Dimensionality Reduction in Multiple Ordinal Regression

Jiabei Zeng, Yang Liu, *Member, IEEE*, Biao Leng, Zhang Xiong, and Yiu-ming Cheung, *Senior Member, IEEE*

***Abstract*— Supervised dimensionality reduction (DR) plays an important role in learning systems with high-dimensional data. It projects the data into a low-dimensional subspace and keeps the projected data distinguishable in different classes. In addition to preserving the discriminant information for binary or multiple classes, some real-world applications also require keeping the preference degrees of assigning the data to multiple aspects, e.g., to keep the different intensities for co-occurring facial expressions or the product ratings in different aspects. To address this issue, we propose a novel supervised DR method for *DR in multiple ordinal regression* (DRMOR), whose projected subspace preserves all the ordinal information in multiple aspects or labels. We formulate this problem as a joint optimization framework to simultaneously perform DR and ordinal regression. In contrast to most existing DR methods, which are conducted independently of the subsequent classification or ordinal regression, the proposed framework fully benefits from both of the procedures. We experimentally demonstrate that the proposed DRMOR method (DRMOR-M) well preserves the ordinal information from all the aspects or labels in the learned subspace. Moreover, DRMOR-M exhibits advantages compared with representative DR or ordinal regression algorithms on three standard data sets.**

***Index Terms*— Dimensionality reduction (DR), multiple labels, ordinal regression, supervised.**

I. Introduction

DIMENSIONALITY reduction (DR) is the procedure of mapping high-dimensional data to a lower-dimensional subspace in which the informative characteristics of the original data are well preserved. Since DR effectively mitigates the curse of dimensionality, this procedure is of great importance and has been widely used in facilitating data compression [59], visualization [51], clustering [37], and classification [32] within vast domains where high-dimensional data are prevalent. For instance, in the field of computer vision, visual descriptors such as scale-invariant feature transform (SIFT), histogram of oriented gradient, and Gabor always have high dimensionality.

An ideal DR method transforms the high-dimensional data into a reduced representation with its intrinsic dimensionality, which is the minimum number of parameters needed to account for the observed properties of the data [18]. A vast number of DR methods have been proposed over the past decades to effectively preserve the necessary properties of the data. According to the availability of label/class information, DR methods can be roughly divided into unsupervised (no label/class information), supervised (complete label/class information), and semi-supervised (partial label/class information) methods. The majority of the unsupervised DR methods search for unrelated or independent factors based on statistics (e.g., principal component analysis (PCA) [25] and factor analysis [22]) or information theory (e.g., independent component analysis [11]). However, these unrelated or independent factors are not guaranteed to preserve the key properties of the label information of data from different classes. To address this issue, supervised DR methods have been proposed to incorporate the label information in choosing the projection of the original data. For example, linear discriminant analysis (LDA) [17] is one of the most representative supervised DR methods, and this method aims to find a low-dimensional subspace to minimize the distance between data points that have the same label and simultaneously differentiate those data points that are unalike.

Although supervised DR preserves the key properties that distinguish the data from different classes, one's interests in the data might be beyond the information for a single two-class or multiclass classification but for classification in multiple aspects and various degrees, namely, multiple ordinal regression. In fact, such interests are ubiquitous in real-world applications. To name a few, in facial expression detection, multiple emotions [13] or facial action units (AUs) [36] always arise simultaneously. Each emotion or AU occurs with different intensities that indicate its subtleness or obviousness. We expect the data in the low-dimensional space to be easily distinguishable with respect to different expressions that have a range of intensities. In product evaluation, a dress is rated from several aspects. For example, a dress could be rated five stars for comfort and only three stars for aesthetics. In this case, the expected reduced representation is the one that faithfully

Manuscript received May 13, 2016; revised February 28, 2017 and June 28, 2017; accepted August 31, 2017. Date of publication October 10, 2017; date of current version August 20, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61472023, Grant 61503317, Grant 61272366, Grant 61672444, and Grant 61702481, in part by the SZSTI Grant under Project JCYJ20160531194006833, and in part by the Faculty Research Grant of Hong Kong Baptist University under Project FRG2/16-17/032, Project FRG2/15-16/049, and Project FRG2/16-17/051. *(Corresponding author: Biao Leng.)*

J. Zeng is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zjb1990@gmail.com; jiabei.zeng@vipl.ict.ac.cn).

Y. Liu is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, and also with the Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen 518057, China (e-mail: csygliu@comp.hkbu.edu.hk).

B. Leng and Z. Xiong are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: lengbiao@buaa.edu.cn; xiongz@buaa.edu.cn).

Y.-M. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, and also with the Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen 518057, China, and also with the United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai 519087, China (e-mail: ymc@comp.hkbu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2752003

reflects the different ratings for a number of different aspects. To learn the low-dimensional subspace for data with different rankings in multiple labels, we propose a new method for *DR in multiple ordinal regression* (DRMOR), the projected subspace of which preserves all the ordinal information in multiple aspects or labels.

Despite the high dimensionality of data in many applications of multiple ordinal regression, DRMOR remains an almost unexplored problem. There are two key issues in DRMOR: 1) how does the learned subspace capture the ordinal information within each individual label and 2) how does the learned subspace preserve all the ordinal information of multiple labels simultaneously? To address these two issues, a novel DRMOR method (DRMOR-M) is proposed. To address the first issue, we force the projected data to be separated by a set of parallel hyperplanes, where the data between two contiguous hyperplanes belong to the same rating or ordinal category. To address the second issue, we ensure that we can find a corresponding set of such hyperplanes for each label. Thus, the proposed framework preserves all the ordinal information of multiple labels. Note that the proposed learning framework optimizes the projected subspace and multiple ordinal regressors simultaneously. Rather than addressing DR and classification/regression separately, the proposed framework combines the two works and can, thus, fully benefit from the merits of the two procedures. The contributions of this paper are summarized as follows.

1) We investigate an important but relatively unexplored problem in learning with high-dimensional data, i.e., DRMOR, which needs to find a low-dimensional subspace that captures the ordinal information in multiple aspects/labels.
2) We propose a novel method named DRMOR-M to address the DRMOR problem by jointly conducting DR and multiple ordinal regression, which takes advantage of both procedures. To solve the problem, we propose a simple but effective heuristic algorithm.
3) We perform extensive experiments on standard data sets and provide a detailed analysis, showing the effectiveness of the proposed method.

II. RELATED WORKS

The proposed DRMOR-M projects data onto a subspace that preserves ordinal information in multiple aspects or labels. This section, therefore, reviews some related works in ordinal regression and multilabel dimensionality reduction.

A. Ordinal Regression

Ranking information can be used to construct models that are more accurate than those constructed from binary yes-or-no information. The necessity of taking ranking information into account was evaluated in the previous works. For example, Hühn and Hüllermeier [27] showed that ordinal metamethods do exploit ordinal information and yield better performance by comparing ordinal metamodels to their nominal counterparts. The results with respect to metamodel approaches can be further improved using specifically designed ordinal regression methods [20].

Most existing works address ordering information in classification problems, namely, ordinal regression. The objective of ordinal regression is to classify patterns using a categorical scale that shows a natural order between the categories. To rank data into discrete ordered scales, a number of ordinal regression methods have been proposed, which can be classified into three typical approaches [20]: *naive approaches* that use other standard machine learning prediction algorithms, e.g., standard regression [23], [43], nominal classification [1], and cost-sensitive classification [31], [52]; *ordinal binary decomposition approaches* that decompose the ordinal problem into several binary problems and then separately solve them using multiple models [6], [41], [57] or using one multiple output model [12], [16]; and *threshold approaches* that approximate a real value predictor and then divide the real line into intervals (see [63]). Among the three approaches, the threshold approaches are the most prevalent approaches for solving ordinal regression, including support vector machine (SVM) formulations [10], [24], Gaussian processes [9], discriminant learning [34], [46], ensemble learning [15], and so forth. To achieve an incremental version of ordinal regression, Gu *et al.* [19] extended the online vSVC algorithm to a modified support vector ordinal regression. Hamsici and Martinez [21] proposed a multiple ordinal regression method by maximizing the sum of the margins between every consecutive class with respect to one or more rankings.

To the best of our knowledge, few works have addressed the ranking information explicitly in dimensionality reduction. Sun *et al.* [46] reformulated the discriminant analysis as a classification technique to tackle ordinal regression, which is known as kernel discriminant learning for ordinal regression (KDLOR). Although KDLOR is not designed to be a DR method, it computes an optimal 1-D mapping for the data. To maintain the projected data in order, KDLOR imposes an ordering constraint over contiguous classes on the averages of projected patterns of each class. In a later work, Sun *et al.* [47] extended KDLOR to learn multiple orthogonal mapping directions and then combined them into a final decision function. Additionally, Liu *et al.* [34], [35] extended KDLOR by preserving the intrinsic geometry of the data in the embedded manifold structure. Pérez-Ortiz *et al.* [40] proposed reformulating the proportional odds model to have a low-dimensional feature space and nonlinear boundaries. They also proposed a method to select the optimal dimensionality. Li *et al.* [30] proposed an ordinal distance metric learning method for image ranking, which preserves both the local geometry and the ordinal relationship of the data.

B. Multilabel Dimensionality Reduction

In traditional supervised dimensionality reduction, each labeled data sample generally belongs to only one class. However, in many real-world applications, such as image classification [42], emotion recognition [33], and text categorization [53], each data sample might be associated with multiple labels. To overcome the curse of dimensionality in such types of multilabel scenarios, multilabel DR techniques

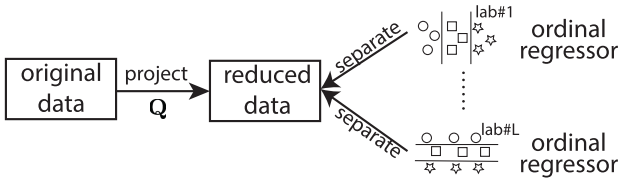


Fig. 1. Framework for joint DR and multiple ordinal regression. The original data are mapped onto a reduced subspace by the projection $\mathbf{Q}$. Meanwhile, the reduced data can be separated by multiple sets of parallel hyperplanes.

have attracted considerable attention in recent years [50]. For instance, Yu *et al.* [60] proposed a method called multilabel informed latent semantic indexing to preserve the information of data while capturing the correlations between multiple labels. Arenas-garcía *et al.* [2] presented the sparse kernel orthonormalized partial least squares approach to handle multilabel data. Zhang and Zhou [62] introduced a multilabel DR algorithm by maximizing the dependence between data and corresponding labels. Park and Lee [39] extended the traditional LDA to a multilabel version by applying the copy transformation. Furthermore, another multilabel LDA was formulated by taking advantage of label correlations [55]. In addition, Yuan *et al.* [61] further extended the above multilabel LDA by incorporating a local consistency term into the objective function. Chang *et al.* [7] proposed a convex formulation for semi-supervised multilabel feature selection. Sun *et al.* [50] proposed a least squares framework to unify several multilabel DR algorithms, including hypergraph spectral learning [48], shared-subspace learning [28], and canonical correlation analysis (CCA) [49], among others.

In addition to the above general multilabel DR algorithms, some methods have also been introduced for more specific applications. For instance, Wang *et al.* [54] presented multilabel sparse coding for image annotation. A multilabel feature transform algorithm was proposed for image classification [56]. Furthermore, another algorithm called block-row sparse multiview multilabel learning was also proposed for the task of image classification [64]. Panagakis *et al.* [38] proposed a sparse multilabel linear embedding nonnegative tensor factorization method for music tagging. Liu *et al.* [33] presented an algorithm called multiemotion similarity preserving embedding for music emotion recognition.

III. Dimensionality Reduction in Multiple Ordinal Regression

In this section, we propose a framework (DRMOR-M) to jointly solve the problems of DR and multiple ordinal regression. Fig. 1 illustrates this joint framework. This framework projects the original data onto a low-dimensional subspace and forces the reduced data to be separated by several sets of parallel hyperplanes, where each set corresponds to an individual label or aspect.

