Ordinal Margin Metric Learning and Its Extension for Cross-Distribution Image Data

Qing Tian[a], Songcan Chen[a,*], Lishan Qiao[b]

[a]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[b]School of Mathematical Science, Liaocheng University, Liaocheng 252059, China

Abstract

In machine learning and computer vision fields, a wide range of applications, such as human age estimation and head pose recognition, are related to ordinal data in which there exists an order relationship. To perform such ordinal estimations in a desired metric space, in this work we first propose a novel *ordinal margin metric learning* (ORMML) method by separating the data classes with a sequence of margins, which makes the classes distribute orderly in the learned metric space. Then, to cope with more realistic scenarios where the data are sampled with each class across multiple distributions, we present a cross-distribution variant of ORMML, coined as CD-ORMML, by maximizing the correlation between distributions within each class when conducting metric learning. Finally, extensive experiments on synthetic and publicly available image datasets demonstrate the superiority of the proposed methods in performance to the state-of-the-art methods.

Keywords: Ordinal metric learning, Human age estimation, Ordinal relationship, Cross-distribution.

1. Introduction

In machine learning and computer vision fields, most of the popular methods, such as the Support Vector Machines [11], decision tree [35], k-nearest-neighbor classifier [3], logistic regression [13] and k-means clustering [4], are performed in the traditional Euclidean space. Although many of them can work well and usually yield competitive results, their learning procedures essentially rely on the distance metric, especially for the k-nearest-neighbor classifier and k-means clustering. In real world applications, the dimensionality of data is usually high with diverse and complex distributions, which might deteriorate the estimation performance if using the Euclidean metric, since the high-dimensional spatial relationship of the data in the Euclidean distance space cannot be properly depicted and tends to be distorted [46].

To handle the aforementioned drawbacks of the Euclidean metric, researchers have developed a variety of metric learning methods. In early 2002, Xing et al. [46] proposed a pioneering metric learning method called MMC by enlarging the dissimilar pairwise distances, and meanwhile compacting the similar ones. Schultz et al. [36] presented a triplet-based metric learning method, in which the metric matrix is decomposed as the product of a predefined matrix and a diagonal metric matrix to make the optimization tractable. Goldberger et al. [8] proposed Neighborhood Component Analysis (NCA) to learn a similarity metric by maximizing the overall stochastic nearest neighbor probabilities. However, the objective function of NCA is nonconvex, and thus easily results in local optima. To improve the performance of NCA, Hong et al. [12] attempted to learn a metric by combining multiple objectives of NCA, and Tarlow et al. [38] extended NCA to the unsupervised scenarios. Later, Globerson et al. [7] developed a Maximally Collapsing Metric Learning (MCML) method by encouraging the distribution of data towards the ideal state based on KL-divergence regularization. Weinberger et al. [44] proposed Large Margin Nearest Neighbors (LMNN) learning, one of the most well-known metric learning methods, by making the distances between neighboring similar

*Corresponding author
Email address: s.chen@nuaa.edu.cn (Songcan Chen)

samples larger than those of dissimilar ones. Besides, Bar-Hillel et al. [1] presented Relevant Component Analysis (RCA) by introducing the "chunklet" to accommodate the similar samples and taking the inverse matrix of the average within-chunklet covariance as the metric matrix to perform distance measure. However, in practice, RCA is not so competitive because the discriminative between-chunklet information is ignored. To address this problem, Yeung et al. [49] extended RCA by putting the between- and within- chunklet information together into the learning algorithm. After that, Davis et al. [2] proposed the well-known Information-Theoretic Metric Learning (ITML) by incorporating Bregman divergence term into the objective function as well as imposing pairwise distance constraints between similar and dissimilar samples, respectively. Additionally, Qi et al. [34] introduced the L_1-regularization on the Mahalanobis matrix to conduct Sparse Distance Metric Learning (SDML) to handle high-dimensional data. More recently, Mei et al. [26] performed metric learning by minimizing LogDet divergence with triplet constraints, and achieved competitive results on several pattern recognition tasks. Wan et al. [40] proposed a robust metric learning approach to address the facial expression recognition problem. To utilize abundant unlabeled data for metric learning, Niu et al. [30] developed a semi-supervised variant of ITML through entropy regularization. More importantly, to encode the semantic contents of data into metric learning apart from the feature similarity, Yu et al. [50] proposed to learn a metric space by taking both the feature similarity and label-based semantic relations into account for more competitive image clustering. Zhu et al. [54] proposed to learn Neighborhood Distance Metric Learning (NDML) by considering the locality, compactness and consistency of training data simultaneously, and then extended NDML to its ensemble counterpart by integrating multi-granularity neighborhood distance metrics. Besides the aforementioned works, many application-specific metric learning methods have also been proposed to cater for object reidentification [31], [33], [53], [21], cartoon synthesis [51], image classification [18], [27], kinship verification [23], object tracking [15], [37], and bioinformatics [41], etc.

