, we rearranged the emotion recognition accuracies in the form of confusion matrix. In Fig. 4, we show the confusion matrices of JCSFE on the two data sets, from which we easily obtain the average recognition accuracy on each state. Taking SEED-IV for example, JCSFE acts the best recognition accuracy, 82.33%, on the *neutral* state. There are only 7.58%, 4.1% and 6% *neutral* EEG samples which are incorrectly recognized as *sad*, *fear* and *happy*, respectively.



(a) SEED-IV (b) SEED-V

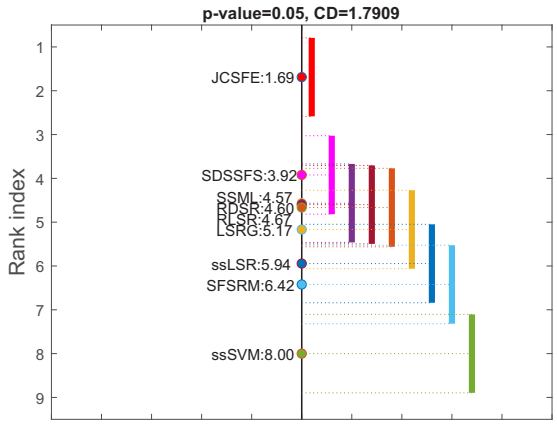
Fig. 4. The accuracies (%) of JCSFE represented by confusion matrices.

Moreover, we performed the Friedman test on the emotion recognition results to perform the statistical analysis among the compared models. The null hypothesis is that all these models share the same performance in emotion

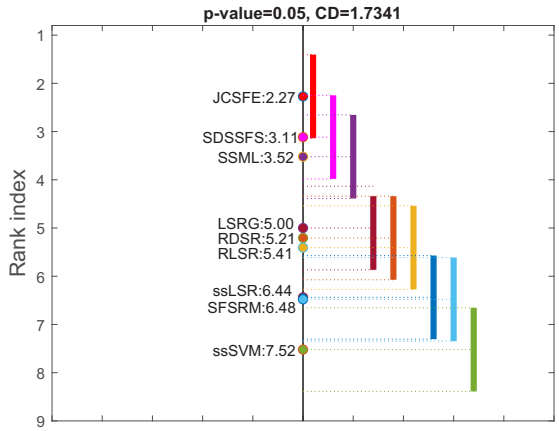
TABLE II
EMOTION RECOGNITION RESULTS (%) ON THE SEED-V DATA SET.