A. Notations

Let N instances $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times D}$ with D-dimensional features, and the label indicator matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top \in \mathbb{R}^{N\times L}$, where $\mathbf{y}_i \in \mathbb{N}^L$ encodes

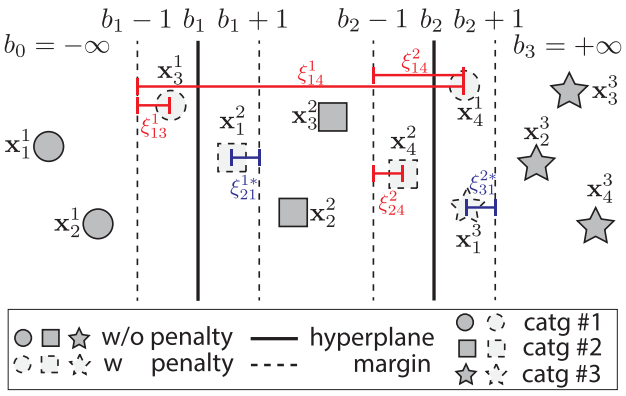


Fig. 2. Illustration of a single label ordinal regression with parallel hyperplanes. The samples with dashed edges and light gray are penalized by errors ξ.

information in L labels for instance $\mathbf{x}_i$. Each label l assigns the samples to r_l ordinal categories $\mathbf{C}_l = \{1, \ldots, r_l\}$. Suppose that there are n_{k_l} instances in category $k_l \in \mathbf{C}_l$; thus, we have $N = \sum_{k_l=1}^{r_l} n_{k_l}$, $\forall l = 1, \ldots, L$. For notational convenience, we also denote a sample $\mathbf{x}_i$ as $\mathbf{x}_{i_l}^{k_l}$, if $\mathbf{x}_i$ is the i_lth sample in category $k_l \in \mathbf{C}_l$ for $i_l \in \{1, \ldots, n_{k_l}\}$.

B. Formulation

The proposed DRMOR-M framework projects the original data onto a low-dimensional subspace, which removes irrelevant and redundant information from the raw features. On the reduced data, DRMOR-M simultaneously solves L ordinal regression problems, each for an individual label, to differentiate the multiple ordinal information.

Specifically, DRMOR-M projects the original data onto a d-dimensional space by a projection matrix $\mathbf{Q} \in \mathcal{R}^{D\times d}$. In the new subspace, DRMOR-M ensures that for each individual label, the projected data can be separated by an ordinal regressor. In our formulation, we use the support vector ordinal regression with implicit constraints (SVORIM) [10] as our basic ordinal regressor. In fact, the proposed learning framework can also easily be built on other maximum-margin-based ordinal regressors, e.g., SVOREX [10]. For label l, it attempts to find $r_l - 1$ parallel hyperplanes to separate the r_l ordered categories. These parallel hyperplanes are described as $h_l^j = \{\mathbf{x}|\mathbf{w}_l^\top \mathbf{Q}^\top \mathbf{x} = b_l^j\}, j = 1, \ldots, r_l - 1$, where $\mathbf{w}_l$ is the normal vector to the hyperplanes, and thresholds $\mathbf{b}_l = [b_l^1, \ldots, b_l^{r-1}]$ determine the locations. By introducing two auxiliary hyperplanes h_l^0 and $h_l^{r_l}$ with $b_l^0 = -\infty$ and $b_l^{r_l} = +\infty$, respectively, samples in category $k_l \in \mathbf{C}_l$ should lie between hyperplanes $h_l^{k_l-1}$ and $h_l^{k_l}$. Fig. 2 illustrates an example of the hyperplanes. In this figure, the solid black lines are two parallel hyperplanes that separate the data into three ordinal categories. The dashed lines denote soft margins. We will penalize the samples within the margins (e.g., $\mathbf{x}_3^1$) and the misclassified ones (e.g., $\mathbf{x}_4^1$). In Fig. 2, these penalized samples have dashed edges and are filled with light gray. We will discuss the details about the penalty later in this section.

To simultaneously seek the projection matrix $\mathbf{Q}$ and L sets of hyperplanes $\{h_l^j\}_{j=1}^{r_l-1}$ $(l = 1, \ldots, L)$, we formulate the

objective function of DRMOR-M as follows:

$$\min_{\{\mathbf{w}_l,\mathbf{b}_l,\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*\}_l^L,\mathbf{Q}} \sum_{l=1}^{L}\left(\frac{1}{2}\mathbf{w}_l^\top\mathbf{w}_l+\kappa_l\text{Cost}\big(\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*\big)\right) \tag{1}$$

$$\text{s.t.: } \mathbf{w}_l^\top\mathbf{Q}^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\le -1+\xi_{l,k_li_l}^{j},\quad \xi_{l,k_li_l}^{j}\ge 0 \tag{2}$$
$$\forall k_l=1,\ldots,j;\quad i_l=1,\ldots,n_{k_l}$$
$$\forall l=1,\ldots,L;\quad j=1,\ldots,r_l-1$$
$$\mathbf{w}_l^\top\mathbf{Q}^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\ge +1-\xi_{l,k_li_l}^{*j},\quad \xi_{l,k_li_l}^{*j}\ge 0 \tag{3}$$
$$\forall k_l=j+1,\ldots,r_l;\quad i_l=1,\ldots,n_{k_l}$$
$$\forall l=1,\ldots,L;\quad j=1,\ldots,r_l-1$$
$$\mathbf{Q}^\top\mathbf{Q}=\mathbf{I} \tag{4}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ are slack variables that represent the errors for misclassifying the samples into higher and lower ordered categories, respectively; $\text{Cost}(\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*)=\sum_{j,k_l,i_l}\xi_{k_li_l}^{j}+\sum_{j,k_l,i_l}\xi_{k_li_l}^{*j}$ denotes the errors on label l; and $\kappa_l>0$ is the balancing parameter.

Fig. 2 briefly illustrates the errors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ under a single label. The red lines denote $\boldsymbol{\xi}$, and the blue lines denote $\boldsymbol{\xi}^*$. The circle, rectangle, and star shapes denote samples from three ordinal categories. Shapes in light gray and with a dashed edge should be penalized according to constraints (13) and (14). With the maximum-margin settings, the constraints in (13) require that for each threshold b_l^j, any sample $\mathbf{x}_{i_l}^{k_l}$ should satisfy $\mathbf{w}_l^\top\mathbf{Q}^\top\mathbf{x}_{i_l}^{k_l}\le b_l^j-1$ if it belongs to a lower or equally ordered category than $j\in\mathbf{C}_l$. For example, in Fig. 2, all the circles $\mathbf{x}_i^1, i=1,\ldots,4$, in category 1 should lie to the left of b_1-1, and all the circles $\mathbf{x}_i^1$ in category 1 and rectangles $\mathbf{x}_i^2$ in category 2 should lie to the left of b_2-1. We penalize $\mathbf{x}_3^1$ and $\mathbf{x}_4^1$ with ξ_{13}^1 and ξ_{14}^1 for not being on the left side of b_1-1. Similarly, we penalize $\mathbf{x}_4^2$ with ξ_{24}^2 because it does not lie to the left of b_2-1. We also penalize x_4^1 with ξ_{14}^2 because it does not lie to the left of b_2-1 as it should. The constraints in (14) require any sample to satisfy $\mathbf{w}_l^\top\mathbf{Q}^\top\mathbf{x}_{i_l}^{k_l}\ge b_l^j+1$, if it has a higher ordered category than $j\in\mathbf{C}_l$. We penalize the samples that violate the rules. For instance, in Fig. 2, $\mathbf{x}_1^2$ is penalized with ξ_{21}^{1*} for not lying to the left of b_1+1 because it belongs to category 2. Similarly, $\mathbf{x}_1^3$ is penalized with ξ_{31}^{2*} for not being on the left side of b_2+1 because it belongs to category 3.

Note that every sample has $\sum_{l=1}^{L}(r_l-1)$ inequality constraints in total (one for each threshold b_l^j). Moreover, (15) restricts an orthonormal transformation $\mathbf{Q}$ to the low-dimensional subspace shared by all the labels.

C. Optimization Procedure

The optimization problem characterized by (1)–(15) is nonconvex with respect to $\mathbf{Q}$ due to the nonconvex constraint $\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}$. In this section, we show that this nonconvex problem can be decomposed and be heuristically solved in two steps: computing the projection $\mathbf{Q}$ and computing the ordinal regressor $\mathbf{w}_l$ for the reduced data.

1) Problem Reformulation: Let us define $\mathbf{u}_l\in\mathbb{R}^D$ as follows:

$$\mathbf{u}_l=\mathbf{Q}\mathbf{w}_l\quad\forall l=1,\ldots,L. \tag{5}$$

By adding (5) as a constraint, the problem in (1)–(15) can be reformulated as follows:

$$\min_{\{\mathbf{u}_l,\mathbf{b}_l,\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*\}_l^L,\mathbf{Q}} \sum_{l=1}^{L}\left(\frac{1}{2}\mathbf{u}_l^\top\mathbf{u}_l+\kappa_l\text{Cost}\big(\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*\big)\right)$$
$$\text{s.t.: } \mathbf{u}_l^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\le -1+\xi_{l,k_li_l}^{j},\quad \xi_{l,k_li_l}^{j}\ge 0$$
$$\forall k_l=1,\ldots,j;\quad i_l=1,\ldots,n_{k_l}$$
$$\forall l=1,\ldots,L;\quad j=1,\ldots,r_l-1$$
$$\mathbf{u}_l^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\ge +1-\xi_{l,k_li_l}^{*j},\quad \xi_{l,k_li_l}^{*j}\ge 0$$
$$\forall k_l=j+1,\ldots,r_l;\quad i_l=1,\ldots,n_{k_l}$$
$$\forall l=1,\ldots,L;\quad j=1,\ldots,r_l-1$$
$$\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}$$
$$\mathbf{u}_l=\mathbf{Q}\mathbf{Q}^\top\mathbf{u}_l\quad\forall l=1,\ldots,L. \tag{6}$$

Before presenting the details of the equivalent transformation from (1)–(15) to (6), we first prove the following theorem.

Theorem 1: For $\mathbf{w}\in\mathbb{R}^d$, $\mathbf{u}\in\mathbb{R}^D$, and $\mathbf{Q}\in\mathbb{R}^{D\times d}$, given that $\mathbf{u}=\mathbf{Q}\mathbf{w}$ and $\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}$, we can have the following:

1) $\mathbf{w}=\mathbf{Q}^\top\mathbf{u}$;
2) $\mathbf{u}=\mathbf{Q}\mathbf{Q}^\top\mathbf{u}$;
3) $\mathbf{w}^\top\mathbf{w}=\mathbf{u}^\top\mathbf{u}$;

Proof: The derivation is straightforward, as follows.

1) $\mathbf{w}=\mathbf{Q}^\top\mathbf{Q}\mathbf{w}=\mathbf{Q}^\top\mathbf{u}$.
2) $\mathbf{u}=\mathbf{Q}\mathbf{w}=\mathbf{Q}(\mathbf{Q}^\top\mathbf{u})=\mathbf{Q}\mathbf{Q}^\top\mathbf{u}$.
3) $\mathbf{w}^\top\mathbf{w}=(\mathbf{Q}^\top\mathbf{u})^\top\mathbf{Q}^\top\mathbf{u}=\mathbf{u}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{u}=\mathbf{u}^\top\mathbf{u}$. □

Since the solution $\mathbf{Q}$ to the problem (1)–(15) yields the constraint $\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}$, we are free to apply Theorem 1 to transform (1)–(15) to (6). Specifically, we first rewrite (1) as the objective in (6) according to Theorem 1(3). Then, we replace $\mathbf{w}_l^\top\mathbf{Q}^\top$ as $\mathbf{u}_l^\top$ in constraints (13) and (14) according to the add-on constraint (5). Finally, we rewrite the add-on constraint (5) as the one in Theorem 1(2).