Actually, in recent years a number of multi-task metric methods have also been proposed, such as mt-LMNN [32], MLCS [48] and GPML [47]. However, multi-task scenario is not our concern here and thus we ignore their details (please interested readers refer to the related literature).

Although most of the metric learning methods above can be directly employed to handle ordinal data concerned in this work, they may lead to suboptimal performance because the ordinal information of data is not considered in them. For instance, for human age estimation, age 15 is elder than age 10 but younger than age 20. If we conduct age estimation as a traditional multi-class classification problem, such an ordinal relationship will be ignored, thus making the age estimation less reliable in age-based security surveillance (as we know, individuals at ages of 10, 15, and 20 should undertake definitely diverse criminal liability in crime). To this end, Xiao et al. [45] presented an ordinal metric leaning method, named mkNN, for ordinal age estimation. In mkNN, the distances from one sample to the samples within its k nearest neighbors are scaled proportional to their label difference, and meanwhile the distances between pairs of samples outside are enlarged. While Fouad et al. [6] extended the ITML to its ordinal counterpart, named OITML, by incorporating the ordinality information into the constraints. More recently, Li et al. [19] proposed a ranking metric learning approach, named LDMLR, by preserving distances of samples within a local neighborhood while enlarging the distances between other samples according to their rank differences, and then further extended LDMLR to multiple kernel version to promote its generalization ability. As a result, the estimations made in the LDMLR metric space on many problems, such as human age estimation, head pose regression and image retrieval, achieve state-of-the-art.

Although the aforementioned ordinal metric learning methods, especially LDMLR, can yield competitive results on ordinal problems, there still exist some shortcomings with them. Concretely, in mkNN, only the ordinal relationships within a local neighborhood of samples are considered, but the ordinal relationships outside the neighborhood are neglected. In OITML, a predefined distance relationship on the training data in the original Euclidean space is introduced to the metric learning process, however, such a predefinition misleads the learning since the distance relationships in the Euclidean space are usually not consistent with the ones defined in the desired metric space. Finally, in LDMLR, although it attempts to preserve the global ordinal relationships between data by weighting the between-rank distances according to their rank difference, maximizing the sum of weighted distances is not powerful enough to enforce the global ordinal relationship. To compensate for such drawbacks of the existing ordinal metric learning methods and more effectively preserve the global ordinal relationship of the data, we propose a novel ordinal metric learning method, coined as *ORdinal Margin Metric Learning* (ORMML), by enforcing global distance constraints on all the classes jointly. Then, to cope with realistic estimation scenarios where the ordinal data classes are drawn across multiple distributions, we further extend ORMML to its cross-distribution counterpart by incorporating the

correlations between the data distributions within data classes into metric learning. Finally, extensive experiments on synthetic and real-world image datasets demonstrate the superiority of the proposed methods.

The rest of this paper is organized as follows. In Section 2, we briefly review several related works. In Section 3, we first propose our novel ORdinal Margin Metric Learning (ORMML), and then extend it to its cross-distribution counterpart (i.e., CD-ORMML). In Section 4, we provide time complexity analysis of the proposed methods. In Section 5, we report the experimental results to demonstrate the superiority of the proposed methods. Finally, we conclude this paper in Section 6.

For the sake of understanding the mathematical symbols involved in this paper, we list their meanings in Table 1.

Table 1: Meanings of mathematical symbols involved in this paper.

