Exploiting Low-Rank Latent Gaussian Graphical Model Estimation for Visual Sentiment Distributions

Yuting Su, Wei Zhao, Peiguang Jing*, Liqiang Nie

***Abstract*—Currently, an increasing number of applications and services has encouraged users to openly express their emotions via images. Unlike visual sentiment classification, visual sentiment distribution learning exploits the overall distribution to represent the relative importance of sentiment labels. Considering that most relevant studies have failed to completely model correlation structures or explicitly apply them to unknown instances, in this paper, we proposed a low-rank latent Gaussian graphical model estimation (LGGME) method for visual sentiment distribution learning tasks. There are three main characteristics of LGGME: 1) an integrated inverse covariance matrix whose parameters characterize the latent correlation structures between and within features and sentiments is estimated based on the sparse Gaussian graphical model; 2) a multivariate normal assumption is assigned on the concatenated latent feature representations and the estimated sentiment distributions instead of the original observations for a reasonable surrogate; 3) the latent feature representations are projected from a low-rank subspace, which is also available for unseen instances, and the estimated sentiment distributions are evaluated by KL divergence to ensure a suitable setting for distribution learning. We further developed an effective optimization algorithm based on the alternating direction method of multipliers (ADMM) for our objective function. The experimental results obtained on three publicly available datasets demonstrate the superiority of our proposed method.**

***Index Terms*—Visual sentiment analysis, Gaussian graphical model, low-rank representation.**

I. Introduction

Currently, the tremendous increase in the use of mobile devices and applications provides users with more convenient ways to share and view images whenever and wherever they require. Images have become an indispensable medium to express users' emotions and opinions in daily life. Under these circumstances, a growing number of approaches have been proposed in visual sentiment analysis due to their potential value in practical applications, such as facial multimedia retrieval [1][2], emotion recognition [3][4], image annotation [5][6], and multimodal text analysis [7][8][9]. For instance, Yang *et al.* [1] proposed a multitask deep framework to jointly solve visual sentiment retrieval and classification tasks. Farzaneh *et al.* [4] proposed a discriminant distribution-agnostic loss approach to address the category imbalance problem in facial expression recognition. Ji *et al.* [8] proposed a two-layer multimodal hypergraph framework to address the challenge of modality missing in microblog sentiment prediction.

In recent studies on sentiment analysis, researchers focused primarily on assigning an instance with one dominant sentiment or multiple sentiments with initially equal intensity [10][11][12][13]. For instance, Borth *et al.* [14] proposed a set of visual sentiment concepts and associated classifiers to offer images with mid-level semantic representations. She *et al.* [11] developed a weakly supervised coupled convolutional network for sentiment multi-label learning, in which the detection and the classification are integrated together. Zhao *et al.* [12] proposed an attention-based polarity-consistent deep attention network to solve fine-grained visual sentiment regression tasks. Although remarkable achievements have been realized, images are reasonably characterized by a mixture of multiple sentiments with various intensities in reality, *i.e.,* label ambiguity, attributed to the complexity of human emotion. That is, users are concerned not only with which sentiments are assigned to an instance but also with the relative importance of assigned sentiments.

In response to this issue, label distribution learning (LDL) [15] is considered in relation to the label ambiguity problem in sentiment analysis. LDL provides a general framework of estimating real-valued description degrees for each label, in which both single-label learning (SIL) and multi-label learning (MLL) can be regarded as a special case. Recently, increasing attention has been devoted to exploring sentiment distribution learning of images [16][17][18]. For example, Yang *et al.* [16] proposed two extended versions of the conditional probability neural network (CPNN) by taking label binary encoding and label enhancement into account. Ren *et al.* [18] proposed selecting label-shared and label-specific features to enhance the performance of visual sentiment distribution learning. Distinct correlations exist between sentiment labels. Plutchik *et al.* [19] also revealed that sentiments are correlated with each other from the psychology perspective. Inspired by the positive effects of label correlation in various earlier studies, many methods have been developed to solve sentiment distribution learning by mining sentiment correlations. For instance, Zhou *et al.* [17] proposed to explore associations of sentiment pairs with Pearson's correlation coefficients. Zhang *et al.* [20] presented a sentiment Bayesian network to characterize the relationship between sentiments and object semantics simultaneously.

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61802277) and the Tianjin Municipal Natural Science Foundation (Grant No. 20JCQNJC01210).

Yuting Su, Wei Zhao, and Peiguang Jing are with the School of Electrical and Information Engineering, Tianjin University, 300072, China (e-mail:ytsu@tju.edu.cn; echozhaowei@gmail.com, pgjing@tju.edu.cn). Liqiang Nie is with the School of Computer Science and Technology, Shandong University, Shandong 250101, China (e-mail: nieliqiang@gmail.com). Peiguang Jing is the corresponding author (pgjing@tju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3140892, IEEE Transactions on Multimedia

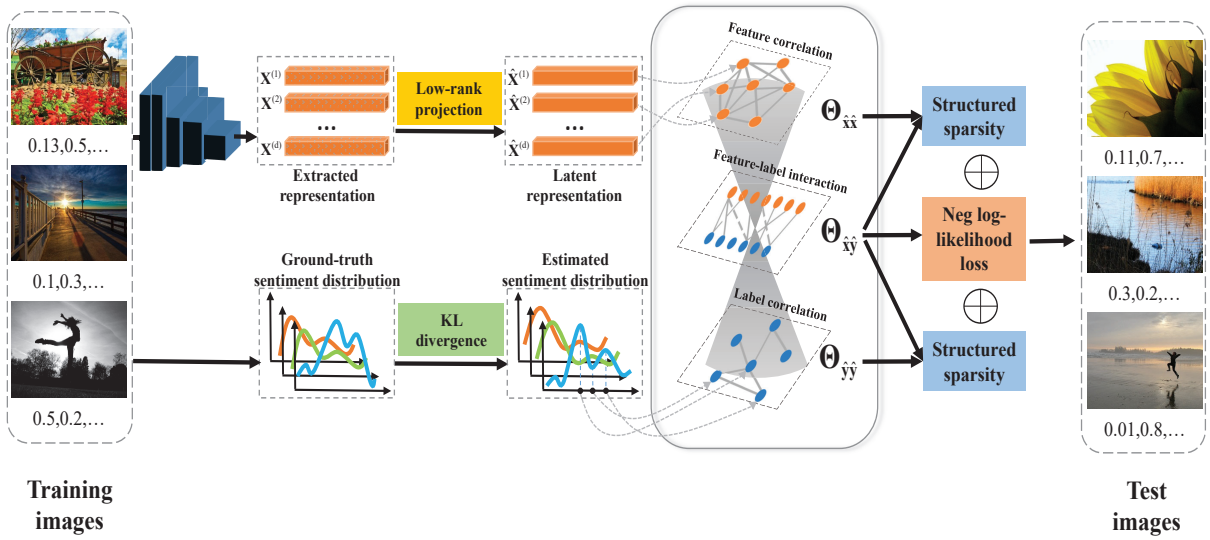


Fig. 1: Illustration of the proposed LGGME method for visual sentiment distribution learning. LGGME characterizes correlation structures between and within latent feature representations and sentiment distributions through an integrated inverse covariance matrix, in which the involved parameter matrices $\mathbf{\Theta}_{\hat{x}\hat{x}}$, $\mathbf{\Theta}_{\hat{x}\hat{y}}$, and $\mathbf{\Theta}_{\hat{y}\hat{y}}$ reflect the inner- and interdependency patterns of the latent representations and the estimated sentiment distributions. Specifically, the latent feature representations are obtained with a low-rank projection matrix, and the estimated sentiment distributions are measured via KL divergence.

More recently, Yang *et al.* [21] proposed a novel well-grounded circular-structured representation method for visual emotion distribution learning, which constructs an emotion circle to represent emotion distributions through three emotion attributes and two emotion properties. The low-rank technique is also considered an effective manner to learn more intrinsic representations and capture various correlation structures of sentiments [22][23][24]. With such considerations, Xu *et al.* [23] proposed to exploit global low-rank structure to learn the overall correlation pattern of labels; Ren *et al.* [24] proposed capturing the global and local label correlation structures by exploiting low-rank approximation and clustering strategies, respectively. Although different criteria are used to improve the performance of sentiment distribution learning, they tended to explore these structures in indirect forms, resulting in an inability to reuse them for unseen instances. Moreover, few studies have considered comprehensive correlation structures between and within features and sentiments simultaneously.

Regarding the abovementioned motivations, in this paper, we proposed a low-rank latent Gaussian graphical model estimation (LGGME) method for visual sentiment distribution learning, in which multiple correlation structures between and within features and sentiments can be intuitively characterized and reused for new instances. Specifically, we first considered assigning a multivariate normal assumption on the concatenated feature representations and sentiment distributions. On this basis, we not only replaced the original inputs and outputs with the latent feature representations and the estimated sentiment distributions for a reasonable surrogate but also characterized different types of correlation structures by estimating different parameter matrices of an integrated inverse covariance matrix with a sparse Gaussian graphical model. Notably, the latent feature representations are projected from a low-rank subspace for intrinsic low-dimensional representations, and the estimated sentiment distributions are evaluated by the KL divergence to ensure suitability for the distribution learning setting. We further developed an effective optimization algorithm based on the alternating direction method of multipliers (ADMM). The experimental results obtained on different datasets demonstrate the superiority of our proposed method. Fig. 1 illustrates the framework of our proposed method for visual sentiment distribution learning.