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	Avg.
ssSVM	76.71	52.87	57.30	60.63	61.18	64.51	67.65	63.40	69.32	52.50	59.52	61.18	71.16	68.21	56.38	50.09	62.04
ssLSR	74.68	58.96	57.49	77.82	72.27	61.00	61.55	65.62	72.09	63.40	62.29	74.49	86.32	88.17	59.70	77.63	69.59
RLSR	75.97	69.87	62.11	78.19	68.76	69.69	69.13	74.12	68.95	57.49	67.28	80.41	87.43	87.43	59.33	67.47	71.48
LSRG	84.47	66.54	62.29	85.21	82.07	73.94	65.25	75.79	70.98	63.40	65.43	76.71	76.89	89.09	65.80	80.04	73.99
SFSRM	75.60	51.76	54.71	69.32	58.23	56.19	65.99	58.96	**95.56**	64.51	61.55	62.48	74.31	68.39	63.22	59.89	65.04
RDSR	67.65	69.69	55.45	79.11	66.54	73.57	71.90	75.60	68.76	71.90	61.37	78.19	86.32	**91.68**	58.60	60.26	71.04
SDSSFS	87.80	70.98	61.37	93.90	79.30	75.97	**82.44**	80.59	79.85	60.44	75.79	81.52	89.65	88.17	56.56	80.94	77.83
SSML	87.80	73.01	77.06	**94.27**	**82.99**	**84.66**	81.55	77.45	77.63	**76.34**	**79.48**	62.48	86.14	81.33	76.89	73.01	79.51
JCSFE	**89.28**	**85.21**	**78.00**	82.62	74.86	75.23	81.89	**85.77**	85.95	68.02	70.24	**83.55**	**97.41**	86.14	**83.55**	**82.62**	**81.90**
ssSVM	57.74	53.91	52.41	65.56	51.08	67.55	75.87	75.71	80.70	56.57	51.91	49.92	61.56	53.41	36.77	52.41	58.94
ssLSR	65.89	65.72	52.25	86.02	63.56	45.09	82.70	70.38	81.53	46.59	80.03	86.02	71.38	58.90	40.27	59.23	65.97
RLSR	68.39	71.05	66.56	85.86	74.88	44.59	91.35	67.72	89.35	50.25	88.85	79.37	72.88	67.22	54.08	54.74	70.45
LSRG	72.38	66.39	50.25	90.18	69.05	62.06	86.36	71.88	80.20	46.09	82.20	84.36	**83.86**	60.73	48.75	61.90	69.79
SFSRM	73.71	49.75	66.56	84.86	53.91	48.09	73.38	54.08	80.20	55.91	85.02	61.06	64.56	59.73	65.06	64.39	65.02
RDSR	67.05	68.55	58.07	80.37	72.55	59.40	79.70	70.88	82.70	55.07	83.69	81.86	73.38	61.06	55.57	54.91	69.05
SDSSFS	75.04	68.05	78.20	**93.18**	74.88	52.58	91.18	67.89	**91.01**	54.24	**89.85**	**92.85**	77.04	64.89	54.74	68.55	74.64
SSML	74.88	77.37	74.88	83.03	74.38	64.39	69.22	81.53	89.18	**72.88**	82.36	68.22	83.36	71.88	71.38	74.54	75.84
JCSFE	**83.69**	**88.52**	**85.02**	92.01	**78.87**	**73.21**	**97.84**	**89.18**	79.03	64.56	84.53	78.04	80.87	**79.70**	**72.05**	**79.20**	**81.65**
ssSVM	53.91	51.25	74.88	60.40	57.74	64.73	60.40	64.23	73.88	56.24	60.90	80.03	63.39	43.09	52.41	**79.20**	62.29
ssLSR	90.02	86.69	64.06	63.89	52.75	66.56	78.37	87.19	87.02	41.26	72.21	77.37	73.21	45.09	60.73	69.72	69.76
RLSR	91.18	89.35	71.21	63.39	54.24	64.56	84.03	83.53	84.69	53.91	65.39	76.37	72.55	53.41	67.55	64.56	71.25
LSRG	91.18	79.37	75.54	68.72	55.74	68.05	81.20	86.86	87.02	44.59	80.53	76.21	78.37	54.74	62.90	70.22	72.58
SFSRM	89.85	88.02	70.05	67.72	60.90	72.21	63.56	81.70	75.37	60.90	73.04	86.19	74.54	49.75	61.56	62.73	71.13
RDSR	91.01	88.85	69.22	78.87	63.23	69.05	81.53	90.52	80.37	50.42	63.89	76.21	81.20	65.22	64.73	72.05	74.15
SDSSFS	**96.51**	89.35	75.54	**85.36**	66.56	68.89	**97.84**	84.86	88.35	47.75	79.37	**88.35**	**84.36**	56.41	72.21	71.15	78.30
SSML	89.02	**94.68**	74.21	72.55	72.55	75.87	76.37	**93.01**	88.02	57.74	76.37	73.04	83.19	64.89	**82.36**	69.55	77.71
JCSFE	**96.51**	88.52	**77.87**	80.03	**79.37**	**87.02**	84.36	87.69	**90.52**	**71.21**	**82.03**	76.21	79.37	**76.54**	72.05	75.04	**81.52**

recognition. If such hypothesis is rejected, we use the Nemenyi post-hoc test to tell whether two among all the nine models have significantly different performance. In this work, we have nine models and 45 cases in SEED-IV (*i.e.*, K=9, N=45). We rank the accuracies in each case in descending order and then mark the highest one as 1, and the lowest one as 9. In case of tiers, the related models share the average rank. Therefore, the average ranks of ssSVM, ssLSR, RLSR, LSRG, SFSRM, RDSR, SDSSFS, SSML, and JCSFE are 8.00, 5.94, 4.67, 5.17, 6.42, 4.60, 3.92, 4.57, and 1.69, respectively, as shown in Fig. 5a). The length of these vertical bars is termed as the critical distance, which is calculated as $\text{CD} = q_\alpha\sqrt{\frac{K(K+1)}{6N}}$, where q_α is the critical value in Tukey distribution (q_α is 3.102 when $K = 9$). We set the significance level as 0.05. Since the average ranks of JCSFE and SDSSFS are 1.69 and 3.92, their difference 2.23 is larger than the CD value 1.7909. Therefore, we conclude that there exists significant difference between their results. Intuitively, there is no overlap between the red and the purple bars in Fig. 5a). For SEED-V, the CD value is 1.7341 since $K = 9$ and $N = 48$; Accordingly, we have the statistical analysis results in Fig. 5b).



(a) SEED-IV



(b) SEED-V

Fig. 5. The statistical analysis of the compared models on the two data sets.

D. Label-common EEG Spatial-frequency Patterns

In section II-E, we explained how to quantitatively measure a feature to be a label-common one. In this section, we first show the correspondence between an EEG feature dimension and its frequency bands (channels), based on which we then investigate the label-common EEG spatial-frequency patterns in cross-session emotion recognition. Consider that θ_i is importance descriptor to define the contribution of the i-

th feature in classifying all the involved emotional states and we are given an EEG data set with p frequency bands and q channels. According to the correspondence between EEG frequency bands and feature dimensions [29], the importance of the $i|_{i=1}^{p}$-th frequency band can be calculated by

$$\omega(i) = \theta_{(i-1)\times q+1} + \theta_{(i-1)\times q+2} + \cdots + \theta_{i\times q}. \quad (12)$$

For the $j|_{j=1}^{q}$-th channel, its importance can be measured by

$$\psi(i) = \theta_j + \theta_{j+q} + \cdots + \theta_{j+(p-1)\times q}. \quad (13)$$

In both SEED-IV and SEED-V, we have five frequency bands and 62 EEG channels. Therefore, by respectively setting p to 5 and q to 62 in both rules (12) and (13), we automatically identify the critical EEG frequency bands and channels in classifying the emotional states, as illustrated in Fig. 6.