Notation	Meaning
λ, λ_1, λ_2:	Non-negative trade-off parameters
ϵ, β, ξ:	Slack variables
η:	Margin scale coefficient of (CD-)ORMML
w:	Weighting coefficient
x_i:	The i-th training sample
x_i^k:	The i-th training sample from the k-th class
$A \succcurlyeq 0$:	Positive semi-definite metric matrix
$d_A(x_i, x_j)$:	Mahalanobis distance between x_i and x_j characterized by A
C_k:	The mean vector of the k-th class
C_k^j:	The mean vector of the j-th distribution of the k-th class
$\widetilde{C}_k$:	The mean vector of the k-th class with multiple distributions
S_+ (S_-):	The similarity (dissimilarity) constraint set

2. Related Works

Before presenting our methods, we briefly review mkNN [45], OITML [6], and LDMLR [19], which are three representative ordinal metric learning methods mostly related to our work.

2.1. mkNN

In [45], Xiao et al. proposed an ordinal distance metric learning method, named mkNN, to preserve the local semantic neighborhood under the assumption that data within the same semantic neighborhood should have more similar labels than those outside. Based on such an idea, they formulated the mkNN as follows:

$$\begin{aligned} max_A \ & \sum_{i,j} tr(AM_{ij}) - \lambda \sum_{i,j} \epsilon_{ij} \\ s.t. \quad & tr(AM_{ij}) + \epsilon_{ij} = \hat{d}_{ij}^2, \ if \ \eta_{ij} = 1 \\ & A \succeq 0 \\ & \epsilon_{ij} \geq 0, \end{aligned} \tag{1}$$

where A is the Mahalanobis metric matrix, $tr(AM_{ij}) = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$ stands for the Mahalanobis distance between two samples, $M_{ij} = (x_i - x_j)(x_i - x_j)^T$, λ is a nonnegative trade-off parameter, ϵ_{ij} is the slack variable, $\eta_{ij} = 1$ denotes that x_i and x_j are in the same neighborhood and $\eta_{ij} = 0$ otherwise. Further,

$$\hat{d}_{ij} = \Big(\frac{L(x_i, x_j) + \gamma}{C - L(x_i, x_j)}\Big)^p \times d(x_i, x_j), \tag{2}$$

in which $L(x_i, x_j)$ denotes the absolute value of label difference between x_i and x_j, γ refers to the labelling noise, $C = max\{L(x_i, x_j)\} + \epsilon, \epsilon > 0$ ensures the denominator not equal to zero, p is a predefined parameter, and $d(x_i, x_j)$ denotes the Euclidean distance between samples x_i and x_j.

From the above Eqs. (1) and (2), it can be found that mkNN attempts to learn an ordinal metric space by rescaling the pairwise distances of data within the same neighborhood while enlarging the ones outside.

2.2. OITML

In [6], Fouad et al. developed an ordinal counterpart of ITML, named OITML, by weighting the pairwise similarity and dissimilarity constraints on samples with weights proportional to their label difference, and formulated the model as follows:

$$\begin{aligned} &min_{\{A>0,\xi\}}\ D_{Burg}(A, M_0) + \lambda \cdot D_{Burg}(diag(\xi), diag(\xi_0)) \\ &s.t. \qquad d_A(x_i, x_j) \le \xi_{s(i,j)} \cdot \vartheta^{+}_{ij},\ \textbf{if}\ (x_i, x_j) \in S_+, \\ &\qquad\ \ d_A(x_i, x_j) \ge \xi_{s(i,j)} \cdot \vartheta^{-}_{ij},\ \textbf{if}\ (x_i, x_j) \in S_-, \end{aligned} \tag{3}$$

where $D_{Burg}(A, M_0) = tr(AM_0^{-1}) - logdet(AM_0^{-1}) - d$ represents the Bregman divergence defining the similarity between metric matrix A and a predefined metric matrix M_0, and similarly, $D_{Burg}(diag(\xi), diag(\xi_0))$ measures the similarity between slack variables ξ and their initial values ξ_0, $d_A(x_i, x_j)$ represents the distance parameterized by A between two samples x_i and x_j from either similarity constraint set S_+ or dissimilarity constraint set S_-, and ϑ^{+}_{ij} (ϑ^{-}_{ij}) denotes the weight coefficient for corresponding similarity (dissimilarity) constraint whose value is proportional to the class label difference of x_i and x_j, and λ is a nonnegative trade-off parameter balancing the two divergence terms.

From Eq.(3), it can be found that the ordinality weight information is incorporated in the constraints, by which the original ITML is extended to its ordinal counterpart.