The main contributions are summarized as follows:

- We proposed a low-rank latent Gaussian graphical model estimation method to estimate sentiment distributions of images. In our method, the correlation structures embedded in features and sentiment labels are characterized by different parts of an intergraded inverse covariance matrix with a sparse Gaussian graphical model.
- To seek a reasonable surrogate, we assigned a multivariate normal assumption on the concatenated latent feature representations and the estimated sentiment distributions instead of the original features and distributions. Specifically, we learned a low-rank subspace to derive the low-dimensional latent feature representations and exploited the KL divergence to restrict the estimated sentiment distributions for a suitable measurement.
- To ensure fast convergence, we developed an effective optimization algorithm based on ADMM. The experimental results obtained on three datasets demonstrate the effectiveness of our proposed method.

The remainder of this paper is organized as follows: Section II first briefly reviews the related research with respect to multivariate regression and label distribution learning. Section III then describes the proposed algorithm. Finally, the

experimental results are reported in section IV, followed by the conclusion in Section V.

II. Related Work

A. Multivariate Regression

Multivariate regression is the generalization of the classical univariate regression, which is widely used in data mining and computer vision fields. Given D-dimensional input $\mathbf{X} \in R^{N\times D}$, the corresponding M-dimensional output is $\mathbf{Y} \in R^{N\times M}$, where N is the number of instances. When adding the sparsity constraint, the multivariate regression aims to learn the coefficient matrix $\mathbf{B} \in R^{D\times M}$ as follows:

$$\min_{\mathbf{B}} \frac{1}{2}\mathrm{Tr}((\mathbf{Y}-\mathbf{XB})(\mathbf{Y}-\mathbf{XB})^T) + \lambda \left\|\mathbf{B}\right\|_1, \tag{1}$$

where λ controls the sparsity degree and $\left\|\cdot\right\|_1$ is the $\mathcal{L}_1$-norm.

Intrinsically, the multivariate regression is still an integration of several separate univariate regression problems; thus, it fails to take the correlations of outputs into account. Taking the correlation structure into account, one representative method is multivariate regression with covariance estimation (MRCE) [25], which aims to estimate the sparse coefficient matrix $\mathbf{B}$ and the noise inverse covariance matrix $\mathbf{\Omega}$ by minimizing the negative log-likelihood function as follows:

$$\begin{aligned}\min_{\mathbf{B},\mathbf{\Omega}} & -N\log|\mathbf{\Omega}| + \frac{1}{2}\mathrm{Tr}((\mathbf{Y}-\mathbf{XB})\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB})^T) \\ & + \lambda_1\left\|\mathbf{B}\right\|_1 + \lambda_2\left\|\mathbf{\Omega}\right\|_1,\end{aligned} \tag{2}$$

where $|\cdot|$ denotes the determinant of the matrix and λ_1 and λ_2 are fine-tuning parameters. Although MRCE models the prediction error relationship through $\mathbf{\Omega}$, the correlation structures of MRCE are incomplete because the regression matrix $\mathbf{B}$ selects reliant factors of inputs for each output independently.

The probabilistic graphical model (PGM) [26] provides additional insight into characterizing the statistical relationship between variables in the form of a graph. When the $\mathcal{L}_1$-norm penalty for the inverse covariance matrix is considered, the graphical Lasso [27] is formulated as follows:

$$\min_{\mathbf{\Theta}\succ 0} -\log|\mathbf{\Theta}| + \mathrm{Tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) + \lambda\left\|\mathbf{\Theta}\right\|_1, \tag{3}$$

where $\mathbf{\Theta}$ and $\hat{\mathbf{\Sigma}}$ are the empirical and estimated inverse covariance matrices, respectively. Subsequently, different regularization penalties are exploited to extend the graphical Lasso. For instance, Obozinski *et al.* [28] studied the block $\mathcal{L}_1/\mathcal{L}_2$-norm regularization penalty for multivariate linear regression. Danaher *et al.* [29] proposed a fused-graphical constraint to encourage different graphical models to share certain characteristics. For more complicated group structures, tree-guided group Lasso [30] utilizes a weighted norm to ensure all labels are penalized in a balanced manner. Graph-guided fused Lasso (GFlasso) [31] provides a statistical framework to encourage highly correlated labels sharing a common set of features.

B. Label Distribution Learning

Label distribution learning is presented as a generalized paradigm to address the label ambiguity problem, which arises in different applications, such as sentiment distribution recognition [17][18][32][33], age estimation [34][35][36], pose estimation [37][38], and crowd counting [39][40]. For example, Zhao *et al.* [41] provided a comprehensively review on the development of affective image content analysis, in which several representative approaches on emotional distribution learning are summarized. Unlike traditional multilabel learning, LDL not only assigns multiple labels to describe instances, but also identifies the relative importance of assigned labels. Current LDL studies can be roughly classified into three groups: algorithm adaptation, problem transformation, and specialized algorithms.

Algorithm adaptation aims to extend existing models to fit label distribution learning. Problem transformation attempts to transform LDL into classical learning problems, making it conveniently solved with existing classifiers. Both algorithm adaptation and problem transformation are extensions of traditional machine learning, while specialized algorithms are designed to directly match the LDL problem. Representative specialized algorithms include but are not limited to LDLSF [18], LDL-LCLR [24], EDL-LRL [32], and LDLLC [42]. For instance, LDLSF [18] implicitly explores the complicated relationship among features and labels. EDL-LRL [32] implicitly exploits the label correlations locally by enforcing low-rank constraints on the clustered labels. Both LDLSF and EDL-LRL are unable to transform the learned correlation information to unknown instances. LDLLC [42] measures the similarity of labels by encoding the label correlation into a distance measurement. LDL-LCLR [24] constructs a label correlation matrix to capture both global and local correlations explicitly. Although promising performances have been achieved, the correlation structures of features and labels have not been comprehensively or explicitly characterized in LDL algorithms. Noticeably, as for visual sentiment distribution learning, several approaches that focus on characterizing the correlation between and within visual content and sentiment labels have been proposed recently. For example, Yang *et al.* [43] proposed constructing a scene-object interrelated visual emotion reasoning network to capture the emotional relationships. Xu *et al.* [44] proposed a novel emotion distribution learning method by exploring the emotion-related regions of images. Our proposed method differs in that we explore the explicit correlation patterns based on a low-rank regularized Gaussian graphical model.

III. Proposed Methodology

A. Problem Formulation

Assuming a set of N images $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$ is available, together with centralized sentiment distributions $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]^T \in \mathbb{R}^{N\times m}$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the d-dimensional feature vector of the i-th image and $\mathbf{y}_i \in R^m$ is the corresponding m-dimensional sentiment distribution. We regard visual sentiment learning as a general

multivariate regression problem as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{4}$$

where $\mathbf{B} \in \mathbb{R}^{d\times m}$ is the multivariate regression coefficient matrix and $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N]^T \in \mathbb{R}^{N\times m}$ is the error matrix whose rows are distributed with mean $\mathbf{0} \in R^m$ and covariance $\mathbf{\Omega} \in R^{m\times m}$. Actually, solving the multivariate regression problem is equivalent to learning several separate regression models since the correlation of outputs is ignored.

Gaussian graphical models provide insight into the statistical relationship between the variables of interest in the form of a graph. To comprehensively consider the dependencies between and within features and sentiments, we tend to concatenate visual features and sentiment distributions for comprehensive analysis. However, it is not always practical to restrict the original observations into normally distributed cases. To address this problem, we propose using the latent representations $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_N]^T$ and the estimated sentiment distributions $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_N]^T$ instead of the original observations. Without loss of generality, randomly given a concatenated vector $\begin{pmatrix}\hat{\mathbf{x}}_i \\ \hat{\mathbf{y}}_i\end{pmatrix}$, a multivariate Gaussian distribution with the zero mean vector and the covariance matrix $\mathbf{\Sigma}$ is then formulated as follows:

$$\begin{pmatrix}\hat{\mathbf{x}}_i \\ \hat{\mathbf{y}}_i\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right). \tag{5}$$

For simplicity, we define the inverse covariance matrix as $\mathbf{\Sigma}^{-1} = \mathbf{\Theta} = \begin{pmatrix}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} & \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \\ \mathbf{\Theta}^{\mathbf{T}}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} & \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\end{pmatrix} \in R^{(d+m)\times(d+m)}$, in which the inverse covariance parameters $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$, $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$, and $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ reflect the conditional dependencies of variables. The conditional distribution of $\hat{\mathbf{y}}_i|\hat{\mathbf{x}}_i$ in Gaussian form is further derived as follows:

$$\hat{\mathbf{y}}_i|\hat{\mathbf{x}}_i \sim \mathcal{N}(-\mathbf{\Theta}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\mathbf{\Theta}^{T}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\hat{\mathbf{x}}_i, \mathbf{\Theta}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}). \tag{6}$$

By comparing Eq. (4) with Eq. (6), the correspondence between multivariate regression and inverse covariance estimation can be naturally connected with $\mathbf{B} = -\mathbf{\Theta}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\mathbf{\Theta}^{T}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ and $\mathbf{\Omega} = \mathbf{\Theta}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$. We consider the inverse covariance parameters as a refined version of the multivariate regression parameters, which are qualified for capturing the structural sparsity of interactions of features and sentiments as well as the sparse dependencies between labels when further enforcing the $\mathcal{L}_1$-norm.