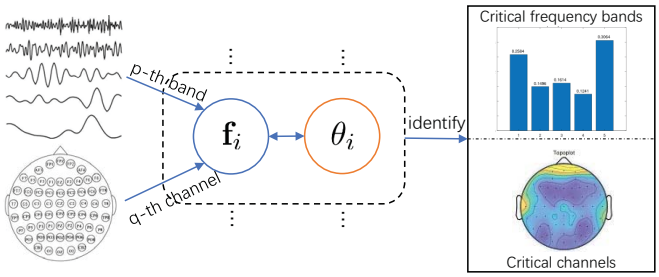
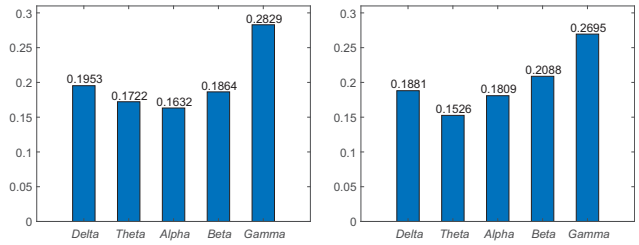


Fig. 6. The correspondence between feature dimensions and EEG frequency bands (channels) [29].

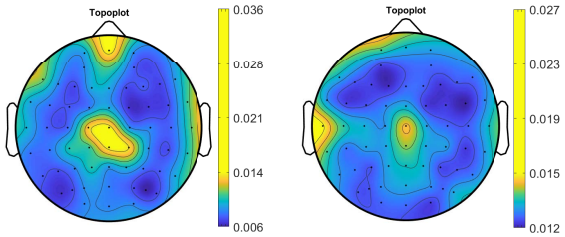
In Fig. 7, bar charts are used to show the importance of different EEG frequency bands on the SEED-IV and SEED-V data sets, and the corresponding values are marked on the top of bars. It can be seen that the *Gamma* frequency band holds the largest value; that is, the *Gamma* band generates more discriminative features than the others on average, which is undoubtedly identified as the most critical frequency band in EEG emotion recognition.



(a) SEED-IV (b) SEED-V

Fig. 7. The analysis of EEG frequency patterns on the two data sets.

Similarly, according to equation (13), it is easy to obtain the quantitative importance measure of different EEG channels. To more intuitively present the importance of different brain regions rather than listing the contributions of all the EEG channels, we use the brain topology to show how the EEG channel importance values distribute on the scalp in Fig. 8, from which we find that the spatial patterns of both data sets are generally consistent. Based on the obtained results, we roughly conclude that the four regions of the prefrontal, the left/right temporal, and the (central) parietal lobes exhibit to be more correlated to emotion recognition. The above EEG spatial-frequency activation patterns identification results are generally consistent with some existing studies [19], [29], [30].



(a) SEED-IV (b) SEED-V

Fig. 8. The analysis of EEG spatial patterns on the two data sets.

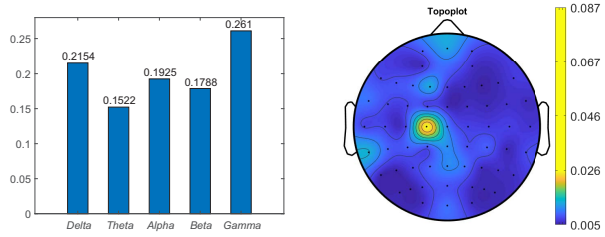
E. Label-specific EEG Spatial-frequency Patterns

Based on the normalized ℓ_1-norm label-specific feature exploration described in section II-E, below we analyze the specific EEG activation patterns associated with each of the emotional states, according to the established rules in the above subsection. Taking the SEED-IV data set for example, we respectively annotated the four emotional states of *sad*, *fear*, *happy* and *neutral* as the first, second, third and fourth classes. By using the one-hot encoding, the label indicator vector of these four emotional states are [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1], respectively. Therefore, the four columns of the projection matrix can be viewed as the feature importance descriptor respectively corresponding to these four emotional states to some extent.