2.3. LDMLR

In [19], Li et al. proposed the *Linear Distance Metric Learning for Ranking*, i.e., LDMLR, to learn an ordinal metric by preserving both the local geometry structure and the ordinal relationship of the data, and formulated the objective function as follows:

$$\begin{aligned} &min_A\ -\sum_{i,j} w_{ij} d_A^2(x_i, x_j) + \lambda \sum_{i,j} \epsilon_{ij} \\ &s.t. \qquad (d_A^2(x_i, x_j) - d_I^2(x_i, x_j)) = \epsilon_{ij},\ if\ \eta_{ij} = 1 \\ &\qquad\ \ A \succeq 0 \\ &\qquad\ \ \epsilon_{ij} \ge 0, \end{aligned} \tag{4}$$

where w_{ij} denotes the ordinal weight factor, $d_A^2(x_i, x_j) = (x_i - x_j)^T A (x_i - x_j)$ denotes the Mahalanobis distance between samples x_i and x_j characterized by metric matrix A, $d_I^2(x_i, x_j)$ denotes the Euclidean distance between x_i and x_j, ϵ_{ij} is a slack variable, $\eta_{ij} = 1$ indicates samples x_i and x_j are in the same target neighborhood, $\eta_{ij} = 0$ otherwise, and λ is a nonnegative trade-off parameter.

From the above formulation (4), it can be found that LDMLR intends to preserve the local geometry relationships of within-rank data in each semantic neighborhood, and meanwhile enlarge the between-rank pairwise distances according to the rank difference.

2.4. Comment

After analyzing formulations (1), (3) and (4) (corresponding to mkNN, OITML and LDMLR, respectively), we can find their limitations in preserving the ordinal relationship of data. Concretely, in mkNN, it can only preserve the local ordinal relationships of data in a local neighborhood since the pairwise distances of data outside the neighborhood are enlarged with identical weights, as stated by the first term of formulation (1). In OITML, although imposing ordinal constraints between pairs of the samples, it drives the metric matrix to be as similar as possible to a predefined distance matrix constructed in the original Euclidean space, doing so will mislead the learning due to that the distance relationships in the Euclidean space are usually not consistent with the ones constructed in the desired metric space. Even worse, when the amount of training data is large, extremely large number of constraints will be involved, thus making the learning prone to infeasible. Finally, in LDMLR, although the between-rank distances have been weighted in terms of the rank difference (as stated by the first term of the objective function in (4)), it may be still not powerful enough to learn a desirable ordinal metric because maximizing a sum of the weighted distances cannot necessarily guarantee the given ordinal relationship of the data. In other words, such a weighting strategy is not preferable to pursue an ordinal metric. We take Figure 1 as an example, where x_k, x_{k+i} and x_{k+i+j} denote three samples from the

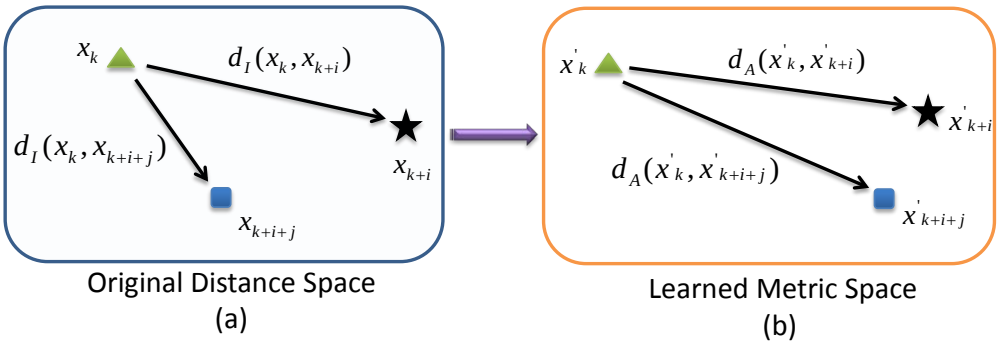


Figure 1: Illustration of the limitation of LDMLR in preserving the ordinal relationship of data. x_k, x_{k+i} and x_{k+i+j} denote samples from the k-th, $(k+i)$-th and $(k+i+j)$-th classes, respectively, while x'_k, x'_{k+i} and x'_{k+i+j} denote their corresponding new distributions in the learned metric space. $d_I(\cdot,\cdot)$ represents the distance between two samples in the original metric space, while $d_A(\cdot,\cdot)$ describes the distance in the new metric space learned by LDMLR.