B. Proposed Method

*1) **Latent feature representation modeling**:* Considering that the original observations may not always satisfy the multivariate Gaussian distribution, we aim to build an embedding that maps the the original observations to a latent subspace. Inspired by the superior performance of the low-rank technique in exploiting low-dimensional intrinsic representations and suppressing the noise interference, we decompose the original observations into a principal part and a sparse error part. Instead of directly imposing the low-rank constraint on the principal part, we learn a low-rank regularized projection matrix to ensure that the original observations are mapped into a latent low-dimensional intrinsic subspace. In this way, the descriptive information and the underlying low-rank structure from seen instances can be propagated and adapted to unseen instances. Thus, we formulate the feature embedding module as follows:

$$\begin{aligned}&\min_{\mathbf{D},\mathbf{E}} \mathrm{Rank}(\mathbf{D}) + \lambda \left\|\mathbf{E}\right\|_1 \\ &s.t.\ \mathbf{X} = \hat{\mathbf{X}} + \mathbf{E}, \hat{\mathbf{X}} = \mathbf{XD},\end{aligned} \tag{7}$$

where $\lambda > 0$ is a trade-off parameter; $\mathbf{D}$ is the projection matrix used to capture the low-rank structure embedded in original observations; and $\mathbf{E}$ is an error term used to relax the tight equality constraint. Here, the $\mathcal{L}_1$-norm is adopted to characterize the error term $\mathbf{E}$ for random corruptions in observations. We replace the rank function $\mathrm{Rank}(\cdot)$ with the nuclear norm $\|\cdot\|_*$, which is defined as the sum of the singular values of the target matrix. This is a commonly used practice in rank minimization to approximate the rank constraint by the nuclear norm [45].

Motivated by the work of Witten and Tibshirani [46], the feature correlation is encouraged by the inverse covariance parameters $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$. We encode the correlation structure of latent feature representations through the regularization term $\mathrm{Tr}(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\hat{\mathbf{X}}^T)$. By further combining with Eq. (7), the latent feature representation modeling component $\mathcal{L}_{\mathbf{X}\to\hat{\mathbf{X}}}$ is formulated as follows:

$$\begin{aligned}&\mathcal{L}_{\mathbf{X}\to\hat{\mathbf{X}}} \\ =&\frac{1}{2}\mathrm{Tr}(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\hat{\mathbf{X}}^T) + \lambda_1 \left\|\mathbf{D}\right\|_* + \lambda_2 \left\|\mathbf{E}\right\|_1 + \lambda_3 \left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\right\|_1,\end{aligned} \tag{8}$$

where $\left\|\mathbf{D}\right\|_*$ could be represented as

$$\left\|\mathbf{D}\right\|_* = \sum_i \sigma_i(\mathbf{D}) \tag{9}$$

where λ_1, λ_2, and λ_3 are trade-off parameters. Noticeably, the $\mathcal{L}_1$-norm imposed on $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ controls the sparsity in the graphical structure, making more distinct latent representations achievable by a few strongly related factors.

*2) **Estimated sentiment distribution modeling**:* In accordance with the distribution learning setting, the KL divergence is exploited to measure the distance between the real sentiment distributions $\mathbf{Y}$ and the estimated sentiment distributions $\hat{\mathbf{Y}}$. Similarly, the inverse covariance parameter matrix $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is also exploited to enhance the label correlations through the regularization term $\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T)$. Furthermore, the $\mathcal{L}_1$-norm is enforced on $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ to extract the significant correlations among labels. Thus, we can formulate the estimated sentiment distribution modeling component as follows:

$$\begin{aligned}&\mathcal{L}_{\mathbf{Y}\to\hat{\mathbf{Y}}} \\ =&\frac{1}{2}\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T) + \gamma\mathrm{KL}(\mathbf{Y} \parallel \hat{\mathbf{Y}}) + \lambda_4 \left\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\right\|_1,\end{aligned} \tag{10}$$

where γ is a trade-off parameter and λ_4 controls the sparsity of $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$.

*3) **Latent Gaussian graphical model**:* After obtaining the latent feature representations and the estimated sentiment distributions, we resort to Eq. (6) and formulate a conditional Gaussian graphical model as follows:

$$p(\hat{\mathbf{y}}_i|\hat{\mathbf{x}}_i) = \frac{\exp\left(\frac{1}{2}\hat{\mathbf{y}}_i\mathbf{\Theta}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_i\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\hat{\mathbf{y}}_i\right)}{\mathbf{Z}(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \hat{\mathbf{x}}_i)}, \tag{11}$$

where

$$\begin{aligned}\mathbf{Z}(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \hat{\mathbf{x}}_i) &= \int \exp\left(-\frac{1}{2}\hat{\mathbf{y}}_i \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \hat{\mathbf{y}}_i - \hat{\mathbf{x}}_i \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \hat{\mathbf{y}}_i\right) \mathrm{d}\hat{\mathbf{y}}_i \\ &= \sqrt{\frac{(2\pi)^M}{|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}|}} \exp\left(\frac{1}{2}\hat{\mathbf{x}}_i^T \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{x}}_i\right)\end{aligned}$$

is the normalization term ensuring that the sum of label description degrees of any instance equals one. Importantly, $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is restricted to be nonnegative definite so that the normalization term is finite and the corresponding conditional probability distribution is well defined.

By minimizing the negative log-likelihood function of Eq. (11) with the inverse covariance parameters $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ and $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$, we have

$$\begin{aligned}\min_{\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}} & -\frac{N}{2}\log|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}| + \frac{1}{2}\mathrm{Tr}(\hat{\mathbf{X}}^T \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T) \\ & + \frac{1}{2}\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T) + \mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T).\end{aligned} \tag{12}$$

We further consider the $\mathcal{L}_1$-norm penalty on $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$, which implies a structured sparsity between features and sentiment labels. Therefore, we have

$$\begin{aligned}&\mathcal{L}_{\hat{\mathbf{X}}\to\hat{\mathbf{Y}}} \\ =& -\frac{N\eta}{2}\log|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}| + \frac{\eta}{2}\mathrm{Tr}(\hat{\mathbf{X}}^T \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T) \\ &+ \frac{\eta}{2}\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T) + \eta\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T) + \lambda_5 \left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\right\|_1,\end{aligned} \tag{13}$$

where η is a trade-off parameter and λ_5 controls the sparsity of $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$.

Thus, the final objective function Eq. (14) is formulated by combining Eq. (8), Eq. (10), and Eq. (13). From Eq. (14), we can observe that the inverse covariance parameter matrices $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$, $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$, and $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ jointly estimate label distributions. $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ influences labels by features directly, and $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ builds a direct influence on labels throughout its nonzero off-diagonal elements. $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ encourages highly related features to be shared by associated labels.

C. Optimization

The alternating direction method of multipliers (ADMM) is exploited to solve the optimization of Eq. (14). Specifically, we first introduce an auxiliary variable $\mathbf{J}$ to relax the original objective function and formulate the augmented Lagrangian function $\mathcal{L}(\cdot)$. We then divide it into several subproblems and alternatively minimize each of the subproblems with other variables fixed. As the proposed LGGME method mainly learns the low-rank projection matrix and the sparse inverse covariance parameters, we adopt a phased-updating rule to ensure a nearly stable projection matrix of nested iterations as well as rapid convergence during outer iterations.

The augmented Lagrangian function is obtained as follows:

$$\begin{aligned}&\mathcal{L}(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}, \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \mathbf{D}, \mathbf{J}, \mathbf{E}, \hat{\mathbf{Y}}) = \\ &\gamma\mathrm{KL}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}}\right) - \frac{N\eta}{2}\log|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}| + \frac{\eta}{2}\mathrm{Tr}\left(\hat{\mathbf{X}}^T \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T\right) \\ &+ \frac{\eta+1}{2}\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T\right) + \frac{1}{2}\mathrm{Tr}\left(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\hat{\mathbf{X}}^T\right) + \eta\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \hat{\mathbf{X}}^T\right) \\ &+ \lambda_1\|\mathbf{J}\|_* + \lambda_2\|\mathbf{E}\|_1 + \lambda_3\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\|_1 + \lambda_4\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\|_1 + \lambda_5\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\|_1 \\ &+ \frac{\rho}{2}\left\|\mathbf{D} - \mathbf{J} + \frac{\mathbf{\Gamma}_1}{\rho}\right\|_F^2 + \frac{\rho}{2}\left\|\mathbf{X} - \mathbf{X}\mathbf{D} - \mathbf{E} + \frac{\mathbf{\Gamma}_2}{\rho}\right\|_F^2,\end{aligned} \tag{15}$$

where $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are Lagrange multipliers; ρ is the penalty parameter for the quadratic term.

1) Update $\mathbf{D}$: with other fixed variables, the update of $\mathbf{D}$ is given by minimizing the following optimization problem:

$$\begin{aligned}\min_{\mathbf{D}}\ & \frac{1}{2}\mathrm{Tr}\left(\hat{\mathbf{X}}(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} + \eta\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T)\hat{\mathbf{X}}^T\right) + \eta\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T\hat{\mathbf{X}}^T) \\ &+ \frac{\rho}{2}\left\|\mathbf{D} - \mathbf{J} + \frac{\mathbf{\Gamma}_1}{\rho}\right\|_F^2 + \frac{\rho}{2}\left\|\mathbf{X} - \mathbf{X}\mathbf{D} - \mathbf{E} + \frac{\mathbf{\Gamma}_2}{\rho}\right\|_F^2.\end{aligned} \tag{16}$$

Note that $\mathbf{D}$ can be effectively solved by the limited-memory BFGS (L-BFGS) algorithm, which only requires the computation of the first-order gradient instead of the inverse Hessian matrix as follows:

$$\begin{aligned}\nabla\mathbf{D} =& \mathbf{X}^T\mathbf{X}\mathbf{D}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^T + \frac{\eta}{2}\mathbf{X}^T\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T + \eta\mathbf{X}^T\mathbf{X}\mathbf{D}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T \\ &+ \rho\left(\mathbf{D} - \mathbf{J} + \frac{\mathbf{\Gamma}_1}{\rho}\right) - \rho\mathbf{X}^T\left(\mathbf{X} - \mathbf{X}\mathbf{D} - \mathbf{E} + \frac{\mathbf{\Gamma}_2}{\rho}\right).\end{aligned} \tag{17}$$

2) Update $\hat{\mathbf{Y}}$: With other fixed variables, the update of $\hat{\mathbf{Y}}$ is given by minimizing the following optimization problem:

$$\min_{\hat{\mathbf{Y}}} \gamma\mathrm{KL}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}}\right) + \frac{\eta+1}{2}\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^T\right) + \eta\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^T\hat{\mathbf{X}}^T\right). \tag{18}$$

Similar to $\mathbf{D}$, $\hat{\mathbf{Y}}$ can be solved via L-BFGS, and its first-order gradient is written as follows:

$$\nabla\hat{\mathbf{Y}} = (\eta+1)\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} + \eta\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} + \gamma\mathbf{Y}\circ\mathbf{H}, \tag{19}$$

where $\circ$ denotes the Hadamard product and $\mathbf{H} \in \mathbb{R}^{d\times m}$ satisfies $H_{ij} = 1/\hat{y}_{ij}$.