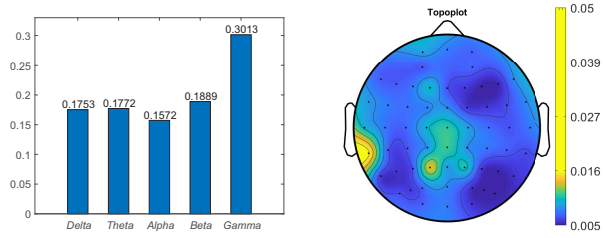
Then, according to equations (12) and (13), the spatial-frequency activation patterns associated with each emotional state are achieved, as shown in Fig. 9. From the obtained results, we find that though some common patterns are shared across different emotional states, their activation patterns are not exactly the same and there have some respective unique patterns. For example, we find the activated occipital region is common for all especially the *fear* and *happy* states in Fig. 9; however, they have differently distributed importance values of frequency bands. Generally, the importance values of frequency bands distribute similarly across the *sad*, *fear* and *neutral* states, which all have the *Gamma* band as the most important one. However, the average contributions of the *Theta* and *Gamma* bands look similar on the *happy* state. Similarly, the label-specific EEG spatial-frequency patterns on the SEED-V data set are provided in Fig. 10. Based on the above analysis, we generally conclude that it is insufficient to emphasize the label-common EEG features only in emotion recognition and it is beneficial to additionally take the label-specific features into consideration.

IV. Discussion

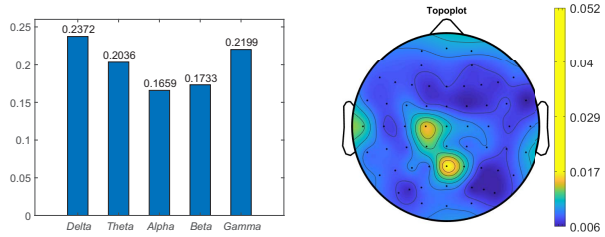
This section discusses the connections and differences between JCSFE and some existing models such as the LLSF [31], JLCLS [32], LFCMLL [33], and CLML [34]. The main common ground among these models is the utilization of the $\ell_{2,1}$-norm and the ℓ_1-norm to respectively learn label-common



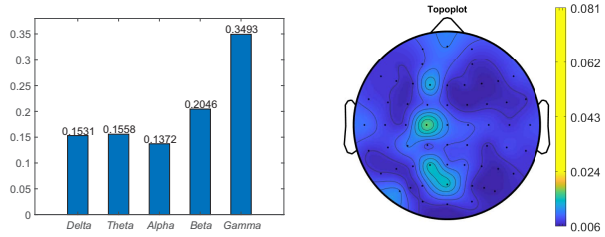
(a) The frequency (left) and spatial (right) patterns of the *sad* state.



(b) The frequency (left) and spatial (right) patterns of the *fear* state.

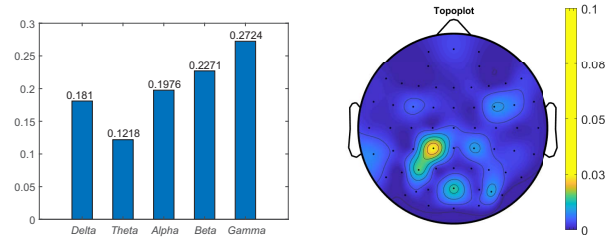


(c) The frequency (left) and spatial (right) patterns of the *happy* state.

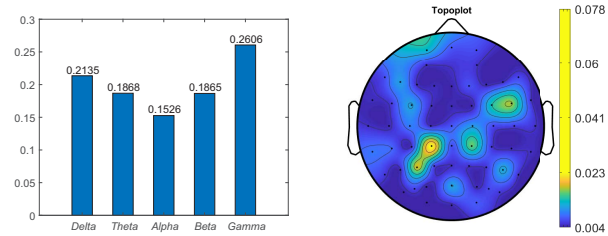


(d) The frequency (left) and spatial (right) patterns of the *neutral* state.

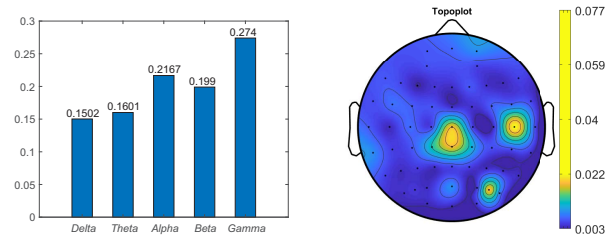
Fig. 9. Spatial-frequency patterns of each emotional state in SEED-IV.



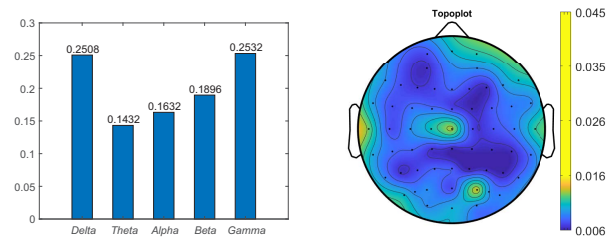
(a) The frequency (left) and spatial (right) patterns of the *sad* state.



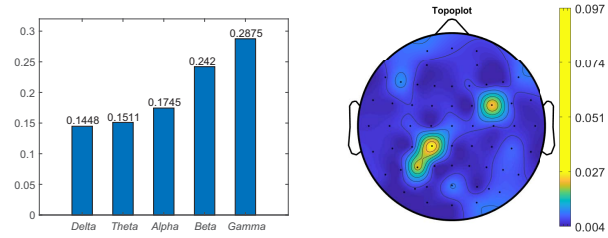
(b) The frequency (left) and spatial (right) patterns of the *fear* state.