k-th, $(k+i)$-th and $(k+i+j)$-th ordinal classes, respectively. From the perspective of ordinal distribution, x_{k+i} is nearer to x_k than to x_{k+i+j} since the label difference between x_k and x_{k+i} is i, smaller than difference $i+j$ between x_k and x_{k+i+j}, the original distance between x_k and x_{k+i}, however, is much larger than that between x_k and x_{k+i+j}. As a result, in the learned metric space the distance between x_k and x_{k+i}, via LDMLR, might be still larger than that between x_k and x_{k+i+j}, implying that the ordinal relationship between x_k, x_{k+i} and x_{k+i+j} cannot be preserved even the distance weight for x_{k+i+j} is larger than that for x_{k+i} (see the first term of the objective function in (4)).

3. Proposed Methodology

To overcome the drawbacks of the ordinal metric learning methods aforementioned, in this section we first propose a novel ordinal metric learning method, coined as ORdinal Margin Metric Learning (ORMML). Then, to cope with the scenario where data are sampled across multiple distributions, we extend ORMML to its cross-distribution counterpart, named CD-ORMML.

3.1. ORdinal Margin Metric Learning (ORMML)

Suppose that we have a total of N data samples $\{x_i, y_i\}_{i=1}^N$, from K ordinal classes with $x_i \in \mathcal{R}^D$ and its corresponding label $y_i \in \{1, 2, ..., K\}$, and denote the mean vectors of the K classes as $\{C_1, C_2, ..., C_K\}$. In what follows, we attempt to construct ORMML by compacting the within-class distances and simultaneously separating the between-class distances, thus making the data classes distribute orderly. For the sake of clarification, we illustrate the key idea behind our methodology in Figure 2.

As shown in Figure 2, for the sample x_{k-i} from the $(k\text{-}i)$-th class, the distance $d_A(x_{k-i}, C_k)$ between it and C_k should be no less than $\eta \times i$ with η being the margin scale, and the distance $d_A(x_{k-i}, C_{k+j})$ between it and C_{k+j} should be at least $\eta \times (i+j)$; for the sample x_k from the k-th class, the distance $d_A(x_k, C_{k-i})$ between it and C_{k-i} should be $\eta \times i$, and the distance $d_A(x_k, C_{k+j})$ between it and C_{k+j} should be at least $\eta \times j$; for the sample x_{k+j} from the $(k+j)$-th class, the distance $d_A(x_{k+j}, C_{k-i})$ between it and C_{k-i} should be $\eta \times (i+j)$, and the distance $d_A(x_{k+j}, C_k)$ between it and C_k should be $\eta \times j$. When we take into account the distance relationships above together, a global ordinal relationship among the classes will emerge automatically, as demonstrated in Figure 2. By this way, an ordinal metric space can

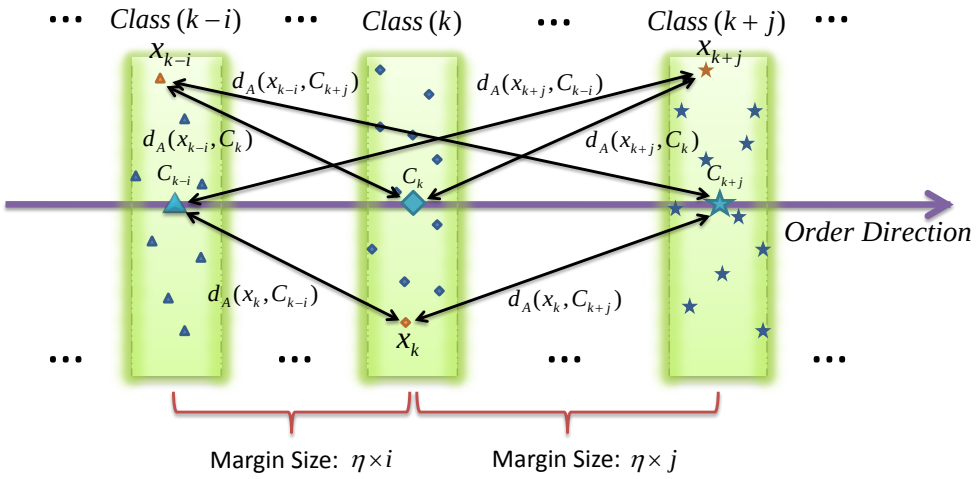


Figure 2: Illustration of ORMML. Here, x_{k-i}, x_k and x_{k+j} respectively denote samples from the $(k\text{-}i)$-th, k-th and $(k+j)$-th classes, η is the margin scale, and the formulation $d_A(\cdot,\cdot)$ denotes the Mahalanobis metric parameterized by metric matrix A.