3) Update $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$: With other fixed variables, the update of inverse covariance parameter matrix $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ is obtained by solving the following problem:

$$\min_{\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}} \mathrm{Tr}\left(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\hat{\mathbf{X}}^T\right) + \lambda_3\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\right\|_1. \tag{20}$$

The $\mathcal{L}_1$-norm constrained $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ can be updated by the shrinkage operator according to the following rules:

$$\begin{aligned}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{(k+1)} &= \arg\min_{\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}} \lambda_3\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\right\|_1 + \frac{1}{2t_k}\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} - \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{(k)} - \frac{t_k}{2}\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right\|^2 \\ &= \mathcal{S}_{\lambda_3 t_k}\left(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{(k)} - \frac{t_k}{2}\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right),\end{aligned} \tag{21}$$

where k is the k-th iteration of the nested optimization of $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$; t_k is a suitable step-size for the k-th iteration, which can be searched by backtracking rules; and $\mathcal{S}_\alpha(\cdot)$ is the shrinkage operator [47]. To accelerate the convergence and the

$$\begin{aligned}\min_{\mathbf{D},\mathbf{\Theta},\mathbf{E},\hat{\mathbf{Y}}} & \gamma \mathrm{KL}(\mathbf{Y} \mid\mid \hat{\mathbf{Y}}) - \frac{N\eta}{2}\log|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}| + \frac{\eta+1}{2}\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^{\mathbf{T}}) + \frac{1}{2}\mathrm{Tr}(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\hat{\mathbf{X}}^{\mathbf{T}}) + \frac{\eta}{2}\mathrm{Tr}(\hat{\mathbf{X}}^{\mathbf{T}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{\mathbf{T}}) \\ & + \eta\mathrm{Tr}(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{T}) + \lambda_1\|\mathbf{D}\|_* + \lambda_2\|\mathbf{E}\|_1 + \lambda_3\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\|_1 + \lambda_4\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\|_1 + \lambda_5\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\|_1 \\ & s.t.\ \mathbf{X} = \hat{\mathbf{X}} + \mathbf{E}, \hat{\mathbf{X}} = \mathbf{X}\mathbf{D}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} \succeq \mathbf{0}\end{aligned} \tag{14}$$

computational simplicity of the $\mathcal{L}_1$-norm regularized problem, we apply the fast iterative shrinkage-thresholding algorithm (FISTA) [48].

4) Update $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$: With fixed other variables, the update of $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ is obtained by solving the following problem:

$$\begin{aligned}\min_{\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}} & \frac{\eta}{2}\mathrm{Tr}\left(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{T}\right) + \eta\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{T}\right) \\ & + \lambda_5\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\right\|_1.\end{aligned} \tag{22}$$

The optimal $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ can be updated as follows:

$$\begin{aligned}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{(k)} &= \arg\min_{\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}} \frac{1}{2t_k}\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}} - (\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{(k-1)} - t_k\mathbf{E}[\hat{\mathbf{x}}\hat{\mathbf{y}}])\right\|^2 + \lambda_5\left\|\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\right\|_1 \\ &= \mathcal{S}_{\lambda_5 t_k}(\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{(k-1)} - t_k\mathbf{E}[\hat{\mathbf{x}}\hat{\mathbf{y}}])\end{aligned} \tag{23}$$

where $\mathbf{E}[\hat{\mathbf{x}}\hat{\mathbf{y}}]$ is the gradient of $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ with the nonsmooth $\mathcal{L}_1$-norm term eliminated:

$$\mathbf{E}[\hat{\mathbf{x}}\hat{\mathbf{y}}] = \eta(\hat{\mathbf{X}}^T\hat{\mathbf{Y}} + \hat{\mathbf{X}}^T\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{(k-1)}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}). \tag{24}$$

5) Update $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$: With fixed other variables, the update of $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is obtained by solving the following problem:

$$\begin{aligned}\min_{\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}} & -\frac{N\eta}{2}\log|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}| + \frac{\eta}{2}\mathrm{Tr}\left(\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{T}\right) \\ & + \frac{\eta+1}{2}\mathrm{Tr}\left(\hat{\mathbf{Y}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Y}}^{\mathbf{T}}\right) + \lambda_4\left\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\right\|_1.\end{aligned} \tag{25}$$

Similar to $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ and $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$, the optimal $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is updated as follows:

$$\begin{aligned}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{(k)} &= \arg\min_{\mathbf{Z}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}} \frac{1}{2t_k}\left\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{(k-1)} + t_k\mathbf{E}[\hat{\mathbf{y}}\hat{\mathbf{y}}]\right\|^2 + \lambda_4\left\|\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\right\|_1 \\ &= \mathcal{S}_{\lambda_4 t_k}(\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{(k-1)} - t_k\mathbf{E}[\hat{\mathbf{y}}\hat{\mathbf{y}}]),\end{aligned} \tag{26}$$

where $\mathbf{E}[\hat{\mathbf{y}}\hat{\mathbf{y}}]$ is the gradient of $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ with the nonsmooth $\mathcal{L}_1$-norm term eliminated:

$$\begin{aligned}\mathbf{E}[\hat{\mathbf{y}}\hat{\mathbf{y}}] = & -\frac{\eta}{2}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}(k-1)}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}^{T}\hat{\mathbf{X}}^{T}\hat{\mathbf{X}}\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}(k-1)} \\ & - \frac{\eta N}{2}\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-\mathbf{1}(k-1)} + \frac{\eta+1}{2}\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}.\end{aligned} \tag{27}$$

6) Update $\mathbf{E}$: The update of $\mathbf{E}$ is obtained by solving the following problem:

$$\min_{\mathbf{E}} \frac{1}{2}\left\|\mathbf{X} - \mathbf{X}\mathbf{D} + \frac{\mathbf{\Gamma}_2}{\rho} - \mathbf{E}\right\|_F^2 + \frac{\lambda_2}{\rho}\|\mathbf{E}\|_1. \tag{28}$$

The problem in Eq. (28) can also be solved using the shrinkage operator.

7) Update $\mathbf{J}$: The update of $\mathbf{J}$ is obtained by solving the following problem:

$$\min_{\mathbf{J}} \frac{1}{2}\left\|\mathbf{D} - \mathbf{J} + \frac{\mathbf{\Gamma}_1}{\rho}\right\|_F^2 + \frac{\lambda_1}{\rho}\|\mathbf{J}\|_*. \tag{29}$$

The above problem can be easily solved by the singular value thresholding (SVT) algorithm [49].

Moreover, the multipliers $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ can be directly updated as follows:

$$\begin{cases}\mathbf{\Gamma}_1 = \mathbf{\Gamma}_1 + \rho(\mathbf{D} - \mathbf{J}) \\ \mathbf{\Gamma}_2 = \mathbf{\Gamma}_2 + \rho(\mathbf{X} - \mathbf{X}\mathbf{D} - \mathbf{E}).\end{cases} \tag{30}$$

The overall iteration procedure of the proposed LGGME is summarized in Algorithm 1.

Algorithm 1 Optimization procedure of LGGME

Require: $\mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{E}, \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}, \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \Gamma_1, \Gamma_2, \rho, \eta, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$.

```
while stopping criterion 1 is not satisfied do
    while stopping criterion 2 is not satisfied do
        update J by solving Eq. (29);
        update Θ_{x̂x̂} by solving Eq. (20);
        update D by solving Eq. (16);
        update E by solving Eq. (28);
        update Γ_1, Γ_2 by solving Eq. (30);
    end while
    update Θ_{x̂ŷ} by solving Eq. (22);
    update Θ_{ŷŷ} by solving Eq. (25);
    update Ŷ by solving Eq. (18);
end while
```

Ensure: $\mathbf{D}, \mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}, \mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$

IV. Experiments

A. Emotion6 Dataset

The Emotion6 dataset [50] contains 1,980 affective images related to different natural scenes. Each of the affective images contains valence arousal values and a seven-dimensional sentiment vector, including six Ekman's basic emotion categories [51] and an additional neutral emotion category. These sentiment vectors are collected via a user study and contain multiple nonzero labels with different degrees. Inspired by the superiority of deep features in visual semantic understanding tasks, the well pretrained VGGNet-19 is applied to extract the deep features. We chose the responses from the last fully connected layer and further reduced them into 1,734 dimensions as the final visual representations. In our experiments, 80% of images are randomly selected as the

training set and the remaining images as the test set. We reported the average performance over 10 random splits. For all the parameters, the optimal values are selected via 5-fold cross-validation on the training set and are set to $\eta = 2.5$, $\gamma = 1$, $\lambda_1 = 1e-6$, $\lambda_2 = 0.1$, $\lambda_3 = 0.5$, $\lambda_4 = 1$, and $\lambda_5 = 0.01$ by default.