2) Computation of $\mathbf{Q}$*:* We heuristically solve $\mathbf{Q}$ based on the initial solutions to the ordinal regressions in the original feature space. Specifically, we first initialize $\mathbf{u}_l$ by solving problem (6) without the constraint $\mathbf{u}_l=\mathbf{Q}\mathbf{Q}^\top\mathbf{u}_l$. Then, with the fixed $\mathbf{u}_l$, we solve $\mathbf{Q}$ according to the last two equations in problem (6).

a) Solving $\mathbf{u}_l$*:* Without the constraints with respect to $\mathbf{Q}$, $\mathbf{u}_1,\ldots,\mathbf{u}_L$ are the solutions to L ordinal regression problems in the original D-dimensional feature space. The ordinal regression problem for label l is a standard SVORIM [10]

$$\min_{\mathbf{u}_l,\mathbf{b}_l,\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*}\ \frac{1}{2}\mathbf{u}_l^\top\mathbf{u}_l+\kappa_l\text{Cost}\big(\boldsymbol{\xi}_l,\boldsymbol{\xi}_l^*\big)$$
$$\text{s.t.: } \mathbf{u}_l^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\le -1+\xi_{l,k_li_l}^{j},\quad \xi_{l,k_li_l}^{j}\ge 0$$
$$\forall j=1,\ldots,r_l-1;\quad k_l=1,\ldots,j;\ i_l=1,\ldots,n_{k_l}$$
$$\mathbf{u}_l^\top\mathbf{x}_{i_l}^{k_l}-b_{lj}\ge +1-\xi_{l,k_li_l}^{*j},\quad \xi_{l,k_li_l}^{*j}\ge 0$$
$$\forall j=1,\ldots,r_l-1;\quad k_l=j+1,\ldots,r_l$$
$$i_l=1,\ldots,n_{k_l}. \tag{7}$$

The standard SVORIM in (7) can be solved in its dual form using the SMO algorithm [10].

b) Solving $\mathbf{Q}$*:* With the optimal $\mathbf{u}_l^* \in \mathbb{R}^D$, $\mathbf{Q} \in \mathbb{R}^{D\times d}$ yields the following equations:

$$\begin{cases} \mathbf{Q}^\top \mathbf{Q} = \mathbf{I} \\ \mathbf{u}_l^* = \mathbf{Q}\mathbf{Q}^\top \mathbf{u}_l^*; \quad l = 1, \ldots, L. \end{cases} \tag{8}$$

Equation (8) has $D \times d$ variables (the degrees of freedom of $\mathbf{Q}$) and $(d \times d + L \times D)$ equations. It is not guaranteed to have a solution if the number of variables is less than the number of equations, say, $D\times d < d\times d + L \times D$. Considering such conditions, we require that $\mathbf{Q}\mathbf{Q}^\top$ should keep $\mathbf{u}_l^*$ as close as possible to its transformation $\mathbf{Q}\mathbf{Q}^\top \mathbf{u}_l^*$. Thus, we obtain $\mathbf{Q}$ by optimizing

$$\min_{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}} \sum_{l=1}^{L} \left|\left|\mathbf{Q}\mathbf{Q}^\top \mathbf{u}_l^* - \mathbf{u}_l^*\right|\right|^2. \tag{9}$$

Following some standard operations in linear algebra, Problem (9) can be rewritten as follows:

$$\max_{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}} \operatorname{tr}(\mathbf{Q}^\top \mathbf{U}\mathbf{U}^\top \mathbf{Q}) \tag{10}$$

where $\mathbf{U} = [\mathbf{u}_1^*, \ldots, \mathbf{u}_L^*]$. Problem (10) is a trace maximization problem and has a closed-form solution, in which the optimal $\mathbf{Q}^*$ is the eigenvectors of matrix $\mathbf{U}\mathbf{U}^\top$ corresponding to the d largest eigenvalues. To avoid numerical problems, we can also compute $\mathbf{Q}^*$ as the left singular vectors of the matrix $\mathbf{U}$.

3) Computation of $\mathbf{w}_l$*:* After obtaining the approximate optimal $\mathbf{Q}^*$, we compute $\mathbf{w}_1, \ldots, \mathbf{w}_L$ in (1) with the fixed $\mathbf{Q} = \mathbf{Q}^*$. The optimal $\mathbf{w}_l$ $(l = 1, \ldots, L)$ is the solution to the l-th ordinal regression problem on the projected data $\tilde{\mathbf{x}} = \mathbf{Q}^{*\top}\mathbf{x} \in \mathbb{R}^d$. Specifically, the ordinal regression problem is

$$\begin{aligned} \min_{\mathbf{w}_l, \mathbf{b}_l, \boldsymbol{\xi}_l, \boldsymbol{\xi}_l^*} \quad & \frac{1}{2}\mathbf{w}_l^\top \mathbf{w}_l + \kappa_l \mathrm{Cost}\left(\boldsymbol{\xi}_l, \boldsymbol{\xi}_l^*\right) \\ \text{s.t.: } & \mathbf{w}_l^\top \tilde{\mathbf{x}}_{i_l}^{k_l} - b_{lj} \le -1 + \xi_{l,k_l i_l}^{j}, \quad \xi_{l,k_l i_l}^{j} \ge 0 \\ & \forall j = 1, \ldots, r_l - 1; \quad k_l = 1, \ldots, j; \; i_l = 1, \ldots, n_{k_l} \\ & \mathbf{w}_l^\top \tilde{\mathbf{x}}_{i_l}^{k_l} - b_{lj} \ge +1 - \xi_{l,k_l i_l}^{*j}, \quad \xi_{l,k_l i_l}^{*j} \ge 0 \\ & \forall j = 1, \ldots, r_l - 1; \quad k_l = j + 1, \ldots, r_l \\ & \qquad\qquad i_l = 1, \ldots, n_{k_l}. \end{aligned} \tag{11}$$

The problem in (11) is also a standard SVORIM, which can be solved in its dual form discussed in [10]. Note that we do not compute $\mathbf{w}_l$ as $\mathbf{w}_l = \mathbf{Q}^\top \mathbf{u}_l$ because the second equation in (8) does not always hold.

We summarize the above optimization procedure in Algorithm 1.

D. Nonlinear Extension

Nonlinear DRMOR-M can be obtained by replacing the original feature $\mathbf{x}$ in the linear formulation, (1), with the values of M kernel functions $\boldsymbol{\Phi}(\mathbf{x}) \in \mathbb{R}^M$, e.g., radial basis function kernel, polynomial kernel, and χ square kernel. To simplify the notation, in the kernel space, the corresponding representation of $\mathbf{x}$ is denoted as $\mathbf{k} = [k_1, \ldots, k_M]$, $k_i = \langle \mathbf{x}, \boldsymbol{c}_i \rangle_{\mathcal{K}}$, where c_i is the ith selected point, and $\langle . \rangle_{\mathcal{K}}$ denotes the inner product in kernel space. Then, the D-dimensional data $\mathbf{x}$ are mapped to a d-dimensional subspace by function $f(\mathbf{x}) = \mathbf{Q}^\top \mathbf{k}$, where $\mathbf{Q} \in \mathbb{R}^{M\times d}$ is the projection matrix. The nonlinear DRMOR-M is then formulated as follows:

Algorithm 1 Method of DRMOR (DRMOR-M)

Input: Low dimension d, data $\mathbf{X} \in \mathbb{R}^{N\times D}$ and their labels $\{\mathbf{y}_1 \in \{1, \ldots, r_L\}^N, \ldots, \mathbf{y}_L \in \{1, \ldots, r_L\}^N\}$

Output: Projection matrix $\mathbf{Q} \in \mathbb{R}^{D\times d}$, classifiers parameter $\mathbf{w}_l$, and thresholds $\mathbf{b}_l$, $l = 1, \ldots, L$

1: **for all** $l = 1, \ldots, L$ **do**
2: $(\mathbf{u}_l, \mathbf{b}_l) \leftarrow$ solve problem (7) with known $(\mathbf{X}, \mathbf{y}_l)$;
3: **end for**
4: $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_L]$;
5: $\mathbf{Q} \leftarrow$ left singular vectors of $\mathbf{U}$ with respect to the d largest singular values;
6: $\tilde{\mathbf{X}} \leftarrow \mathbf{X}\mathbf{Q}$;
7: **for all** $l = 1, \ldots, L$ **do**
8: $(\mathbf{w}_l, \mathbf{b}_l) \leftarrow$ solve problem (11) with known $(\tilde{\mathbf{X}}, \mathbf{y}_l)$;
9: **end for**

$$\min_{\{\boldsymbol{\alpha}_l, \mathbf{b}_l, \boldsymbol{\xi}_l, \boldsymbol{\xi}_l^*\}_l^L, \mathbf{Q}} \sum_{l=1}^{L} \left(\frac{1}{2}\boldsymbol{\alpha}_l^\top \boldsymbol{\alpha}_l + \kappa_l \mathrm{Cost}\left(\boldsymbol{\xi}_l, \boldsymbol{\xi}_l^*\right)\right) \tag{12}$$

$$\text{s.t.: } \boldsymbol{\alpha}_l^\top \mathbf{Q}^\top \mathbf{k}_{i_l}^{k_l} - b_{lj} \le -1 + \xi_{l,k_l i_l}^{j}, \quad \xi_{l,k_l i_l}^{j} \ge 0 \tag{13}$$

$$\forall k_l = 1, \ldots, j; \quad i_l = 1, \ldots, n_{k_l}$$

$$\forall l = 1, \ldots, L; \quad j = 1, \ldots, r_l - 1$$

$$\boldsymbol{\alpha}_l^\top \mathbf{Q}^\top \mathbf{k}_{i_l}^{k_l} - b_{lj} \ge +1 - \xi_{l,k_l i_l}^{*j}, \quad \xi_{l,k_l i_l}^{*j} \ge 0 \tag{14}$$

$$\forall k_l = j + 1, \ldots, r_l; \quad i_l = 1, \ldots, n_{k_l}$$

$$\forall l = 1, \ldots, L; \quad j = 1, \ldots, r_l - 1$$

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}. \tag{15}$$

E. Discussion

In fact, several representative algorithms can be considered as special cases of the proposed DRMOR-M. Specifically, if we set $L = 1$, DRMOR-M becomes a single-label problem that jointly performs DR and ordinal regression. It further degenerates to SVORIM [10], if we remove the scheme of projecting the original data to a reduced space. When $r_1 = \ldots = r_L$, DRMOR-M provides an alternative solution to the graded multilabel classification proposed in [8]. DRMOR-M treats the binary classification problem as a special case of ordinal regression with $r = 2$. In particular, if $r_l = 2, \forall l$, DRMOR-M reduces to DR in multilabel classification, which is equivalent to the one discussed in [29].

IV. EXPERIMENTS

In this section, we experimentally validate the effectiveness of the proposed DRMOR-M. First, we utilize a synthetic example to demonstrate that DRMOR-M can preserve the ordinal information on several aspects. Second, we show that DRMOR-M outperforms representative DR algorithms and competitive ordinal regression methods in terms of estimation errors. Third, we investigate how the new subspace is influenced by the dimensions and label numbers.