(c) The frequency (left) and spatial (right) patterns of the *happy* state.



(d) The frequency (left) and spatial (right) patterns of the *neutral* state.



(e) The frequency (left) and spatial (right) patterns of the *disgust* state.

Fig. 10. Spatial-frequency patterns of each emotional state in SEED-V.

and label-specific features. From this point of view, our JCSFE model formulation is inspired by the existing ones. On the model optimization, most of these models use the APG method to solve the model objective functions.

The differences between JCSFE and the above mentioned models consist at least the following three aspects.

• JCSFE is a semi-supervised model by utilizing both labeled and unlabeled EEG samples in model learning, which is more effective in capturing the underlying data properties [35]. Moreover, jointly estimating the emotional states of unlabeled EEG samples and optimizing the remaining model variables can better guide the discriminative feature exploration.

• JCSFE is particularly designed for EEG emotion recognition. In the above experiments, we not only obtained improved emotion recognition performance by JCSFE, but also investigated the EEG spatial-frequency patterns from two aspects, *i.e.*, each feature in terms of all the emotional states and each emotional state in terms of all the features. However, the other models focused only on evaluating their performance on benchmark data sets by standard metrics (*e.g.*, accuracy) but paid less investigation on the problem itself.

• In the present work, the video-evoked EEG emotion recognition is a single-label pattern classification problem and each sample should be uniquely categorized into a specific emotional state. Therefore, we did not take the label correlations into consideration, which is different from these multi-label or label distribution learning models.

V. Conclusion and Future Work

In this paper, we proposed a new model term JCSFE for semi-supervised cross-session EEG emotion recognition,

which jointly explores the label-common and label-specific EEG features by respectively introducing the $\ell_{2,1}$-norm and the ℓ_1-norm based regularization terms. Moreover, a similarity graph was used to characterize the data manifold based on which the local invariance property of data was preserved. Comparative studies were performed on two emotional EEG data sets and the results demonstrated that 1) JCSFE obtained improved emotion recognition performance in comparison with the state-of-the-arts, 2) the EEG spatial-frequency patterns in emotion recognition were extensively analyzed from two aspects, *i.e.*, the patterns across all the emotional states and those associated with each emotional state. It is worth mentioning that the analysis of EEG spatial-frequency patterns in this work is completely data-driven. Though there exist consistencies between our results and some existing studies to some extent, further research from both cognitive neuroscience and information science is still necessary to validate whether they are related to the neural mechanism of affective information processing.

In the present work, we consider the cross-session emotion recognition only, which is much easier than the cross-subject setting due to the existence of inter-subject variabilities. As our future work, we will consider extending the current JCSFE model in dealing with cross-subject EEG emotion recognition. That is, possible transfer learning strategies will be improved and integrated into JCSFE to suppress the inter-subject variabilities.

REFERENCES

[1] R. Adolphs, L. Mlodinow, and L. F. Barrett, "What is an emotion?" *Curr. Biol.*, vol. 29, no. 20, pp. R1060–R1064, 2019.

[2] J. D. Parker, L. J. Summerfeldt, C. Walmsley, R. O'Byrne, H. P. Dave, and A. G. Crane, "Trait emotional intelligence and interpersonal relationships: eesults from a 15-year longitudinal study," *Pers. Indiv. Differ.*, vol. 169, p. 110013, 2021.

[3] M. E. Speer and M. R. Delgado, "Emotion–cognition interactions in memory and decision making," *Stevens' Handbook Exp. Psychol. Cognit. Neurosci.*, vol. 4, pp. 1–26, 2018.

[4] R. Vaughan, S. Laborde, and C. McConville, "The effect of athletic expertise and trait emotional intelligence on decision-making," *Eur. J. Sport Sci.*, vol. 19, no. 2, pp. 225–233, 2019.

[5] B. S. Mitchell, L. Kern, and M. A. Conroy, "Supporting students with emotional or behavioral disorders: State of the field," *Behav. Disorders*, vol. 44, no. 2, pp. 70–84, 2019.

[6] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, and N. Kumar, "EEG based emotion recognition: A tutorial and review," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 79:1–57, 2022.

[7] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, 2017.

[8] Y. Cui, Y. Xu, and D. Wu, "EEG-based driver drowsiness estimation using feature weighted episodic training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2263–2273, 2019.

[9] J. Li, N. Thakor, and A. Bezerianos, "Brain functional connectivity in unconstrained walking with and without an exoskeleton," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 3, pp. 730–739, 2020.

[10] Y. Jiang, D. Wu, Z. Deng, P. Qian, J. Wang, G. Wang, F.-L. Chung, K.-S. Choi, and S. Wang, "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2270–2284, 2017.

[11] Y. Peng, W. Wang, W. Kong, F. Nie, B.-L. Lu, and A. Cichocki, "Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1941–1958, 2022.

[12] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, 2014.