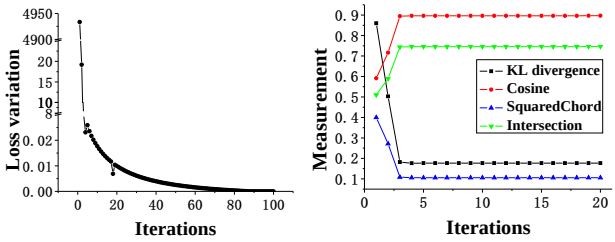
1) Evaluation Measures: To better investigate the prediction performance, we selected six widely used measurements to calculate the distance or similarity between the predicted and real sentiment distributions, including SquaredChord, KL divergence, Intersection, Cosine, Sorensendist, and Chebyshev. TABLE I summarizes the selected measurements and their formulations, where $\mathbf{p} \in \mathbb{R}^m$ and $\mathbf{q} \in \mathbb{R}^m$ denote the real and predicted distributions, respectively. Moreover, "↓" indicates the smaller the value is, the better the performance, while "↑" indicates the larger the value is, the better the performance.

TABLE I: The definitions of different types of measurements.

Distance/Similarity	Family	Formulation
SquaredChord ↓	Squared chord	$\sum_{i=1}^{m} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$
KL divergence↓	Shannons entropy	$\sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i}$
Cosine↑	Inner product	$\frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{i=1}^{m} p_i^2}\sqrt{\sum_{i=1}^{m} q_i^2}}$
Intersection↑	Intersection	$\sum_{i=1}^{m} \min(p_i, q_i)$
Sorensendist↓	$\mathcal{L}_1$-norm	$\frac{\sum_{i=1}^{m} \lvert p_i - q_i \rvert}{\sum_{i=1}^{m} (p_i + q_i)}$
Chebyshev↓	$\mathcal{L}_p$ Minkowski	$\max_i \lvert p_i - q_i \rvert$

2) Convergence analysis: In this section, we investigate the convergence of our objective function with the absolute variations of the objective function and 4 representative measurements. Fig. 2(a) presents absolute variations of the overall loss function with respect to iterations. From the figure, we can observe that the curve has a dramatic drop during the first few iterations and smoothly approaches zero after 80 iterations. To comprehensively evaluate the effectiveness of our proposed method, we further collected the values of 4 measurements during the training process, including KL divergence, SquaredChord, Cosine, and intersection. As shown in Fig. 2(b), the measurements of KL divergence and SquaredChord decrease sharply in the first few iterations and then tend to be stable at a relatively low level. The other two measurements show the opposite trends, as we expected. Both the objective function values and different measurements reveal the good convergence and stability of our proposed method. In our experiments, we used the absolute variation value falling below a threshold of 1e-6 and a maximum of 100 iterations as the stopping criteria.

3) Parameter sensitivity: In this section, we investigated the sensitivity of five parameters, including η, γ, λ_1, λ_4, and λ_5. The parameters η, γ, and λ_1 control the influence of the negative log-likelihood, the KL divergence, and the low-rank representation, respectively. Fig. 3 shows the prediction results with various values of the parameter η on the Emotion6 dataset. From the figure, we can observe that a larger value of η easily generates unsatisfactory prediction results since the correlation patterns between features and between labels



(a) (b)

Fig. 2: Convergence curves of our proposed method in terms of (a) the absolute variations of the objective function and (b) the four measurements.

are weakened by inaccurate $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$. However, a smaller value of η tends to weaken the influence of the Gaussian graphical model, which gives inferences from representations to distributions. Fig. 4 presents the prediction performance

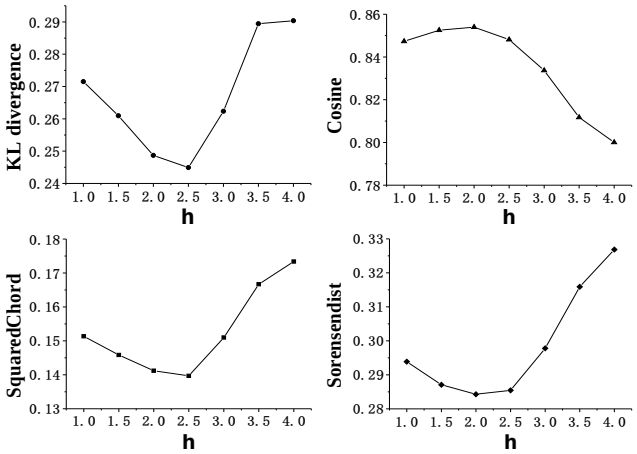


Fig. 3: The prediction performance with various values of the parameter η on the Emotion6 dataset.

with various values of the parameter γ on the Emotion6 dataset. From the figure, we can see that our proposed method achieves poor performance with relatively small values of the parameter γ. With the increasing values of the parameter γ, the prediction performance is significantly improved as the importance of supervised information increases. However, when the value of the parameter γ further increases, the prediction performance begins to deteriorate, indicating the necessity of relaxing the original sentiment distributions. Fig. 5 illustrates the prediction performance with various values of the parameter λ_1 on the Emotion6 dataset. From the figure, we can see that the best prediction performance is achieved when $\lambda_1 = 1e-6$. Although our proposed method shows insensitivity to the parameter λ_1, the prediction performance will not be satisfactory if the low-rank projection matrix $\mathbf{D}$ is replaced with the identify matrix, as shown in TABLE II.

The parameters λ_4 and λ_5 control the sparsity of $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$, respectively. Fig.6 and Fig.7 show the influence of the parameters λ_4 and λ_5 on the Emotion6 dataset. From the two figures, we can observe that the best performance is achieved when $\lambda_4 = 1$ and $\lambda_5 = 0.01$. It is noteworthy that our

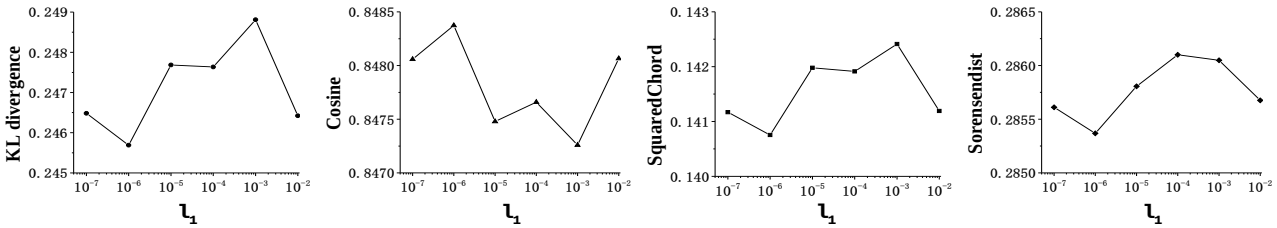


Fig. 5: The prediction performance with various values of the parameter λ_1 on the Emotion6 dataset.

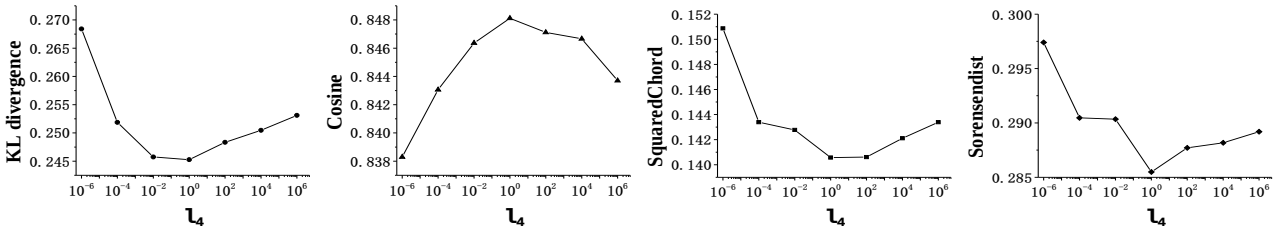


Fig. 6: The prediction performance with various values of the parameter λ_4 on the Emotion6 dataset.

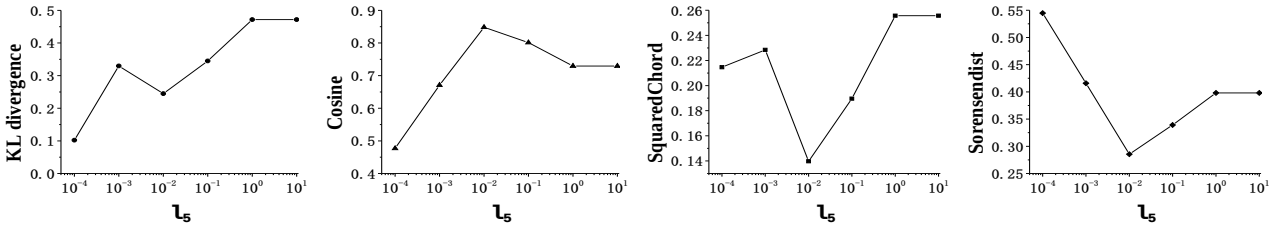


Fig. 7: The prediction performance with various values of the parameter λ_5 on the Emotion6 dataset.