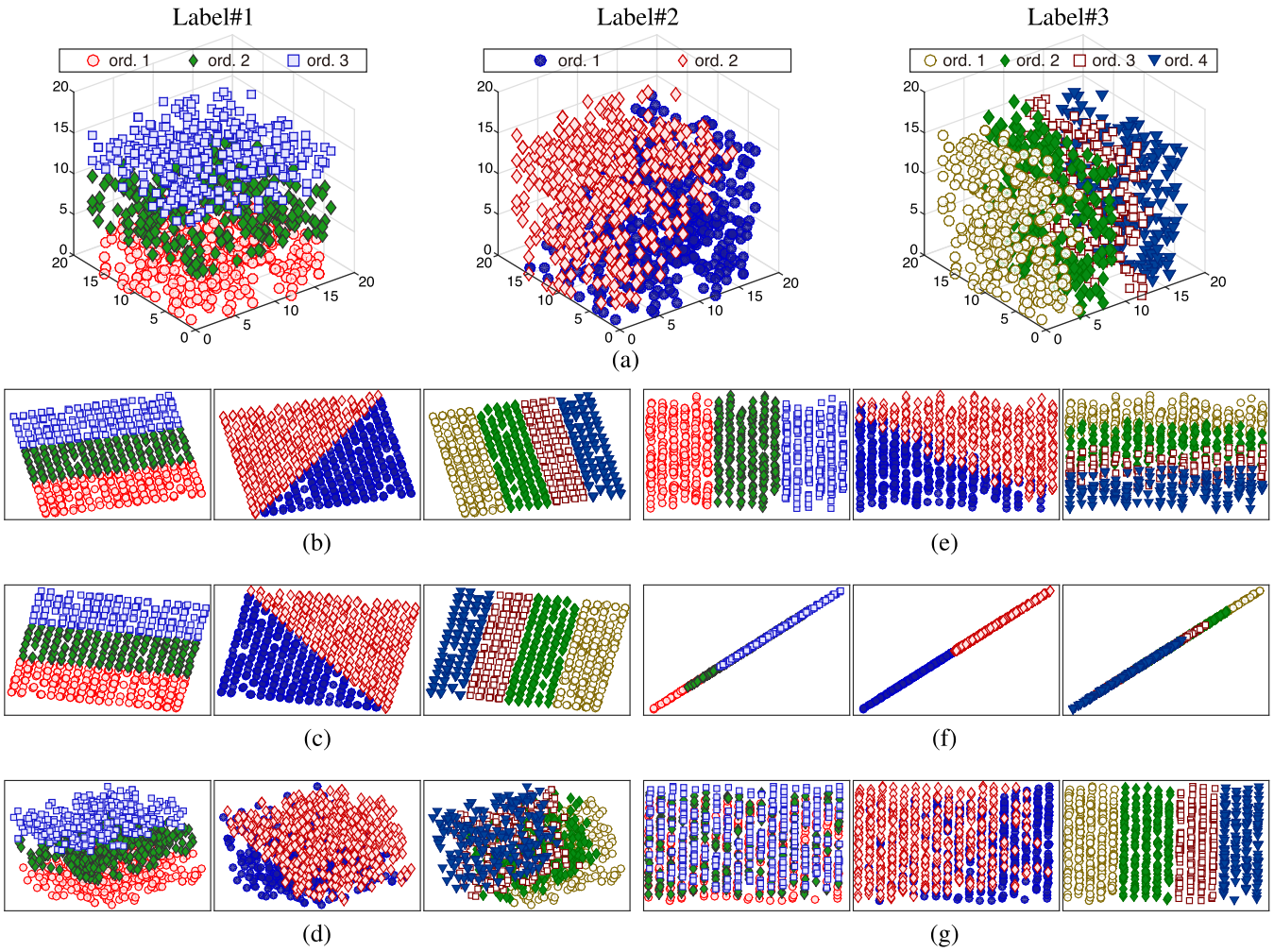


Fig. 3. (a) Synthetic data and (b) 2-D projections by DRMOR-M. (c) CCA. (d) PCA. (e)–(g) LDA supervised by different labels. Each column shows the annotation with one label. DRMOR-M preserves all the ordinal information. Best viewed in color.

A. Schematic Illustration

To show that the DRMOR-learned subspace can simultaneously preserve multiple ordinal information, we illustrate the projections of synthetic examples by DRMOR-M, as well as other typical supervised and unsupervised DR methods.

We uniformly generated 1000 data points in an $18 \times 18 \times 18$ cube and annotated them with three labels. Under all the labels, the points are linearly separable in the original feature space. Fig. 3(a) presents the 3-D synthetic data with annotations under the three labels. As shown in this figure, the points are horizontally divided into three categories under Label#1, obliquely divided into two categories under Label#2, and vertically divided into four categories under Label#3.

Fig. 3(b)–(g) present the 2-D projections of the synthetic data by DRMOR-M, CCA [49], PCA [25], and LDA [17] under the supervision of three labels, respectively. As shown in Fig. 3(b) and (c), the projections on both DRMOR-learned and CCA-learned subspaces are perfectly separated into the ordinal categories under all three labels. This result occurs because the proposed DRMOR-M considers all the ordinal information on different labels, which makes the learned subspace capable of reflecting the correct rankings of the data on each label. CCA maximizes the correlations between the data and labels; thus, the data are projected onto a lower-dimensional space directed by the label information. In Fig. 3(d), the 2-D projections of PCA failed to represent the data well in all the three labels, because the principal components are not guaranteed to distinguish the data well in different categories. Fig. 3(e)–(g) indicate that LDA is competent in separating categories in one label but leaving the categories in other labels inseparable. As shown in Fig. 3(e), the categories in Label#1 are perfectly separated only when LDA is supervised by Label#1. Similar observations are obtained in Fig. 3(f) and (g) for Label#2 and Label#3, respectively.

Second, we show how the nonlinear DR methods perform on the synthetic data that cannot be linearly separated in the original feature space. Fig. 4 (the first row) illustrates the 3-D synthetic data with three types of annotations. As shown, we uniformly generated 1000 data points in a 4-radius sphere and divided them into different categories in three ways. In the first way (Label#1), we divided the points into concentric spheres as four orders from inside to outside. In the second way (Label#2), we constructed two categories: a cylinder

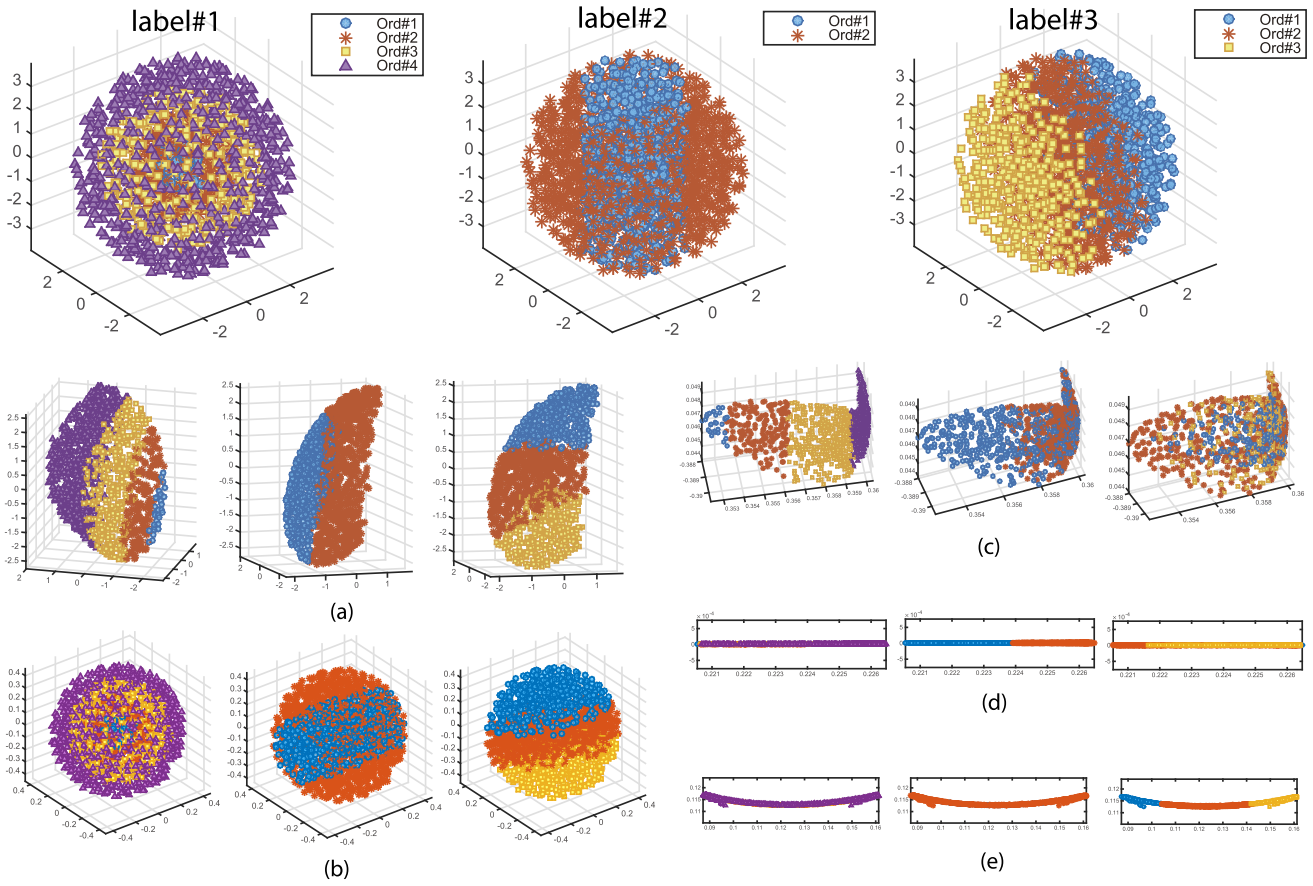


Fig. 4. Synthetic data (first row) and projections by (a) nonlinear DRMOR-M, (b) KPCA, and (c)–(e) KDA supervised by different labels. In (a)–(e), each subfigure shows the annotation with one label. Nonlinear DRMOR-M linearizes all the ordinal information. Best viewed in color.

surrounded by the other category. The third way (Label#3) consists of three horizontal categories.

We plot the projections of the synthetic data by nonlinear DRMOR-M, kernel PCA (KPCA) [44], and kernel discriminant analysis (KDA) [3] under the supervision of the three labels in Fig. 4(a)–(e), respectively. In nonlinear DRMOR-M, the radial basis function kernel was employed: $\Phi(\mathbf{x}) = \exp((1/2)||(\mathbf{x} - \mathbf{c}_m)/\sigma||^2)$, where σ is a predefined parameter and c_m is one of the $M = 10$ selected points. We determined the centers $\mathbf{c}_m$ using k-means on $\mathbf{X}$. Both KPCA and KDA adopted Gaussian kernels and shared the same parameters σ with nonlinear DRMOR-M. As shown in Fig. 4(a), the projection by nonlinear DRMOR-M is linearly separable under all the three labels, despite the original data being not linearly separable on Label #1 or Label #2. As shown in Fig. 4(b), the KPCA-projected data cannot be linearly separated into the categories under Label #1 or Label #2. As shown in Fig. 4(c)–(e), KDA is capable of separating the data only with the label by which KDA is supervised. However, KDA fails to simultaneously separate the data linearly on all the labels.

B. Comparisons With Other Methods

Recall that the proposed DRMOR-M simultaneously conducts DR and ordinal regression. This section demonstrates how DRMOR-M performs in terms of classification errors through comparisons with other DR and ordinal regression methods. In the remainder of this section, we present the details on the data sets, metrics, methods for comparison, and results in our experiments.

TABLE I

OVERVIEW OF THE DATA SETS IN THE EXPERIMENTS. N: NUMBER OF INSTANCES. D: DIMENSION OF FEATURES. L: NUMBER OF LABELS

dataset	N	D	L	ordinal information
BeLa-E	1918	46	4	Gender: {female, male} Age: $\{[21,25],[25,30],[30,35],[35,40]\}$ Prop#1, #2: $\{1 \prec 2 \prec 3 \prec 4 \prec 5\}$
DISFA	130815	6272	8	all: $\{0 \prec \mathsf{A} \prec \mathsf{B} \prec \mathsf{C} \prec \mathsf{D} \prec \mathsf{E}\}$
OES10	403	298	16	1 label: $\{1 \prec 2 \prec 3 \prec 4\}$ 10 labels: $\{1 \prec 2 \prec 3 \prec 4 \prec 5\}$ 5 labels: $\{1 \prec 2 \prec 3 \prec 4 \prec 5 \prec 6\}$

1) Data Sets: We conducted comparisons on three data sets. The overview of the data sets is presented in Table I. Each of the data sets is annotated with multiple labels. Under each label, an instance is assigned to several rating scales or a set of discrete and ordinal categories.