[13] G. Zhang, M. Yu, G. Chen, Y. Han, D. Zhang, G. Zhao, and Y. Liu, "A review of EEG features for emotion recognition (in Chinese)," *Scientia Sinica-Informationis*, vol. 49, no. 9, pp. 1097–1118, 2019.

[14] D. Dadebayev, W. W. Goh, and E. X. Tan, "EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4385–4401, 2022.

[15] X. Li, F. Shen, Y. Peng, W. Kong, and B.-L. Lu, "Efficient sample and feature importance mining in semi-supervised EEG emotion recognition," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 69, no. 7, pp. 3349–3353, 2022.

[16] Y. Peng, W. Kong, F. Qin, F. Nie, J. Fang, B.-L. Lu, and A. Cichocki, "Self-weighted semi-supervised classification for joint EEG-based emotion recognition and affective activation patterns mining," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[17] X. Gu, W. Cai, M. Gao, Y. Jiang, X. Ning, and P. Qian, "Multi-source domain transfer discriminative dictionary learning modeling for electroencephalogram-based emotion recognition," *IEEE Trans. Comput. Social Syst.*, vol. 9, no. 6, pp. 1604–1612, 2022.

[18] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in EEG: Advance of the last ten-year critical period," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 348–365, 2022.

[19] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, 2015.

[20] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell 2,1$-norms minimization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, pp. 1813–1821, 2010.

[21] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression." in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[22] Y. Peng, X. Zhu, F. Nie, W. Kong, and Y. Ge, "Fuzzy graph clustering," *Inf. Sci.*, vol. 571, pp. 38–49, 2021.

[23] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 379–387.

[24] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, 2018.

[25] S. Xu, J. Dai, and H. Shi, "Semi-supervised feature selection based on least square regression with redundancy minimization," in *Proc. Int. J. Conf. Neural Netw.*, 2018, pp. 1–8.

[26] P. Song, W. Zheng, Y. Yu, and S. Ou, "Speech emotion recognition based on robust discriminative sparse regression," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 343–353, 2020.

[27] X. Chen, R. Chen, Q. Wu, F. Nie, M. Yang, and R. Mao, "Semisupervised feature selection via structured manifold learning," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5756–5766, 2022.

[28] C. Wang, X. Chen, G. Yuan, F. Nie, and M. Yang, "Semisupervised feature selection with sparse discriminative least squares regression," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8413–8424, 2022.

[29] Y. Peng, F. Qin, W. Kong, Y. Ge, F. Nie, and A. Cichocki, "GFIL: a unified framework for the importance analysis of features, frequency bands and channels in EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 935–947, 2022.

[30] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, 2019.

[31] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label specific features for multi-label classification," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 181–190.

[32] Y. Wang, W. Zheng, Y. Cheng, and D. Zhao, "Joint label completion and label-specific features for multi-label learning algorithm," *Soft Comput.*, vol. 24, no. 9, pp. 6553–6569, 2020.

[33] X.-Y. Jia, S.-S. Zhu, and W.-W. Li, "Joint label-specific features and correlation information for multi-label learning," *J. Comput. Sci. Techn.*, vol. 35, no. 2, pp. 247–258, 2020.

[34] J. Li, P. Li, X. Hu, and K. Yu, "Learning common and label-specific features for multi-label classification with correlation information," *Pattern Recogn.*, vol. 121, no. 108259, pp. 1–15, 2022.

[35] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, 2019.

Cross-session Emotion Recognition by Joint Label-common and Label-specific EEG Features Exploration - Supplementary Material

Yong Peng, Honggang Liu, Junhua Li, Jun Huang, Bao-Liang Lu, and Wanzeng Kong*

APPENDIX: OPTIMIZATION TO JCSFE MODEL OBJECTIVE FUNCTION

Due to the page limit, the optimization to the JCSFE model objective function is provided in this supplementary material.

■ **The $\mathbf{Y}_u$ step**. Here we provide the detailed derivation to objective function defined on variable $\mathbf{y}^i|_{i=l+1}^n$. By denoting $\mathbf{m}^i \triangleq \mathbf{x}_i^T\mathbf{W}$, $i = l+1, l+2, \cdots, l+u$, we have

$$\min_{\mathbf{y}^i\geq\mathbf{0},\ \mathbf{y}^i\mathbf{1}_c=1} \|\mathbf{y}^i - \mathbf{m}^i\|_2^2. \tag{1}$$