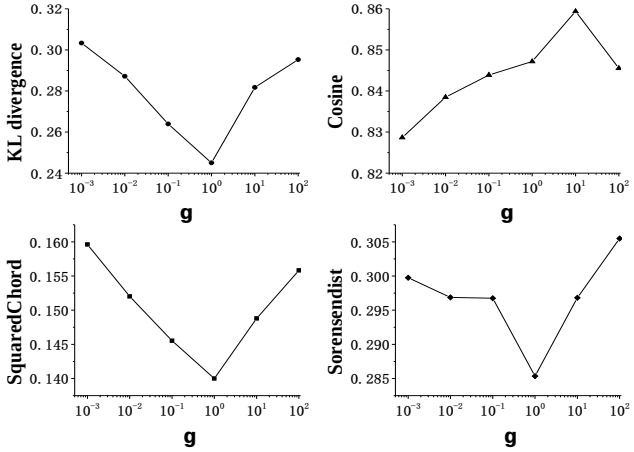


Fig. 4: The prediction performance with various values of the parameter γ on the Emotion6 dataset.

proposed method is more sensitive to the parameter λ_5 than the parameter λ_4. One possible reason is that $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{y}}}$ encodes richer information from features than $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ from sentiments. In addition, this phenomenon reflects that the interactions within labels are less important than the interactions between features and labels, as expected.

4) Ablation analysis: To illustrate the effectiveness of the involved components in our proposed method, we carried out experiments from the following perspectives:

• **NoLC**: We considered the effect of label correlations by replacing $\mathbf{\Theta}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ with the identity matrix.

• **NoLS**: We considered the influence of the sparsity of label correlations by setting $\lambda_4 = 0$.

• **NoFS**: We considered the effect of the sparsity of feature-label correlations by setting $\lambda_5 = 0$.

• **NoLR**: We considered the influence of the low-rank projection matrix by replacing $\mathbf{D}$ with the identity matrix.

• **NoFC**: We considered the influence of feature correlations by replacing $\mathbf{\Theta}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ with the identity matrix.

• **NoKC**: We considered the influence of the KL divergence constraint by setting $\gamma = 0$.

Table II illustrates the performance comparison of the involved components in terms of four measurements. The importance of different components quantized by the average ranking of four measurements is ranked from crucial to subtle as **NoKL** > **NoLC** > **NoLR** > **NoFS** > **NoLS** > **NoFC**. **NoKL** has the greatest impact on the prediction performance, indicating that supervised information is essential to improve the prediction performance of our method. **NoLC** leads to inferior prediction performance, demonstrating that label correlation analysis plays an indispensable role in visual sentiment distribution learning. The prediction results of **NoFS** and **NoLS** illustrate that enforcing structural sparsity on inverse covariance parameters not only reduces the redundancy parameters but also contributes to the selection of significant dependencies. Moreover, we can infer that exploiting low-rank latent representation is superior to the raw features, as **NoLR** achieves less desirable results.

TABLE III: Performance comparison of our proposed method and state-of-the-art methods on the Emotion6 dataset.

Method	SquaredChord↓	KL divergence↓	Intersection↑	Cosine↑	Sorensendist↓	Chebyshev↓	Average Rank
PT-SVM[15]	0.561(13)	0.889(12)	0.477(13)	0.585(13)	0.523(13)	0.356(13)	12.83
AA-KNN[15]	0.356(11)	0.708(11)	0.538(12)	0.602(12)	0.462(12)	0.353(12)	11.67
GFLasso[31]	0.295(9)	0.549(9)	0.565(11)	0.683(11)	0.430(10)	0.323(10)	10.00
MRCE[25]	0.182(2)	0.250(2)	0.661(4)	0.808(2)	0.345(6)	0.238(3)	3.17
LDSVR[52]	0.224(6)	0.415(5)	0.630(9)	0.769(9)	0.370(9)	0.285(9)	7.83
CPNN[34]	0.295(10)	0.564(10)	0.569(10)	0.685(10)	0.431(11)	0.331(11)	10.33
EDL-LRL[32]	0.404(12)	3.699(14)	0.653(7)	0.780(8)	0.327(3)	0.279(8)	8.67
LDL-LCLR[24]	0.217(4)	0.444(8)	0.659(5)	0.790(6)	0.341(5)	0.250(5)	5.50
LDLLC[42]	0.210(3)	0.424(6)	0.664(3)	0.796(5)	0.336(4)	0.247(4)	4.17
LDL-SCL[53]	0.219(5)	0.405(4)	0.637(8)	0.788(7)	0.363(7)	0.268(7)	6.33
ACPNN[16]*	0.701(14)	1.950(13)	0.403(14)	0.475(14)	0.597(14)	0.476(14)	13.83
JCDL[54]*	0.260(8)	0.438(7)	0.668(2)	0.805(3)	0.325(2)	0.251(6)	4.67
SSDL[55]*	0.242(7)	0.400(3)	0.658(6)	0.803(4)	0.369(8)	0.237(2)	5.00
LGGME	**0.140(1)**	**0.244(1)**	**0.700(1)**	**0.848(1)**	**0.286(1)**	**0.211(1)**	**1.00**

TABLE II: Performance comparison with the removal of various components of our proposed method.

	SquaredChord↓	KL divergence↓	Intersection↑	Sorensendist↓
NoLC	0.1540(6)	0.2759(6)	0.6989(3)	0.2976(6)
NoFS	0.1496(4)	0.2666(4)	0.6896(6)	0.2947(4)
NoLS	0.1493(3)	0.2645(3)	0.6904(5)	0.2942(3)
NoLR	0.1532(5)	0.2720(5)	0.6965(4)	0.2950(5)
NoFC	0.1407(2)	0.2457(2)	0.7003(2)	0.2860(2)
NoKL	0.2960(7)	0.5505(7)	0.5637(7)	0.4363(7)
LGGME	0.1397(1)	0.2444(1)	0.7004(1)	0.2856(1)

5) Comparison with state-of-the-arts: In this section, our proposed LGGME method is compared with 10 representative learning methods, including problem transformation on support vector machine (PT-SVM) [15], algorithm adaptation on K-nearest neighbor (AA-KNN) [15], GFLasso [31], MRCE [25], label distribution support vector regression (LDSVR) [52], conditional probability neural network (CPNN) [34], emotion distribution learning by exploiting low-rank label correlations locally (EDL-LRL) [32], label distribution learning with label correlations via low-rank approximation (LDL-LCLR) [24], label distribution learning by exploiting label correlations (LDLLC) [42], and label distribution learning with label correlations on local samples (LDL-SCL) [53]. The comparison methods are briefly described as follows:

- **PT-SVM**: transforms LDL into single-label learning based on a support vector machine (SVM).
- **AA-KNN**: adapts the stable training-free algorithm KNN to the LDL problem.
- **GFLasso**: utilizes a special graph-guided fused penalty to capture correlation patterns.
- **MRCE**: estimates the sparse coefficients and noise inverse covariance matrix in a multivariate regression setting.
- **CPNN**: is specifically designed for LDL by learning a conditional probability neural network.
- **LDSVR**: utilizes multivariate support vector regression followed by a sigmoid function to predict values of label distribution.
- **LDL-SCL**: introduces extra local correlation vectors to encode the influence of clustered instances.
- **EDL-LRL**: utilizes several clustered label groups to capture label correlation structures locally.
- **LDLLC**: applies Person's correlations to capture label relevance in a global manner.
- **LDL-LCLR**: models the global and the local correlations by low-rank approximation and clustering strategies.

We reproduced three algorithms for fair comparisons with the same training and test splits, including augmented conditional probability neural network (ACPNN)[16], joint image emotion classification and distribution learning (JCDL)[54], and structured and sparse annotations for image emotion distribution learning (SSDL)[55]. In our experiments, the average prediction performance of 10 random splits is reported with the same training and test splitting. All comparison methods are fine-tuned with suggested parameters for fair comparisons. Table III shows the performances of our proposed method and other state-of-the-art algorithms on the Emotion6 dataset. From the table, we can make the following observations: 1) Our proposed method achieves the best prediction performance on 6 measurements; 2) In contrast to specialized label distribution learning methods such as CPNN, EDL-LRL, LDLLC, and LDL-LCLR, PT-SVM and AA-KNN obtain poor performance, indicating that the adaptation strategies of traditional learning methods are insufficient for our sentiment distribution learning; 3) Among all specialized LDL methods, CPNN and LDSVR obtain less satisfactory results due to their deficiency in exploiting correlation patterns among labels; 4) GFlasso is significantly inferior to our proposed LGGME, indicating that the intrinsic feature representations and relaxed sentiment distributions are of vital importance for graphical models; 5) LDL-EDL, LDL-LCLR, LDLLC, and LDL-SCL show fairly desirable results. This indicates that exploiting low-rank techniques to capture correlation patterns is beneficial for sentiment distribution learning. 6) MRCE achieves superior performance compared to most of the other methods. Despite

TABLE VI: Running time comparisons of our proposed method and state-of-the-art methods on the Flickr-LDL dataset.