be generated. Along this line, we can mathematically model the ORMML as follows:

$$min_A \sum_{k,i} \beta_{ki} + \lambda \sum_{p,q,i} \xi_{pqi} \cdot w_{pq} \tag{5a}$$

$$s.t. \quad d_A^2(x_i^p, C_q) \geq \eta|p-q| - \xi_{pqi}, \ (p \neq q) \tag{5b}$$

$$d_A^2(x_i^k, C_k) \leq \beta_{ki} \tag{5c}$$

$$A \geq 0 \tag{5d}$$

$$\beta_{ki} \geq 0 \tag{5e}$$

$$\xi_{pqi} \geq 0 \tag{5f}$$

where λ is a nonnegative trade-off parameter, x_i^p denotes the i-th sample from the p-th class, C_q stands for the mean vector of the q-th class, ξ_{pqi} is a slack variable, η is a manually defined parameter to control the margin size, $|\cdot|$ denotes the absolute-value operator, $w_{pq} = |p-q|^h$ is the weighting factor with a predefined exponential parameter h, and

$$d_A^2(x_i^p, C_q) = (x_i^p - C_q)^T A (x_i^p - C_q). \tag{6}$$

From (6), it can be found that when the metric matrix A is set to the identity matrix, the metric reduces to the Euclidean metric.

As demonstrated in (5), by means of such distance constraints between- (corresponding to (5b)) and within- classes (corresponding to (5c)), the distance space is enforced to be ordinal in terms of the class order.

Clearly, the problem in (5) is a semi-definite programming (SDP), so in this work we employ the projected gradient descent method [22] to solve it, due to its simplicity and effectiveness. More specifically, for the sake of computation,

we eliminate all the slack and auxiliary variables from (5) and reformulate it as

$$\begin{aligned}
&min\ \mathcal{J}_{ORMML}(A)\\
&= \sum_{k,i} d_A^2(x_i^k, C_k) + \lambda \sum_{p,q,i} max\{0,\ \eta|p-q| - d_A^2(x_i^p, C_q)\} \cdot w_{pq}\\
&= \sum_{k,i} (x_i^k - C_k)^T A (x_i^k - C_k)\\
&\quad + \lambda \sum_{p,q,i} max\{0,\ \eta|p-q| - (x_i^p - C_q)^T A (x_i^p - C_q)\} \cdot w_{pq}\\
&= \sum_{k,i} tr\big((x_i^k - C_k)^T A (x_i^k - C_k)\big)\\
&\quad + \lambda \sum_{p,q,i} max\{0,\ \eta|p-q| - tr\big((x_i^p - C_q)^T A (x_i^p - C_q)\big)\} \cdot w_{pq}
\end{aligned} \tag{7}$$

Taking a derivative of $\mathcal{J}_{ORMML}(A)$ with respect to A, we then obtain

$$\frac{\partial \mathcal{J}_{ORMML}(A)}{\partial A} = \begin{cases} \displaystyle\sum_{k,i} (x_i^k - C_k)(x_i^k - C_k)^T, \\ \qquad \textbf{if } \eta|p-q| \le tr\big((x_i^p - C_q)^T A (x_i^p - C_q)\big); \\ \displaystyle\sum_{k,i} (x_i^k - C_k)(x_i^k - C_k)^T - \lambda \sum_{p,q,i} (x_i^p - C_q)(x_i^p - C_q)^T \cdot w_{pq}, \\ \qquad \textbf{if } \eta|p-q| > tr\big((x_i^p - C_q)^T A (x_i^p - C_q)\big). \end{cases} \tag{8}$$

Now, let A_t denote the value of A after the t-th iteration of projected gradient descent, then A_{t+1} can be computed by

$$A_{t+1} = A_t - \sigma \frac{\partial \mathcal{J}_{ORMML}(A)}{\partial A}\Big|_{A_t} \tag{9}$$

where σ is the learning rate computed via linear search [14].