The corresponding Lagrangian function is

$$\mathcal{L}(\mathbf{y}^i, \eta, \boldsymbol{\delta}) = \|\mathbf{y}^i - \mathbf{m}^i\|_2^2 - \eta(\mathbf{y}^i\mathbf{1}_c - 1) - \mathbf{y}^i\boldsymbol{\delta}^T, \tag{2}$$

where η and $\boldsymbol{\delta} \in \mathbb{R}^{1\times c}$ are two Lagrange multipliers respectively in scalar and vector forms. Below we show how both the Lagrange multipliers are determined. Suppose that the optimal solution to problem (1) is $\mathbf{y}^{i*}$, and the corresponding Lagrange multipliers are η^* and $\boldsymbol{\delta}^*$. Then, according to the Karush-Kuhn-Tucker (KKT) condition, we have the following equations and inequalities

$$\begin{cases} \forall j, \quad y_{ij}^* - m_{ij} - \eta^* - \delta_j^* = 0, & (3)\\ \forall j, \quad y_{ij}^* \geq 0, & (4)\\ \forall j, \quad \delta_j^* \geq 0, & (5)\\ \forall j, \quad y_{ij}^*\beta_j^* = 0, & (6) \end{cases}$$

where y_{ij}^* is the j-th element of vector $\mathbf{y}^{i*}$. The vector form of (3) is

$$\mathbf{y}^{i*} - \mathbf{m}^i - \eta^*\mathbf{1}_c^T - \boldsymbol{\delta}^* = \mathbf{0}. \tag{7}$$

Since we have the constraint $\mathbf{y}^i\mathbf{1}_c = 1$, the above equation can be reformulated into

$$\eta^* = \frac{1 - \mathbf{m}^i\mathbf{1}_c - \boldsymbol{\delta}^*\mathbf{1}_c}{c}. \tag{8}$$

By replacing η^* in (7) with (8), we have

$$\mathbf{y}^{i*} = \mathbf{m}^i - \frac{\mathbf{m}_i\mathbf{1}_c}{c}\mathbf{1}_c^T + \frac{1}{c}\mathbf{1}_c^T - \frac{\boldsymbol{\delta}^*\mathbf{1}_c}{c}\mathbf{1}_c^T + \boldsymbol{\delta}^*. \tag{9}$$

By denoting $\bar{\delta}^* = \frac{\boldsymbol{\delta}^*\mathbf{1}_c}{c}$ and $\mathbf{q} = \mathbf{m}^i - \frac{\mathbf{m}_i\mathbf{1}_c}{c}\mathbf{1}_c^T + \frac{1}{c}\mathbf{1}_c^T$, we can rewrite the above equation as

$$\mathbf{y}^{i*} = \mathbf{q} + \boldsymbol{\delta}^* - \bar{\delta}^*\mathbf{1}_c^T. \tag{10}$$

Accordingly, for each $j = 1, 2, \cdots, c$, we have

$$y_{ij}^* = q_j + \delta_j^* - \bar{\delta}^*. \tag{11}$$

Considering equations (4), (5), (6), and (11) together, we know that $q_j + \delta_j^* - \bar{\delta}^* = (q_j - \bar{\delta}^*)_+$, where $(f(\cdot))_+ = \max(f(\cdot), 0)$. Therefore, we have

$$y_{ij}^* = (q_j - \bar{\delta}^*)_+. \tag{12}$$

Till now, if $\bar{\delta}^*$ could be determined, $\mathbf{y}_i^*$ will be accordingly determined by (12). From (11), we have $\delta_j^* = y_{ij}^* + \bar{\delta}^* - q_j$ such that $\delta_j^* = (\bar{\delta}^* - q_j)_+$. Therefore, $\bar{\delta}^*$ can be calculated as

$$\bar{\delta}^* = \frac{1}{c}\sum_{j=1}^{c}(\bar{\delta}^* - q_j)_+. \tag{13}$$

According to the constraint $\mathbf{y}^i\mathbf{1}_c = 1$ and (12), we define the following function

$$f(\bar{\delta}) = \sum_{j=1}^{c}(q_j - \bar{\beta})_+ - 1, \tag{14}$$

and the optimal $\bar{\delta}^*$ should satisfy $f(\bar{\delta}^*) = 0$. When (14) equals to zero, the optimal $\bar{\delta}^*$ can be obtained via Newton method, namely,

$$\bar{\delta}^{(k+1)} = \bar{\delta}^{(k)} - \frac{f(\bar{\beta}^{(k)})}{f'(\bar{\beta}^{(k)})}. \tag{15}$$

It is obvious that $f(\bar{\delta})$ is a piecewise linear and monotonically increasing function. When $q_j \geq \bar{\delta}$, we have $f(\bar{\delta}) = \sum_{j=1}^{c} q_j - \bar{\delta} - 1$ and $f'(\bar{\delta}) = -1$. When $q_j \leq \bar{\delta}$, we have $f(\bar{\delta}) = -1$ and its derivative $f'(\bar{\delta}) = 0$. As a result, we obtain $f'(\bar{\delta})$ by counting the number of positive values in $(q_j - \bar{\delta})|_{j=1}^{c}$.