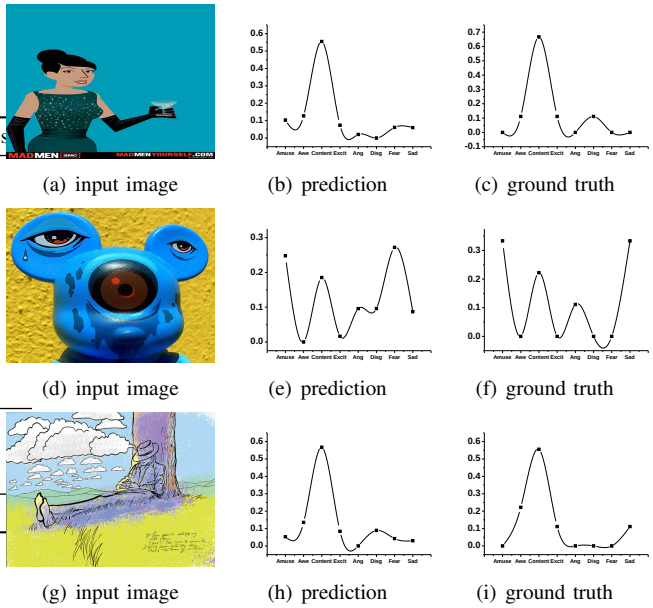
Method	Inner Loop Process (secs)	Overall Training Process (secs)
PTSVM[15]	—	89264.49
MRCE[25]	—	80.05
GFLasso[31]	—	120.08
AA-KNN[15]	—	107.72
CPNN [34]	—	603.12
LDL-LCLR [24]	—	572.37
LDLLC [42]	—	52.12
LDSVR [52]	—	50.96
EDL-LRL [32]	32.22	6505.56
LDL-SCL[53]	1969.54	98476.95
ACPNN[16]*	—	1116.90
JCDL[54]*	142.68	7735.51
SSDL[55]*	214.15	11309.01
LGGME	31.13	428.56

this, our proposed LGGME still outperforms MRCE owing to the ability to characterize the various correlation structures embedded in features and labels. 7) JCDL, which jointly integrates label distribution learning and classification tasks together, achieves desirable performance. Moreover, SSDL also achieves relatively good prediction performance owing to the joint utilization of Earth Mover's distance and KL divergence.

B. Flickr-LDL Dataset

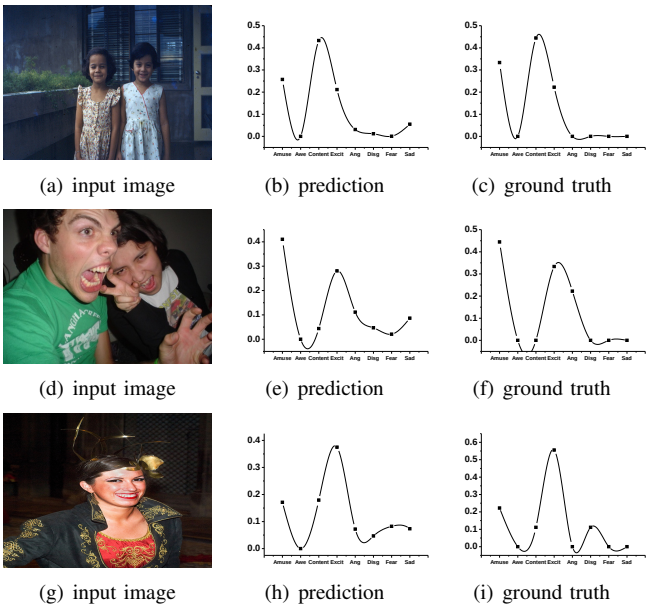
The Flickr-LDL dataset [16] is a large-scale dataset containing 10,700 images labeled with 8 commonly used emotions, including amusement, contentment, excitement, awe, anger, disgust, fear, and sadness. Eighty percent of the images are randomly selected as the training set, and the remaining images are selected as the test set. The average performance over 10 random splits is reported. Similar to the Emotion6 dataset, we extracted features from the last fully connected layer in VGG19-Net and reduced them to 175 dimensions. All optimal parameters are selected by the grid-search strategy and set to $\eta = 50$, $\gamma = 29$, $\lambda_1 = 5e-4$, $\lambda_2 = 0.01$, $\lambda_3 = 0.5$, $\lambda_4 = 0.7$, and $\lambda_5 = 0.01$ by default.

1) Time complexity analysis: To ensure comprehensive analysis of our proposed method, we reported the running time to show the time complexity on the Flickr-LDL dataset, which is a large-scale dataset containing more emotional images than others. TABLE VI shows the average single inner loop time and the overall training time. The simulations of our proposed method and other MATLAB-based methods are carried out in the MATLAB 2017a environment running on a Core 8 Quad, 3.5 GHZ CPU with 16 GB RAM. The CNN-based methods are carried out in a Python 3.7 environment running in an NVIDIA GeForce RTX 3090Ti. The R-based methods are carried out in R 3.5.3. As several experiments are trained as a whole process, it is difficult to record their inner loop time. From the table, we can see that the fast convergence of our proposed method is ensured by the lower inner loop time to some degree. Overall, our proposed method maintains a lower training time, which can provide better support for real-time applications.



(a) input image (b) prediction (c) ground truth
(d) input image (e) prediction (f) ground truth
(g) input image (h) prediction (i) ground truth

Fig. 8: The predicted and ground truth distributions under the cartoon scenario.



(a) input image (b) prediction (c) ground truth
(d) input image (e) prediction (f) ground truth
(g) input image (h) prediction (i) ground truth

Fig. 9: The predicted and ground truth distributions under the human scenario.

2) Subjective analysis: We conducted subjective experiments under different scenarios to show the prediction results of our proposed method. In detail, we first randomly extracted several emotional images related to animals, people, cartoons, indoor scenes and natural scenes from the Flickr-LDL dataset, where each category contains 3 images. We then fed the extracted features of those images into the well-trained model to obtain the predicted distributions. These predicted results were then compared with the ground truth distribution, as shown in Fig. 8-Fig. 11.

From Fig. 9 to Fig. 11, we can see that our proposed

TABLE IV: Performance comparison of our proposed method and state-of-the-art methods on the Flickr-LDL dataset.

Method	SquaredChord↓	KLDiv↓	Intersect↑	Cosine↑	Sorensendist↓	Chebyshev↓	Average Rank
PT-SVM[15]	0.969(14)	1.876(14)	0.307(14)	0.364(14)	0.693(14)	0.532(14)	14.00
AA-KNN[15]	0.447(6)	0.737(7)	0.599(6)	0.777(7)	0.401(6)	0.308(7)	6.50
GFLasso[31]	0.385(5)	0.540(5)	0.634(5)	0.832(5)	0.365(5)	0.276(5)	5.00
MRCE[25]	0.374(4)	0.512(3)	0.641(4)	0.835(4)	0.361(4)	0.273(4)	3.83
LDSVR[52]	0.828(13)	1.184(13)	0.374(13)	0.529(13)	0.626(13)	0.480(13)	13.00
CPNN[34]	0.555(11)	1.001(11)	0.538(10)	0.695(11)	0.462(10)	0.353(10)	10.50
EDL-LRL[32]	0.463(7)	0.864(10)	0.596(7)	0.791(6)	0.402(7)	0.303(6)	7.17
LDL-LCLR[24]	0.503(9)	0.786(9)	0.571(8)	0.767(10)	0.429(8)	0.329(9)	8.83
LDLLC[42]	0.503(8)	0.785(8)	0.570(9)	0.768(9)	0.430(9)	0.329(8)	8.50
LDL-SCL[53]	0.555(10)	0.731(6)	0.529(11)	0.769(8)	0.471(11)	0.357(11)	9.50
ACPNN[16]*	0.614(12)	1.179(12)	0.506(12)	0.650(12)	0.494(12)	0.378(12)	12.00
JCDL[54]*	**0.292(1)**	0.528(4)	**0.676(1)**	0.837(3)	0.338(2)	0.266(2)	2.17
SSDL[55]*	0.356(3)	0.450(2)	0.646(3)	0.849(2)	0.349(3)	0.267(3)	2.67
LGGME	0.299(2)	**0.432(1)**	0.669(2)	**0.856(1)**	**0.317(1)**	**0.243(1)**	**1.33**

TABLE V: Performance comparison of our proposed method and state-of-art methods on the SUB-3DFE dataset.

Method	SquaredChord↓	KLDiv↓	Intersect↑	Cosine↑	Sorensendist↓	Chebyshev↓	Average Rank
PT-SVM[15]	0.042(12)	0.090(12)	0.836(12)	0.913(12)	0.164(12)	0.141(12)	12.00
AA-KNN[15]	0.036(5)	0.074(5)	0.851(4)	0.927(5)	0.149(4)	0.125(4)	4.50
GFLasso[31]	0.040(11)	0.083(11)	0.838(11)	0.919(11)	0.162(11)	0.136(11)	11.00
MRCE[25]	0.036(7)	0.077(7)	0.847(6)	0.925(7)	0.153(6)	0.128(6)	6.50
LDSVR[52]	0.035(3)	0.073(3)	0.852(3)	0.929(3)	0.148(3)	0.126(5)	3.33
CPNN[34]	0.038(8)	0.080(9)	0.841(8)	0.922(9)	0.159(8)	0.135(9)	8.50
EDL-LRL[32]	0.035(4)	0.074(4)	0.850(5)	0.928(4)	0.150(5)	0.125(3)	4.17
LDL-LCLR[24]	0.036(6)	0.076(6)	0.845(7)	0.926(6)	0.155(7)	0.129(7)	6.50
LDLLC[42]	0.039(10)	0.083(10)	0.838(10)	0.920(10)	0.162(10)	0.135(10)	10.00
LDL-SCL[53]	0.030(2)	0.063(2)	0.858(2)	0.938(2)	0.142(2)	0.119(2)	2.00
ACPNN[16]*	0.038(9)	0.079(8)	0.841(9)	0.922(8)	0.159(9)	0.134(8)	8.5
LGGME	**0.029(1)**	**0.060(1)**	**0.862(1)**	**0.942(1)**	**0.138(1)**	**0.112(1)**	**1.00**

(a) input image (b) prediction (c) ground truth

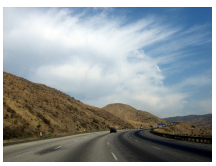
(d) input image (e) prediction (f) ground truth

(g) input image (h) prediction (i) ground truth

Fig. 10: The predicted and ground truth distributions under the indoor scene scenario.