The *BeLa-E* data set [8] records 1930 graduate students' attitudes toward their future jobs. Each record contains 50 attributes. The first 2 attributes are the gender and the age of the student. The remaining 48 attributes grade the importance

TABLE II

COMPARISONS OF DRMOR-M, REPRESENTATIVE DR METHODS (PCA, KPCA, LDA, KDA, AND CCA), AND COMPETITIVE ORDINAL REGRESSION METHODS (SVORIM AND KDLOR WITH LINEAR OR RBF KERNELS) ON THE BELA-E DATA SET. THE DOT "·" MEANS THAT ITS DIFFERENCE FROM THE PROPOSED DRMOR-M IS SIGNIFICANT BASED ON THE PAIRED t-TEST WITH A SIGNIFICANCE LEVEL OF 0.05

	Gender	Age	Prop#1	Prop#2	*Avg.*	Gender	Age	Prop#1	Prop#2	*Avg.*
method	MZE (×100)					MAE (×100)				
DRMOR-M	30.23±1.19	**32.94**±0.71	52.71±1.06	**31.85**±0.91	**36.93**±0.97	30.23±1.19	**33.52**±0.72	61.39±1.25	**34.42**±1.04	**39.89**±1.05
PCA	·48.88±7.90	·34.07±7.91	·89.93±2.44	·36.94±0.60	·52.46±4.71	·48.88±7.90	·35.07±10.79	·169.50±14.95	·45.34±0.84	·74.70±8.62
KPCA	·34.80±0.85	·**32.94**±0.71	·54.23±1.10	·36.84±0.74	·39.70±0.85	· 34.80±0.85	·**33.52**±0.72	62.82±1.27	·44.85±1.60	·44.00±1.11
LDA	30.84±1.07	·36.84±0.94	·57.11±1.07	33.11±1.10	·39.47±1.04	30.84±1.07	·39.03±1.14	·72.24±1.72	37.22±1.40	·44.83±1.33
KDA	33.22±1.10	**32.94**±0.71	·58.76±0.96	36.44±1.29	·40.34±1.02	33.22±1.10	**33.52**±0.72	·69.59±1.21	38.93±1.15	43.82±1.05
CCA	·30.83±1.10	·33.22±0.71	·53.47±1.13	·32.12±0.92	·37.41±0.96	· 30.83±1.10	·33.85±0.73	·62.76±1.33	·34.84±1.03	·40.57±1.05
SVORIM	·30.45±1.19	·**32.94**±0.71	·**52.70**±1.06	·32.39±0.91	·37.12±0.97	· 30.45±1.19	·**33.52**±0.72	·**61.19**±1.25	·35.40±1.04	·40.14±1.05
SVORIM(n)	·**30.21**±1.14	·**32.94**±0.71	·54.08±1.08	·32.09±1.01	·37.33±0.99	· **30.21**±1.14	·**33.52**±0.72	·63.05±1.20	·35.14±1.19	·40.48±1.06
KDLOR(l)	·31.01±1.04	·72.14±1.66	·64.20±1.59	·38.04±1.75	·51.35±1.51	· 31.01±1.04	·98.60±3.69	·84.91±2.68	·46.30±2.96	·65.20±2.59
KDLOR(n)	· 30.85±0.98	·**32.94**±0.71	·58.10±1.24	·35.92±1.37	·39.45±1.07	· 30.85±0.98	·**33.52**±0.72	·71.11±1.55	·38.88±1.47	·43.59±1.18

of different properties of their future job, evaluated by the students on an ordinal scale with five levels. We used 1918 of these 1930 records (excluding 12 records with invalid ages). We selected the first 4 attributes (Gender, Age, Prop#1, and Prop#2) as labels and the remaining 46 properties as features. Gender consists of two classes. DRMOR-M treats the binary classification problem as a special case of ordinal regression with $r = 2$. Age was divided into four time-ordered intervals. Prop#1 and Prop#2 were two of the graded properties with five levels.

The Denver Intensity of Spontaneous Facial Action (*DISFA*) data set [36] records 27 subjects' spontaneous expressions when they are watching video clips. This data set consists of 27 videos with 4845 frames each. Each frame is encoded with 12 facial AUs [14], and each AU is annotated by 0 (absent) or intensities A–E. Our experiments selected eight AUs as labels, and the features were the concatenated SIFT descriptors around 49 landmarks detected by active appearance model [58].

The *OES10* data set [45] was collected from the annual Occupation Employment Survey compiled by the U.S. Bureau of Labor Statistics for the year 2010. Each instance provides the estimated number of full-time equivalent employees across many employment types for a specific metropolitan area. The 298-d features were a randomly sequenced subset of employment types, and the 16 labels were randomly selected from the entire set of categories above the 50% threshold. OES10 was used to validate multitarget regression problems [4], [45]. To adapt OES10 to our experiments, we divided each of the target variables into four to six ordinal categories according to their values.

2) Metrics: We utilized two standard evaluation criteria, i.e., mean zero-one error (MZE) and mean absolute error (MAE), in our experiments. For all the methods, we also report the average results obtained over all the labels.

MZE is the error rate of the single-label classifier, i.e., $\text{MZE} = (1/N)\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i \neq y_i)$, where y_i is the ground truth, $\hat{y}_i$ is the predicted one, and the indicator $\mathbf{1}(\hat{y}_i \neq y_i) = 1$ if the prediction conflicts with the ground truth; otherwise, $\mathbf{1}(\hat{y}_i \neq y_i) = 0$. MZE reflects the global performance without considering the order.

MAE measures the average deviation of the prediction from the true order, i.e., $\text{MZE} = (1/N)\sum_{i=1}^{N} |\hat{y}_i - y_i|$, where the ordinal scales are treated as consecutive integers.

In our experiments, each reported value was averaged over several repeated runs, i.e., 50 runs in Bela-E, 5 runs in DISFA, and 10 runs in OES10. During each run of BeLa-E, 30% of the instances (approximately 570 samples) were randomly selected as training samples, and the remainder were used for testing. Similarly, in DISFA, 2% (approximately 2700 samples) of the instances were selected for training, and the remainder were used for testing; in OES10, approximately 33% (130–140 samples) of the instances were for training, and the remaining instances were used for testing. For fairness, we fixed these training/test splits for each comparison on all the data sets in our experiments.

3) Compared Methods: First, we compared the proposed DRMOR-M with the typical unsupervised or supervised dimensionality reduction as follows.

PCA/KPCA: PCA [25] is a prevalent unsupervised DR method. It finds a new orthogonal coordinate system that optimally describes variance for the original data. In our experiments, we adopted PCA to reduce the data dimension by preserving 98% of the energy. KPCA [44] extends PCA using techniques of kernel methods.

LDA/KDA: LDA [5], [17] is a supervised DR method that finds a linear combination of features to separate two or more classes of data points. KDA [3] extends LDA using kernel methods. In our experiments, the LDA or KDA was learned and evaluated on each individual label.

CCA: CCA [26], [49] finds correlations between two sets of multidimensional variables. It can be applied as a multilabel DR tool, in which the two sets of variables are derived from the data and the class labels, respectively. By maximizing the correlation between the data and the associated labels, the data are projected onto a lower-dimensional space directed by the label information.

The projected data generated by the above DR methods were fed into an SVORIM [10]. In KPCA and KDA, a Gaussian kernel was used.

TABLE III

COMPARISONS OF DRMOR-M, REPRESENTATIVE DR METHODS (PCA, LDA, AND CCA), AND COMPETITIVE ORDINAL REGRESSION METHODS (SVORIM AND KDLOR) ON THE DISFA DATA SET

	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	*Avg.*
Method	MZE ($\times 100$)								
DRMOR-M	**4.31**±0.13	**3.15**±0.11	**10.68**±0.09	**9.30**±0.20	**3.80**±0.17	**13.04**±0.10	**12.99**±0.21	**11.31**±0.41	**8.58**±0.18
SVORIM	5.02±0.05	3.43±0.06	11.71±0.11	10.75±0.17	3.96±0.06	15.55±0.15	15.16±0.29	13.63±0.29	9.90±0.15
KDLOR	6.62±0.01	5.56±0.01	42.30±29.53	14.83±0.01	5.38±0.01	33.57±18.56	92.52±0.00	19.07±0.95	27.48±6.13
PCA	6.49±0.28	4.71±0.97	18.71±0.02	14.83±0.01	23.25±37.50	19.79±1.69	24.93±4.08	19.13±10.40	16.48±6.87
LDA	6.62±0.01	5.56±0.01	18.72±0.00	14.83±0.01	5.38±0.01	23.49±0.01	35.09±0.01	19.02±0.01	16.09±0.01
CCA	5.46±0.16	3.75±0.22	11.93±0.28	9.22±0.14	3.84±0.20	14.84±0.29	13.36±0.40	13.59±0.33	9.50±0.25
Method	MAE ($\times 100$)								
DRMOR-M	**5.78**±0.20	**4.40**±0.21	**13.06**±0.29	**10.23**±0.21	**5.32**±0.06	**13.80**±0.11	**13.52**±0.26	**12.46**±0.44	**9.81**±0.22
SVORIM	8.09±0.16	5.49±0.10	15.63±0.42	12.35±0.10	6.57±0.14	16.81±0.16	16.77±0.44	15.28±0.32	12.12±0.23
KDLOR	17.01±0.03	14.58±0.03	63.52±24.74	26.61±0.02	12.42±0.01	44.77±17.85	115.29±0.01	26.72±2.47	40.12±5.65
PCA	15.63±2.49	10.39±3.89	47.34±2.49	26.58±0.04	27.13±36.56	26.48±4.33	35.59±11.19	22.35±12.55	26.43±9.19
LDA	17.01±0.03	14.58±0.03	48.60±0.02	26.61±0.02	12.42±0.01	46.11±0.02	85.47±0.03	31.39±0.02	35.27±0.02
CCA	6.39±0.15	4.39±0.24	14.08±0.42	9.88±0.11	4.54±0.18	15.87±0.40	14.00±0.48	14.65±0.31	10.48±0.29

TABLE IV

COMPARISONS OF DRMOR-M, REPRESENTATIVE DR METHODS (PCA, KPCA, LDA, KDA, AND CCA), AND COMPETITIVE ORDINAL REGRESSION METHODS (SVORIM AND KDLOR WITH LINEAR OR RBF KERNELS) ON THE OES10 DATA SET. THE DOT "·" MEANS THAT ITS DIFFERENCE FROM THE PROPOSED DRMOR-M IS SIGNIFICANT BASED ON THE PAIRED t-TEST WITH A SIGNIFICANCE LEVEL OF 0.05