■ **The W step.** First, the convex optimization problem in the general APG method is defined as

$$\min_{\mathbf{W}\in\mathcal{H}} F(\mathbf{W}) = f(\mathbf{W}) + g(\mathbf{W}), \tag{16}$$

where $\mathcal{H}$ indicates the real Hilbert space. $f(\mathbf{W})$ is convex and smooth, and $g(\mathbf{W})$ is convex but typically non-smooth. $f(\mathbf{W})$ further satisfies the Lipschitz continuous condition; that is

$$\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\|_2 \leq L_f\|\Delta\mathbf{W}\|_2, \tag{17}$$

where L_f is termed as the Lipschitz constant and $\Delta\mathbf{W} = \mathbf{W}_1 - \mathbf{W}_2$. Below we propose to minimize the separable quadratic approximation sequence of $f(\mathbf{W})$ by the proximal gradient algorithm rather than minimizing it directly, which is expressed as

$$\begin{aligned} Q(\mathbf{W}, \mathbf{W}^{(t)}) =& f(\mathbf{W}^{(t)}) + \langle\nabla f(\mathbf{W}^{(t)}), \mathbf{W} - \mathbf{W}^{(t)}\rangle \\ &+ \frac{L_f}{2}\|\mathbf{W} - \mathbf{W}^{(t)}\|_2^2 + g(\mathbf{W}). \end{aligned} \tag{18}$$

By denoting $\mathbf{G}^{(t)} = \mathbf{W}^{(t)} - \frac{1}{L_f}\nabla f(\mathbf{W}^{(t)})$, we rewrite the above expression as

$$Q(\mathbf{W}, \mathbf{W}^{(t)}) = g(\mathbf{W}) + \frac{L_f}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2. \tag{19}$$

According to the JCSFE objective function and equation (16), we have

$$f(\mathbf{W}) = \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_2^2 + \gamma \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \beta \mathrm{Tr}(\mathbf{W}^T\mathbf{A}\mathbf{W}), \tag{20}$$

and

$$g(\mathbf{W}) = \alpha\|\mathbf{W}\|_1. \tag{21}$$

By combining equations (19), (20) and (21) together, we obtain the objective function in terms of variable $\mathbf{W}$ as

$$\mathbf{W} = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2 + \frac{\alpha}{L_f}\|\mathbf{W}\|_1. \tag{22}$$

According to the existing studies [1], [2], we set $\mathbf{W}^{(t)} = \mathbf{W}^{(t)} + \frac{b^{(t-1)}-1}{b^{(t)}}(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)})$ and then the convergence speed of the proximal gradient method can be accelerated to $\mathcal{O}(t^{-2})$, where sequence $b^{(t)}$ satisfies $(b^t)^2 - b^t \leqslant (b^{(t-1)})^2$ and $\mathbf{W}^{(t)}$ is the updated result at t-th iteration. It is obvious that (22) is an ℓ_1-norm regularized problem which can be solved by the following soft-shrinkage operator

$$\mathcal{S}_\varepsilon[x] = \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The ε above is usually a small positive value. This operator can be extended to vectors and matrices by applying it element-wisely. Then, by setting $\varepsilon = \frac{\alpha}{L_f}$, we can obtain $\mathbf{W}^{(t+1)}$ by solving

$$\mathcal{S}_\varepsilon[\mathbf{G}^{(t)}] = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2 + \varepsilon\|\mathbf{W}\|_1. \tag{24}$$

For $\nabla f(\mathbf{W})$, it can be obtained by taking the derivative of equation (20) with respect to $\mathbf{W}$. That is

$$\nabla f(\mathbf{W}) = \mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y} + \gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W} + \beta\mathbf{A}\mathbf{W}. \tag{25}$$

When $\mathbf{W}_1$ and $\mathbf{W}_2$ are given, we have

$$\begin{aligned}
&\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\|_2^2 \\
=&\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W} + \gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W} + \beta\mathbf{A}\Delta\mathbf{W}\|_2^2 \\
\leq&(\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| + \|\beta\mathbf{A}\Delta\mathbf{W}\|)^2 \\
=&\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\Delta\mathbf{W}\|^2 \\
&+ 2\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| \\
&+ 2\|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\beta\mathbf{A}\Delta\mathbf{W}\| \\
&+ 2\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\beta\mathbf{A}\Delta\mathbf{W}\| \\
\leq&3(\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\Delta\mathbf{W}\|^2) \\
\leq&3(\|\mathbf{X}\mathbf{X}^T\|^2\|\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|^2\|\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\|^2\|\Delta\mathbf{W}\|^2) \\
=&3(\|\mathbf{X}\mathbf{X}^T\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|^2 + \|\beta\mathbf{A}\|^2)\|\Delta\mathbf{W}\|^2
\end{aligned} \tag{26}$$

By comparing inequalities (17) and (26), the Lipschitz constant L_f can be set as

$$L_f = \sqrt{3(\|\mathbf{X}\mathbf{X}^T\|_2^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|_2^2 + \|\beta\mathbf{A}\|_2^2)}. \tag{27}$$

When $\mathbf{W}$ is given, then $\mathbf{A}$ is fixed and further L_f is a constant value.

REFERENCES

[1] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 379–387.

[2] J. Li, P. Li, X. Hu, and K. Yu, "Learning common and label-specific features for multi-label classification with correlation information," *Pattern Recogn.*, vol. 121, no. 108259, pp. 1–15, 2022.