(a) input image (b) prediction (c) ground truth

(d) input image (e) prediction (f) ground truth



(g) input image (h) prediction (i) ground truth

Fig. 11: The predicted and ground truth distributions under the natural scene scenario.

method achieves desirable prediction results in four types of scenarios. Obvious intensity differences exist among weakly correlated labels. However, our proposed method obtains less satisfactory results under the cartoon scenario, which exhibits more abstract emotional patterns. Although our proposed method can ensure globally consistent intensity among significantly related emotions, it is deficient in mining the relative intensity among highly related emotion labels. In future work, we will devote more attention to encoding the global and local correlations among labels to capture the fine-grained emotional differences and enhance the performance of sentiment distribution learning.

3) Comparison with state-of-the-arts: TABLE IV reports the comparison performance of various methods on the Flickr-LDL dataset. From the table, we can observe that our proposed LGGME method achieves the best prediction performance compared to the others, showing its superiority in exploiting complex correlation structures of emotion data. AA-KNN obtains better prediction performance on Flickr-LDL than Emotion6 with various measurements, which is attributed to its advantages in dealing with large-scale datasets. In addition, LDSVR obtains poor prediction results in sentiment distribution learning due to its deficiency in exploring structural information. Both JCDL and SSDL achieve superior performance on the Flickr-LDL dataset in contrast to the Emotion6 dataset. The main reason is that the Flickr-LDL dataset contains more images to support the training process of the two networks.

C. SUB-3DFE Dataset

The SUB-3DFE dataset [15] is a facial emotional expression dataset containing 2,500 images labeled with 6 basic emotions, including happiness, sadness, surprise, fear, anger, and disgust. The SUB-3DFE dataset is extended from a famous 3D facial expression dataset BU-3DFE [56] and scored by 60 participants with a five-level intensity scale for each emotion. Different from the Emotion6 and Flickr-LDL datasets, the SUB-3DFE dataset provides 243-dimensional facial image representations extracted by the local binary patterns (LBP) descriptor. We randomly split the whole dataset into two parts: 80% of images as the training set and the remaining images as the test set. The average performance over 10 random splits is reported. All optimal parameters are selected by the grid-search strategy and set to $\eta = 0.3$, $\gamma = 3$, $\lambda_1 = 1e-14$, $\lambda_2 = 0.1$, $\lambda_3 = 0.5$, $\lambda_4 = 1e-5$, and $\lambda_5 = 5e-5$ by default.

Table V reports the comparison performance of different methods on the SUB-3DFE dataset. From TABLE I, we can see that both LDL-SCL and EDL-LRL exploit clustering strategies, achieving more desirable performance than LDLLC. In fact, the finding that facial images exhibit significant clustering behavior may imply that facial representations are more likely to be associated with prominent sentiments. MRCE achieves worse prediction performance on the SUB-3DFE than on the Flickr-LDL and Emotion6 datasets. One explanation for this discrepancy is that the intrinsic correlation structure is difficult to capture from the original features in this task. Among all the comparison methods, our proposed LGGME produces the most satisfactory performance, illustrating the effectiveness of our proposed method in facial expression distribution learning.

V. Conclusions and Future Work

In this work, we proposed a low-rank latent Gaussian graphical model estimation method to predict visual sentiment distributions of images. In our proposed method, we enforced a multivariate normal distribution on the latent low-rank representations and the estimated sentiment distributions instead of the original observations. On this basis, we modeled different structural sparse correlation patterns between and within features and sentiments and reused them for new instances directly. The experiments conducted on three datasets demonstrated the effectiveness of modeling comprehensive correlation structures and the necessity of our surrogate strategy for the multivariate norm distribution assumption. In the future, we will focus on joint visual representation learning and sentiment semantics embedding in a deep neural network for visual sentiment distribution learning. Furthermore, we will consider seeking representations of sentiment distributions rather than separate sentiments such that the global and local correlations among sentiments can be preserved well.

References

[1] J. Yang, D. She, Y. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018, pp. 491–498.

[2] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 1140–1150.

[3] L. Zhou, X. Fan, Y. Ma, T. Tjahjadi, and Q. Ye, "Uncertainty-aware cross-dataset facial expression recognition via regularized conditional alignment," in *Proceedings of ACM International Conference on Multimedia*, 2020, pp. 2964–2972.

[4] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 406–407.

[5] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proceedings of ACM International Conference on Multimedia*, 2010, pp. 1187–1190.

[6] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1947–1960, 2017.

[7] F. Huang, K. Wei, J. Weng, and Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 3, pp. 1–19, 2020.

[8] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1062–1075, 2018.

[9] M. Jian, J. Dong, M. Gong, H. Yu, L. Nie, Y. Yin, and K.-M. Lam, "Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 970–979, 2019.

[10] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 2661–2668.

[11] D. She, J. Yang, M. Cheng, Y. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2019.

[12] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proceedings of ACM International Conference on Multimedia*, 2019, pp. 192–201.

[13] M. Jian, K.-M. Lam, J. Dong, and L. Shen, "Visual-patch-attention-aware saliency detection," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1575–1586, 2014.

[14] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of ACM International Conference on Multimedia*, 2013, pp. 223–232.

[15] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[16] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2017, pp. 224–230.

[17] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proceedings of ACM International Conference on Multimedia*, 2015, pp. 1247–1250.

[18] T. Ren, X. Jia, W. Li, L. Chen, and Z. Li, "Label distribution learning with label-specific features." in *Proceedings of International Joint Conference on Artificial Intelligence*, 2019, pp. 3318–3324.

[19] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to Emotion*, vol. 1984, pp. 197–219, 1984.

[20] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y. Chen, and S. Chang, "Object-based visual sentiment concept analysis and application," in *Proceedings of ACM International Conference on Multimedia*, 2014, pp. 367–376.

[21] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, "A circular-structured representation for visual emotion distribution learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4237–4246.

[22] Z. Li and J. Tang, "Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2017, pp. 4154–4160.

[23] M. Xu and Z. Zhou, "Incomplete label distribution learning," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2017, pp. 3175–3181.

[24] T. Ren, X. Jia, W. Li, and S. Zhao, "Label distribution learning with label correlations via low-rank approximation," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2019, pp. 3325–3331.

[25] A. Rothman, E. Levina, and J. Zhu, "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 947–962, 2010.

[26] M. I. Jordan *et al.*, "Graphical models," *Statistical Science*, vol. 19, no. 1, pp. 140–155, 2004.

[27] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2007.

[28] G. R. Obozinski, M. J. Wainwright, and M. Jordan, "High-dimensional support union recovery in multivariate regression," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1217–1224, 2008.

[29] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society Series B (Statistical methodology)*, vol. 76, no. 2, p. 373, 2014.

[30] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of International Conference on Machine Learning*, 2010, pp. 543–550.

[31] S. Kim and E. P. Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLOS Genetics*, vol. 5, no. 8, p. e1000587, 2009.

[32] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9841–9849.

[33] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, "Stimuli-aware visual emotion analysis," *IEEE Transactions on Image Processing*, vol. 30, pp. 7432–7445, 2021.

[34] X. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[35] P. Hou, X. Geng, Z.-W. Huo, and J.-Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 2015–2021.

[36] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning." in *IJCAI*, 2018, pp. 712–718.

[37] X. Geng, X. Qian, Z. Huo, and Y. Zhang, "Head pose estimation based on multivariate label distribution," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, DOI: 10.1109/TPAMI.2020.3029585*.

[38] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7093–7102.

[39] M. Ling and X. Geng, "Indoor crowd counting by mixture of gaussians label distribution learning," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5691–5701, 2019.

[40] B. Yang, W. Zhan, N. Wang, X. Liu, and J. Lv, "Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel," *Neurocomputing*, vol. 390, pp. 207–216, 2020.

[41] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021,DOI: 10.1109/TPAMI.2021.3094362.

[42] X. Jia, W. Li, J. Liu, and Y. Zhang, "Label distribution learning by exploiting label correlations," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018, pp. 3310–3317.

[43] J. Yang, X. Gao, L. Li, X. Wang, and J. Ding, "Solver: Scene-object interrelated visual emotion reasoning network," *IEEE Transactions on Image Processing*, vol. 30, pp. 8686–8701, 2021.

[44] Z. Xu and S. Wang, "Emotional attention detection and correlation exploration for image emotion distribution learning," *IEEE Transactions on Affective Computing*, 2021,DOI: 10.1109/TAFFC.2021.3071131.

[45] G. Liu, Z. Lin, Y. Yu *et al.*, "Robust subspace segmentation by low-rank representation." in *Proceedings of International Conference on Machine Learning*, vol. 1, 2010, p. 8.

[46] D. D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high dimensional problems," *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 615–636, 2009.

[47] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[48] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[49] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[50] K. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 860–868.

[51] P. Ekman, "What emotion categories or dimensions can observers judge from facial behavior?" *Emotions in the Human Face*, vol. 96, pp. 39–55, 1982.

[52] X. Geng and P. Hou, "Pre-release prediction of crowd opinion on movies by label distribution learning." in *Proceedings of International Joint Conference on Artificial Intelligence*, 2015, pp. 3511–3517.

[53] X. Jia, Z. Li, X. Zheng, W. Li, and S. Huang, "Label distribution learning with label correlations on local samples," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 8, pp. 1–13, 2019.

[54] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network." in *Proceedings of International Joint Conference on Artificial Intelligenc*, 2017, pp. 3266–3272.

[55] H. Xiong, H. Liu, B. Zhong, and Y. Fu, "Structured and sparse annotations for image emotion distribution learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 363–370.

[56] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 211–216.