	DRMOR-M	PCA	KPCA	LDA	KDA	CCA	SVORIM	SVORIM(n)	KDLOR(l)	KDLOR(n)
	MZE ($\times 100$)									
1	56.7±3.0	·67.2±1.8	·69.6±3.8	·65.7±2.1	·67.8±2.3	·68.0±2.1	60.8±5.7	·59.8±2.8	69.0±2.2	·**53.9**±4.4
2	48.1±7.0	·62.6±5.7	·64.8±4.2	·54.9±2.3	·60.4±3.3	·60.4±3.2	·56.2±6.2	·53.2±8.3	·59.6±4.7	·**36.1**±2.0
3	50.1±3.2	·69.0±2.5	·72.2±1.6	·56.9±2.8	·63.8±3.9	·63.2±3.8	·63.8±3.9	·57.7±7.8	62.5±2.4	·**46.8**±3.9
4	63.8±2.2	·73.8±1.7	·75.8±2.2	**61.0**±3.7	·69.2±2.6	·70.7±3.0	·69.3±3.4	·63.6±1.5	72.9±4.0	·63.1±1.9
5	39.1±2.1	·49.2±2.9	·53.6±5.3	·**34.6**±2.8	·58.7±2.1	·57.6±2.0	·48.3±8.9	·67.0±8.4	·58.0±3.5	·43.2±2.1
6	68.2±1.8	·71.0±2.0	·71.3±1.2	**64.8**±3.2	·71.5±2.6	·74.1±2.9	68.3±1.3	67.5±2.3	·74.8±3.8	·76.2±5.0
7	43.4±2.3	·67.8±5.0	·57.4±5.7	·51.5±3.7	·55.1±1.5	·54.5±1.6	·51.1±4.7	·**42.1**±3.8	55.8±3.7	·43.0±3.6
8	51.9±1.8	·67.2±3.2	·63.9±5.3	·62.0±2.1	·64.1±2.4	·63.7±2.2	52.3±2.7	52.6±2.4	63.7±3.8	·**50.8**±2.2
9	58.3±3.1	·77.5±4.8	·81.3±1.0	·**53.9**±3.3	·73.6±3.3	·73.0±2.9	·77.2±2.8	·65.9±6.7	73.7±1.6	·61.9±3.0
10	**51.6**±1.8	·67.0±3.6	·67.6±5.9	·58.6±3.4	·63.1±3.4	·62.4±3.4	·57.8±6.2	51.6±2.2	·64.3±3.6	·54.7±1.9
11	**54.5**±2.2	·63.5±6.5	·62.9±3.6	·63.7±3.4	·65.8±3.1	·65.4±3.5	·58.9±3.6	·57.9±3.2	·65.9±3.2	·65.0±1.9
12	32.3±2.7	·52.5±10.3	·43.6±4.5	·46.6±2.3	·45.3±3.8	·46.0±3.5	·40.3±4.4	·**29.4**±2.2	44.0±3.0	·34.3±4.5
13	65.6±2.8	·72.4±2.4	·74.8±2.3	**62.8**±2.5	·70.3±2.3	·70.5±2.3	·70.6±3.8	·68.2±5.3	·72.5±2.7	·70.7±2.0
14	27.4±2.1	·48.3±10.0	·42.4±4.1	·40.7±2.6	·41.0±2.5	·41.4±2.5	·43.3±7.2	·**26.6**±1.1	40.8±2.2	·27.7±2.3
15	**53.7**±1.4	·73.6±3.7	·71.9±5.3	·64.0±4.1	·70.5±2.9	·70.6±2.3	·59.3±4.0	·61.1±2.1	·69.5±3.8	·63.0±2.4
16	**46.7**±2.3	·64.6±6.5	·59.3±5.6	·59.0±2.4	·61.9±3.4	·61.1±3.0	·52.0±2.2	·53.1±2.8	·61.1±2.9	·56.0±4.1
avg.	**50.7**±11.3	·65.4±8.7	·64.5±11.1	·56.3±9.0	·62.6±9.1	·62.7±9.3	·58.1±10.2	·54.8±12.7	·63.0±9.9	·52.9±13.7
	MAE ($\times 100$)									
1	**75.75**±3.14	·89.96±2.75	·92.49±6.04	·110.84±4.97	·116.47±7.46	·116.55±6.60	79.97±8.05	81.79±4.18	111.47±3.92	·81.54±8.58
2	59.28±4.68	·74.73±9.54	·78.78±5.71	·83.02±5.54	·90.63±6.52	·90.71±6.48	·65.57±5.56	·63.22±5.53	·86.60±8.33	·**51.09**±3.20
3	57.90±1.90	·72.28±4.59	·78.28±2.07	·89.10±5.87	·91.95±8.60	·91.84±8.32	·65.40±3.18	·62.41±3.98	84.92±5.50	·**57.59**±3.73
4	**107.17**±4.32	·127.52±8.59	·131.66±4.58	·137.52±8.92	·145.61±7.25	·143.62±7.63	115.51±8.47	·113.00±3.02	·148.35±12.52	·126.72±5.17
5	**46.46**±2.52	·71.25±4.25	·59.05±4.36	·56.11±3.95	·75.77±3.68	·74.86±3.75	51.62±7.44	·68.60±8.42	·74.09±4.98	·56.84±2.88
6	120.16±4.50	124.68±7.55	124.27±3.97	·150.98±12.45	·160.25±11.73	·160.90±13.14	·**116.80**±3.27	·123.44±5.87	154.69±8.46	·128.68±6.00
7	51.96±2.78	·84.85±11.30	·69.16±6.60	·75.74±4.39	·76.89±4.59	·76.48±4.43	·60.27±5.42	·**50.38**±3.77	·74.45±5.95	·55.70±5.62
8	**65.36**±2.94	·90.11±5.75	·83.00±8.12	·100.19±4.25	·99.07±5.56	·98.65±5.14	65.50±4.53	66.45±3.68	·94.31±6.97	·75.09±4.02
9	**89.18**±4.56	·104.29±2.98	·107.77±1.57	·106.61±4.95	·132.06±6.53	·131.34±6.83	100.30±2.97	·90.06±6.15	·129.89±5.19	·107.27±4.87
10	**65.61**±2.50	·87.57±7.25	·87.16±8.21	·94.58±7.08	·96.80±6.53	·96.09±6.54	·74.65±8.43	·68.66±2.70	·96.53±7.33	·83.01±4.23
11	**71.07**±3.03	·93.09±14.47	·89.30±7.41	·105.59±5.55	·109.69±9.42	·108.41±9.96	·80.38±4.43	·76.69±3.98	·102.13±5.46	·102.59±4.23
12	34.02±2.97	·61.33±19.15	·49.04±6.14	·55.23±4.01	·54.83±5.13	·55.85±5.14	·42.53±4.94	·**31.84**±2.07	51.41±3.74	·38.48±6.73
13	**93.72**±3.39	·104.17±3.41	·107.50±4.41	·120.34±8.06	·126.53±6.50	·126.54±6.22	99.11±4.54	·99.78±4.74	·128.28±9.82	·115.68±7.06
14	27.83±1.98	·54.55±17.08	·45.35±4.22	·48.52±3.91	·48.96±5.30	·49.57±5.44	·44.13±8.08	·**26.89**±1.29	45.73±3.14	·28.07±2.60
15	**71.90**±2.16	·104.63±9.86	·94.08±6.68	·111.56±8.72	·120.61±5.74	·119.38±4.04	·76.51±4.57	·84.58±3.84	·110.98±9.01	·96.51±7.07
16	**53.46**±2.24	·88.32±13.57	·72.83±11.12	·82.93±5.21	·89.32±8.16	·87.88±7.02	·57.94±3.82	·60.69±3.00	·85.90±8.14	·64.37±3.46
avg.	**68.18**±24.95	·89.58±20.41	·85.61±24.39	·95.55±28.67	·102.22±30.77	·101.79±30.51	·74.76±23.09	·73.03±26.01	·98.73±31.13	·79.33±31.10

We also compared the proposed method with state-of-the-art ordinal regression algorithms using original features. Both linear and nonlinear versions (with a Gaussian kernel) of SVORIM and KDLOR were used.

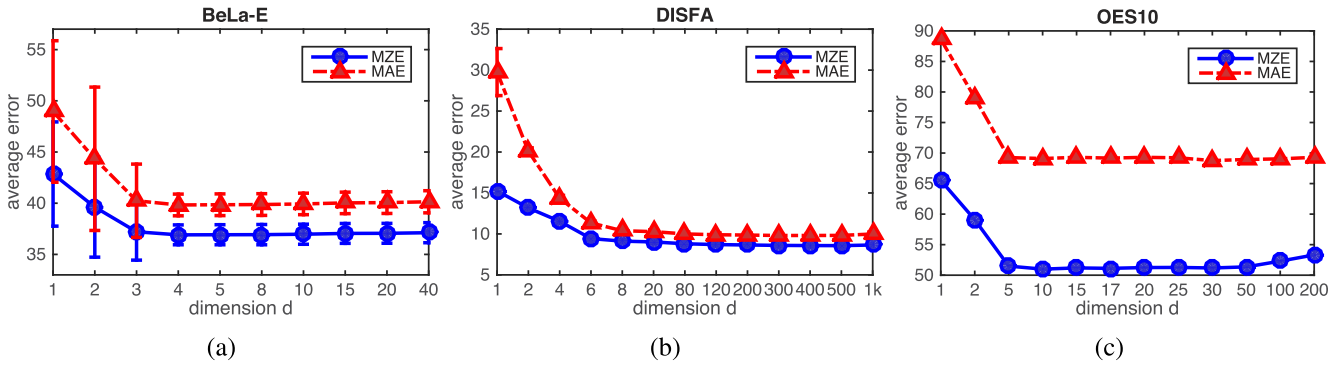


Fig. 5. Average errors of DRMOR-M under various dimensions d on (a) BeLa-E, (b) DISFA, and (c) OES10 data sets.

SVORIM: SVORIM [10] is an SVM-based algorithm that can classify instances into ordinal categories by optimizing some thresholds.

KDLOR: KDLOR [46] reformulates discriminant learning to tackle ordinal regression. KDLOR introduces an ordering constraint on the averages of projected patterns of each ordinal category.

4) Parameter Settings: In DRMOR-M, we set the low dimension d as 5, 400, and 10 for the Bela-E, DISFA, and OES10 data sets, respectively. To reduce the burden of parameter tuning, κ_l was set to be the same for all the ordinal regression problems in the DRMOR-M, SVORIM, and KDLOR methods. The optimal κ_l was selected from the range of $[0.01, 0.03, 0.06, 0.1, 0.3, 0.6, 1, 3, 6, 10]$ via a fivefold cross-validation. For all the nonlinear methods, we used Gaussian kernel $k(\mathbf{x}, \mathbf{z}) = \exp\left((||\mathbf{x} - \mathbf{z}||^2)/(2\delta^2)\right)$ with parameter δ^2 as the averaged square root of the pair-wise Euclidean distance over the training set. In KDLOR, hyperparameter C is searched in the range of $[0.01, 0.03, 0.06, 0.08, 0.1, 0.3, 0.6, 0.8, 1, 3, 6]$, and k is in the range of $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10]$.

5) Results: Tables II–IV report the MZE and MAE on the BeLa-E, DISFA, and OES10 data sets, respectively. The nonlinear versions of the methods were not used in DISFA (Table III) because of their computational limitations on large data sets. The values of MZE and MAE in these three tables are in terms of percentage. From the tables, we summarize the following observations.

First, the proposed DRMOR-M consistently outperforms the other methods in terms of average errors over all the labels, although it is not the best on every individual label. This result is reasonable because DRMOR-M aims to minimize the overall error, as shown in its objective (1). It balances the trade-off among the performances of different labels by extracting a shared low-dimensional subspace. Within the shared subspace, as will be discussed in Section IV-C2, some but not all the labels benefit from it.

Second, DRMOR-M is good at capturing the ranking information compared with other supervised DR methods. On the one hand, as shown in Table IV, DRMOR-M outperforms the other DR methods in all the individual labels in terms of MAE, although it obtains larger MZEs than LDA in several labels. This result indicates that, even though some predictions of DRMOR-M are not exactly correct, DRMOR-M has less prediction discrepancies from the true order than LDA does. The reason is that DRMOR-M requires the samples to be ranked by orders, whereas LDA merely requires them to be separated in different categories. On the other hand, DRMOR-M outperforms CCA on all the data sets. Compared with CCA, DRMOR-M achieved 1.28% and 1.68% improvements in terms of average MZE and MAE on the BeLa-E data set, 9.79% and 6.30% on the DISFA data set, and 19.14% and 33.02% on the OES10 data set. One plausible reason for this result is that CCA does not take into consideration the ranking information within each label, although it explores the relationship between the data and multiple labels.

Third, as shown in Tables II–IV, DRMOR-M outperforms the unsupervised DR methods on all the individual labels in terms of both MZE and MAE. In the BeLa-E data set, DRMOR-M achieves 29.58% and 46.60% reductions in MZE and MAE compared with PCA. These values in the DISFA and OES10 data sets are 48%, 62.85%, 22.48%, and 23.89%, respectively. PCA has relatively high errors because it incorporates no labels; thus, it cannot be guaranteed to separate the data well in different categories or those in ordered categories with different labels. Although KPCA is also unsupervised, KPCA achieves better performance than PCA because KPCA is nonlinear. As shown in Table II, KPCA has the same MZE and MAE as DRMOR-M with label Age. The average reductions in MZE and MAE by DRMOR-M compared with KPCA are 7.5% and 9.34% on the BeLa-E data set and 21.4% and 20.36% on the OES10 data set.

Finally, DRMOR-M is competitive in conducting independent ordinal regression methods on individual labels. On the one hand, DRMOR-M slightly exceeds SVORIM with respect to the average MZE and MAE, even though SVORIM has the smallest errors on a few labels, e.g., Prop#1 in BeLa-E and Labels 7, 13, and 15 in OES10. DRMOR-M achieves 0.51% and 3.11% improvements in terms of the average MZE and MAE on the BeLa-E data set, 13.43% and 18.98% on DISFA, and 12.74% and 8.80% on the OES10 data set. A plausible reason for the slight error reduction is that the ordinal regressions on some labels benefit from the shared subspace, which implicitly incorporates the correlations among different labels. On the other hand, DRMOR-M significantly outperforms KDLOR, particularly the linear version KDLOR(l).

Compared to KDLOR(l), DRMOR-M achieved 28.08% and 38.83% improvements in terms of MZE and MAE on BeLa-E, 68.81% and 75.52% on DISFA, and 19.52% and 30.94% on OES10. The poor performance of KDLOR(l) is attributed to the loss of crucial information because KDLOR(l) reduces the original data to a single dimension without sufficiently capturing the rankings by a linear kernel. When replacing the linear kernel with a Gaussian kernel, KDLOR(n) performs much better than KDLOR(l). Compared to KDLOR(n), DRMOR-M achieved 6.39% and 8.49% improvements in terms of MZE and MAE on BeLa-E and 4.16% and 14.06% on OES10.

C. Analysis

To fully understand the proposed DRMOR-M method, we investigated how DRMOR-M is influenced by the number of reduced dimensions and the number of labels.

1) DRMOR-M With Various Reduced Dimensions: We investigated the performance of DRMOR-M under different reduced dimensions on the Bela-E, DISFA, and OES10 data sets. The original features were 46-D, 6272-D, and 298-D in Bela-E, DISFA, and OES10, respectively.

Fig. 5(a)–(c) plot the average errors over all the labels under various d for the BeLa-E, DISFA, and OES10 data sets, respectively. Both MZE and MAE significantly decrease when d varies from 1 to a turning point, after which the error curves reach a plateau and remain fairly stable. The value of V at the turning point indicates that the largest V eigenvectors might convey most of the useful information in the original feature space. As shown, V is not greater than the number of labels L, e.g., in the BeLa-E and DISFA data sets, $V = L$; in the OES10 data set, $V(=5) < L(=16)$. The reason for this result can be traced back to the computation of the projection matrix $\mathbf{Q}$ in (10). $\mathbf{Q}$ consists of the eigenvectors of $\mathbf{U}\mathbf{U}^\top$, and $\mathbf{U}$ is with rank $V \leq L$. Therefore, the largest V eigenvectors possess the most of the discriminant information. Large errors and variances would be obtained if we reduce the features to an extremely low-dimensional subspace, e.g., $d = 1$, where some useful information was discarded.

Rather than the law of "the more the better," we found that the smallest errors were achieved with a relatively low dimension by examining the exact values in Fig. 5(a)–(c), i.e., the optimal d is approximately five in Bela-E, approximately 400 in DISFA, and 10 in OES10. Note that in Fig. 5(c), the (MZE, MAE) values at 8, 20, 80, 120, 200, 300, 400, 500, and 1 k are (9.12,10.39), (9.02,10.28), (8.80,10.00), (8.70,9.90), (8.65,9.86), (8.60,9.81), (8.58,9.81), (8.57,9.82), and (8.65,9.99), respectively. A possible reason for this result is that extra dimensions may not provide useful information but introduce redundancy or even noise that might cause worse performance.

To further investigate the effects of various d on each label, we plot Fig. 6 to show the errors of each label on the three data sets. As shown, all the labels share the same characteristics: when the reduced dimension d is small, the errors significantly decrease if we increase the dimensions and when d becomes relatively large, the error curves change smoothly. Although

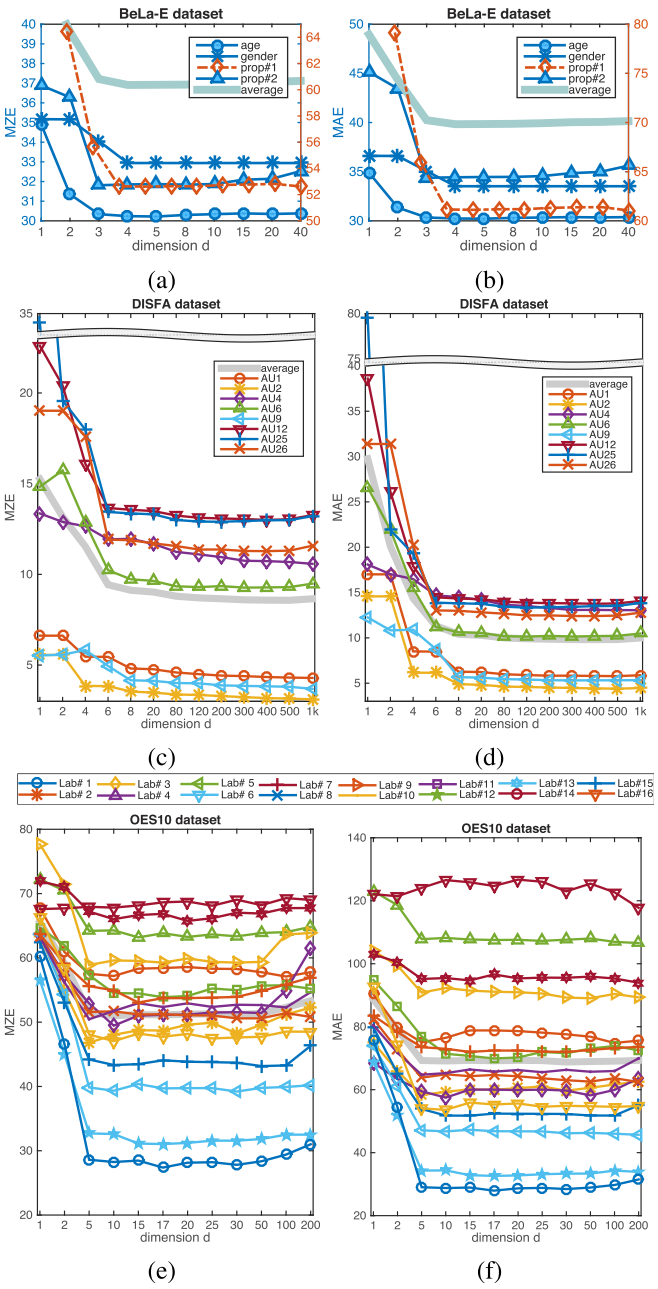


Fig. 6. Errors of DRMOR-M with each label under various dimension d. (a) MZE and (b) MAE of each label on BeLa-E, (c) MZE and (d) MAE of each label on DISFA, and (e) MZE and (f) MAE of each label on OES10.

the changes are not obvious, we can observe different tendencies in different labels. For some of them, the errors continue decreasing with larger d, such as AU1, AU2, AU4, and AU9 in DISFA. For others, the errors accumulate if we continue increasing d, such as Prop#2 in BeLa-E-diff and AU6 in DISFA. This result indicates that, in some labels, the label-related information is inherently complicated and is well represented by high-dimensional features. In others, the label-related information is in the low dimensions and is sensitive to extra noises that are introduced by increasing the dimensionality.

2) DRMOR-M With Various Number of Labels: We designed experiments on the BeLa-E data set to show how L (the number of labels) influences the prediction errors. In our

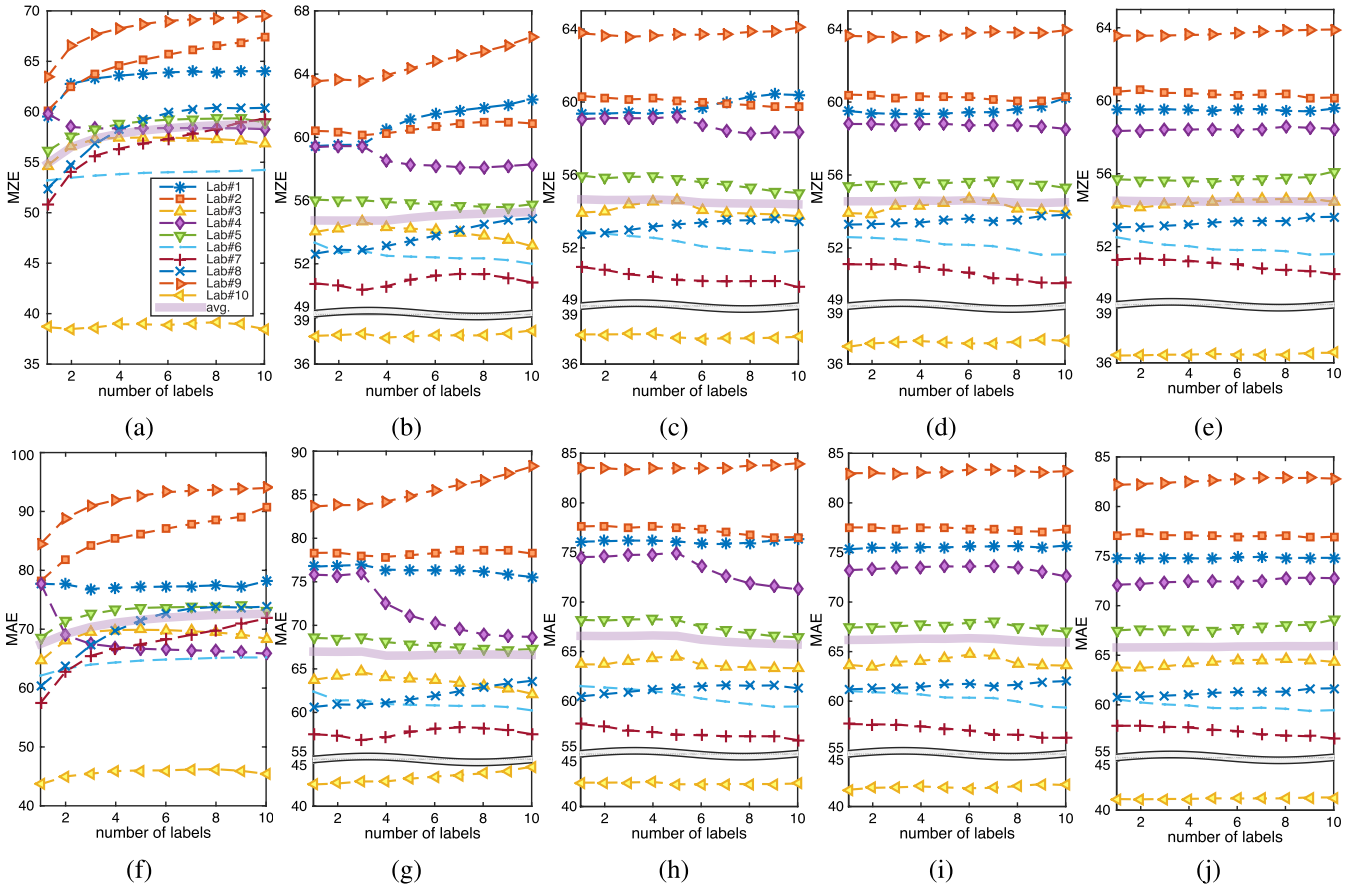


Fig. 7. (a)–(e) MZE and (f)–(j) MAE of DRMOR-M on the BeLa-E data set with various L (number of labels). Each of the five columns is with $d = 1, 3, 5, 7, 10$, respectively.

experiments, we selected 10 five-level properties as candidate labels and the remaining 40 attributes as features. For each L, we selected L labels from the candidate ones and performed experiments with all the C_{10}^{L} combinations of L from the 10 candidate labels. We also prepared 10 splits of the data, in which 50 instances are in the training set and the others are in the test set. We repeated the experiments on 10 splits to obtain the statistics.

Fig. 7 illustrates how the errors change over L (i.e., the number of labels) on each label. Each column corresponds to a fixed reduced dimension d, which equals 1, 3, 5, 7, and 10 from left to right. Each point in the figure denotes the average error of a label if L labels are involved. For example, in Fig. 7(a), a yellow triangle pointing to the left (Lab#10) with $L = 5$ means that Lab#10 achieves a MZE of 0.4, if there are five labels in total involved. The value is averaged over $126(=C_9^4)$ out of the $252(=C_{10}^5)$ combinations where Lab#10 is involved, with 10 repeated trials on different splits in each combination.

As shown in Fig. 7, the performance of DRMOR-M becomes more diverged if we take more labels into consideration. For example, the error curves of Lab#9 buildup but those of Lab#4 shrink when L increases. The average errors over all the labels slightly decrease as L increases when $d = 5, 7, 10$, but the average curves are rapidly or slightly increasing when $d = 1$ and 3. This result suggests that when d is small, considering too many labels provide a little benefit or even decreases the performance.

Furthermore, DRMOR-M is somewhat sensitive to L when the reduced dimension is small. In Fig. 7(a) and (f), where $d = 1$, the curves change in a larger scale than those in Fig. 7(b) and (g), where $d = 3$. In the last two columns of Fig. 7, where $d = 7$ and 10, most of the curves are nearly flat, which indicates that the performance is stable.

V. CONCLUSION

In this paper, we have investigated DRMOR, an important but relatively unexplored issue involving finding an intrinsic subspace that preserves ordered categories in multiple labels. We have formulated DRMOR as a joint optimization problem, which simultaneously takes DR and multiple ordinal regressions into consideration. Specifically, we have proposed DRMOR-M to solve the optimization problem. The experimental results on synthetic examples show that the learned subspace preserves the ordinal information on all the labels. Moreover, empirical comparisons on three standard data sets demonstrate the superiority of DRMOR-M over both representative DR algorithms and competitive ordinal regression methods. We have also investigated how DRMOR-M is influenced by the reduced dimension and the number of involved labels. The investigations show that an optimal reduced dimension can be found for the overall performance

on all the labels, but both the reduced dimension and the number of labels have different influences on individual labels.

For the future work, we are particularly interested in exploring the relations among multiple labels in DRMOR-M as it plays an important role in learning the shared subspace. Moreover, in multilabel scenarios, some labels might be more important than others. How to automatically adjust the weights of different labels in the DRMOR-M framework is another issue worth to further study.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their constructive comments and suggestions.

REFERENCES

[1] A. Agresti, *Analysis of Ordinal Categorical Data*, vol. 656. Hoboken, NJ, USA: Wiley, 2010.

[2] J. Arenas-garcía, K. B. Petersen, and L. K. Hansen, "Sparse kernel orthonormalized PLS for feature extraction in large data sets," in *Proc. NIPS*, 2007, pp. 33–40.

[3] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[4] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Data Mining Knowl. Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[5] D. Cai, X. He, and J. Han, "SRDA: An efficient algorithm for large-scale discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.

[6] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, Jun. 2011, pp. 585–592.

[7] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. 28th AAAI*, 2014, pp. 1171–1177.

[8] W. Cheng and E. Hüllermeier, and K. J. Dembczynski, "Graded multilabel classification: The ordinal case," in *Proc. ICML*, 2010, pp. 223–230.

[9] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1019–1041, Jul. 2005.

[10] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, no. 3, pp. 792–815, 2007.

[11] P. Comon, "Independent component analysis, A new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[12] J. F. P. da Costa, H. Alonso, and J. S. Cardoso, "The unimodal model for the classification of ordinal data," *Neural Netw.*, vol. 21, no. 1, pp. 78–91, 2008.

[13] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[14] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford, U.K.: Oxford Univ. Press, 1997.

[15] F. Fernandez-Navarro, P. A. Gutiérrez, C. Hervas-Martinez, and X. Yao, "Negative correlation ensemble learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 11, pp. 1836–1849, Nov. 2013.

[16] F. Fernández-Navarro, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2075–2085, Nov. 2014.

[17] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Francisco, CA, USA: Academic, 2013.

[19] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.

[20] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.

[21] O. C. Hamsici and A. M. Martinez, "Multiple ordinal regression by maximizing the sum of margins," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2072–2083, Oct. 2016.

[22] H. H. Harman, *Modern Factor Analysis*, 3rd ed. Chicago, IL, USA: Univ. Chicago Press, 1976.

[23] E. F. Harrington *et al.*, "Online ranking/collaborative filtering using the perceptron algorithm," in *Proc. ICML*, vol. 20. 2003, pp. 250–257.

[24] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Proc. NIPS*, 1999, pp. 115–132.

[25] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[26] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[27] J. C. Hühn and E. Hullermeier, "Is an ordinal class structure useful in classifier learning," *Int. J. Data Mining, Model. Manag.*, vol. 1, no. 1, pp. 45–67, 2008.

[28] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 2, pp. 8:1–8:29, 2010.

[29] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proc. IJCAI*, vol. 9. 2009, pp. 1077–1082.

[30] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.

[31] H.-T. Lin and L. Li, "Reduction from cost-sensitive ordinal ranking to weighted binary classification," *Neural Comput.*, vol. 24, no. 5, pp. 1329–1367, 2012.

[32] Y. Liu, Y. Liu, K. C. C. Chan, and K. A. Hua, "Hybrid manifold embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2295–2302, Dec. 2014.

[33] Y. Liu, Y. Liu, Y. Zhao, and K. A. Hua, "What strikes the strings of your heart?—Feature mining for music emotion analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 247–260, Jul. 2015.

[34] Y. Liu, Y. Liu, and K. C. C. Chan, "Ordinal regression via manifold learning," in *Proc. AAAI*, Aug. 2011, p. 1.

[35] Y. Liu, Y. Liu, S. Zhong, and K. C. C. Chan, "Semi-supervised manifold ordinal regression for image ranking," in *Proc. ACM Multimedia*, 2011, pp. 1393–1396.

[36] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[37] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1796–1808, Nov. 2011.

[38] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Sparse multi-label linear embedding within nonnegative tensor factorization applied to music tagging," in *Proc. 11th ISMIR*, 2010, pp. 393–398.

[39] C. H. Park and M. Lee, "On applying linear discriminant analysis for multi-labeled problems," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 878–887, 2008.

[40] M. Pérez-Ortiz, P. A. Gutiérrez, M. Cruz-Ramírez, J. Sánchez-Monedero, and C. Hervás-Martínez, "Kernelising the Proportional Odds Model through kernel learning techniques," *Neurocomputing*, vol. 164, pp. 23–33, Sep. 2015.

[41] M. Pérez-Ortiz, P. A. Gutiérrez, and C. Hervás-Martínez, "Projection-based ensemble learning for ordinal regression," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 681–694, May 2014.

[42] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional multilabel active learning with an efficient online adaptation model for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1880–1897, Oct. 2009.

[43] J. Sánchez-Monedero, P. A. Gutiérrez, P. Tino, and C. Hervás-Martínez, "Exploitation of pairwise class distances for ordinal classification," *Neural Comput.*, vol. 25, no. 9, pp. 2450–2485, Sep. 2013.

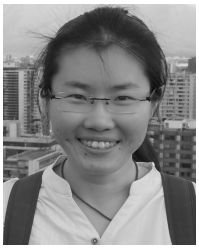
[44] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[45] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas. (2012). "Multi-label classification methods for multi-target regression." [Online]. Available: https://arxiv.org/abs/1211.6581

[46] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 906–910, Jun. 2010.

[47] B.-Y. Sun, H.-L. Wang, W.-B. Li, H.-J. Wang, J. Li, and Z.-Q. Du, "Constructing and combining orthogonal projection vectors for ordinal regression," *Neural Process. Lett.*, vol. 41, no. 1, pp. 139–155, 2015.

[48] L. Sun, S. Ji, and J. Ye, "Hypergraph spectral learning for multi-label classification," in *Proc. 14th ACM SIGKDD*, 2008, pp. 668–676.
[49] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.
[50] L. Sun, S. Ji, and J. Ye, *Multi-Label Dimensionality Reduction*. Boca Raton, FL, USA: CRC Press, 2013.
[51] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[52] H.-H. Tu and H.-T. Lin, "One-sided support vector regression for multiclass cost-sensitive classification," in *Proc. 27th ICML*, 2010, pp. 1095–1102.
[53] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *Proc. NIPS*, 2003, pp. 737–744.
[54] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. CVPR*, 2009, pp. 1643–1650.
[55] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *Proc. 11th ECCV*, 2010, pp. 126–139.
[56] H. Wang, H. Huang, and C. Ding, "Multi-label feature transform for image classifications," in *Proc. 11th ECCV*, 2010, pp. 793–806.
[57] H. Wu, H. Lu, and S. Ma, "A practical svm-based algorithm for ordinal regression in image retrieval," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 612–621.
[58] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proc. CVPR*, 2013, pp. 532–539.
[59] J. Ye, R. Janardan, and Q. Li, "GPCA: An efficient dimension reduction scheme for image compression and retrieval," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004, pp. 354–363.
[60] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proc. 28th ACM SIGIR*, 2005, pp. 258–265.
[61] Y. Yuan, K. Zhao, and H. Lu, "Multi-label linear discriminant analysis with locality consistency," in *Neural Information Processing* (Lecture Notes in Computer Science), vol. 8835, C. K. Loo, K. S. Yap, K. M. Wong, A. Teoh, and K. Huang, Eds. Cham, Springer, 2014.
[62] Y. Zhang and Z.-H. Zhou, "Multi-label dimensionality reduction via dependence maximization," in *Proc. 23rd AAAI*, 2008, pp. 1503–1505.
[63] B. Zhao, F. Wang, and C. Zhang, "Block-quantized support vector ordinal regression," *IEEE Trans. Neural Netw.*, vol. 20, no. 5, pp. 882–890, May 2009.
[64] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 450–461, Feb. 2015.



Jiabei Zeng received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2007 and 2011, respectively.

From 2013 to 2015, she was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA. She is currently an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. Her current research interests include computer vision and affective computing, especially on facial expression analysis.

Yang Liu (M'16) received the B.S. and M.S. degrees in automation from the National University of Defense Technology, Changsha, China, in 2004 and 2007, respectively, and the Ph.D. degree in computing from The Hong Kong Polytechnic University, Hong Kong, in 2011.

Between 2011 and 2012, he was a Post-Doctoral Research Associate with the Department of Statistics, Yale University, New Haven, CT, USA. He is currently a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include cognitive science, machine learning, applied mathematics, as well as their applications in high-dimensional data mining, complex network analysis, health informatics, brain modeling, multimedia content understanding, and music therapy.

Biao Leng received the B.Sc. degree from the School of Computer Science and Technology, National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2009.

He is currently an Associate Professor and a Ph.D. Supervisor with the School of Computer Science and Engineering, Beihang University, Beijing. His current research interests include 3-D model retrieval, image processing, computer vision, and data mining.

Zhang Xiong is a Full Professor with the School of Computer Science of Engineering, Beihang University, Beijing, China, and the Director of the Advanced Computer Application Research Engineering Center of National Educational Ministry, Beihang University. He is currently the Chief Scientist of smart city project supported by the National High Technology Research and Development Program of China. His current research interests include computer vision, wireless sensor networks and information security, where he has authored on the above topics.

Prof. Xiong serves as a member of several national committees, e.g., the National Computer Science and Technology Teaching Steering Committee of Ministry of Education. He was a recipient of the National Science and Technology Progress Award. He is the Executive Editors-in-Chief of *Frontiers of Computer Science*.

Yiu-ming Cheung (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung is an IET Fellow, a BCS Fellow, and an IETI Fellow. He is the Founding Chairman of Computational Intelligence Chapter of the IEEE Hong Kong Section. Also, he is now serving as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, PATTERN RECOGNITION, KNOWLEDGE AND INFORMATION SYSTEMS, and the *International Journal of Pattern Recognition and Artificial Intelligence*, among others.