To ensure the positive semi-definiteness of A_{t+1}, we perform spectral decomposition of A_{t+1} as

$$A_{t+1} = U_{t+1} \Lambda_{t+1} U_{t+1}^T, \tag{10}$$

and then threshold the eigenvalues to make them non-negative by

$$A_{t+1} = U_{t+1} max(0,\ \Lambda_{t+1}) U_{t+1}^T. \tag{11}$$

We iteratively update A_{t+1} based on (8)-(11) until it converges or the maximal iteration number is reached. Table 2 summarizes our whole ORMML algorithm. With the obtained metric matrix A through ORMML algorithm, we can evaluate the distance between two data points in the learned ordinal metric space.

3.2. Cross-Distribution ORMML (CD-ORMML)

In real world applications, usually the data of interest are sampled across multiple distributions. That is, the data within each class can be grouped into two or more clusters with each cluster representing one distribution. For example, in age estimation human aging data are usually sampled across different races such as Caucasian and African. To preferably perform ordinal metric learning in such a cross-distribution scenario, we should take into account the potential correlation between the distributions, and demonstrate our idea in Figure 3. To be specific, because of the positive semi-definiteness of the metric matrix A in formulation (5), it can be decomposed into $A = P^T P$.

Table 2: Algorithm of ORMML.

Input:	Training instances $\{x_i\}_{i=1}^N$ and ordinal labels $\{y_i\}_{i=1}^N$; Parameters λ, h, and η.
Output:	metric matrix A.

1. Initialize $A_0 = I_D$;
2. **for** $t = 1, 2, ..., T_{max}$ **do**
3. Compute $\frac{\partial \mathcal{J}_{ORMML}(A)}{\partial A}$ based on (8);
4. Update A_t based on (9);
5. Project A_t onto the PSD cone based on (10) and (11);
6. **end for**
7. $A \Leftarrow A_t$.

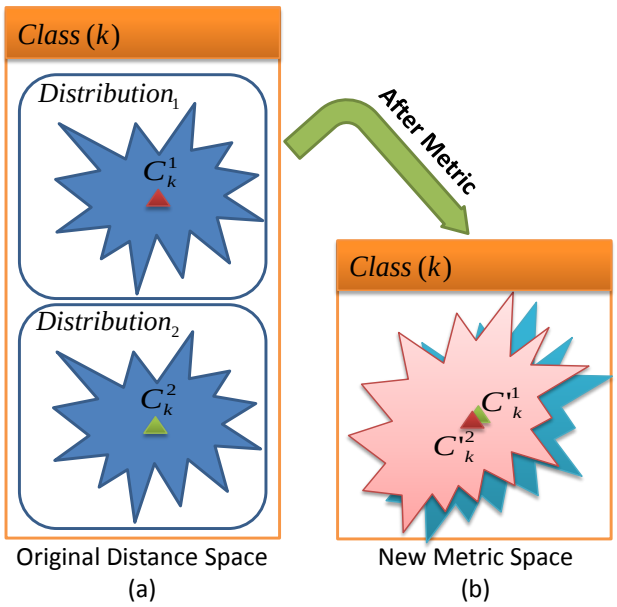


Figure 3: Illustration of cross-distribution metric learning. Here, $Distribution_1$ and $Distribution_2$ denote two different distributions within the same class k, and C_k^1 and C_k^2 stand for their corresponding sample means, respectively. After metric learning, the two distributions are projected into a metric space where their new means, $C_k^{'1}$ and $C_k^{'2}$, are pulled much closer towards each other.

Consequently, the Mahalanobis metric between two data points can be identically treated as a two-step procedure: projecting them by the P and then evaluating their Euclidean distance in the projected space. Now, the Mahalanobis metric (6) can be transformed as follows:

$$\begin{aligned} & d_A^2(x_i^p, C_q) \\ &= (x_i^p - C_q)^T A (x_i^p - C_q) \\ &= (x_i^p - C_q)^T P^T P (x_i^p - C_q) \\ &= (P x_i^p - P C_q)^T (P x_i^p - P C_q) \\ &= (x_i^{'p} - C_q^{'})^T (x_i^{'p} - C_q^{'}) \end{aligned} \tag{12}$$

where $x_i^{'p}$ and $C_q^{'}$, respectively, denote the projected representations of the x_i^p and C_q in the Euclidean space. Using the transformation (12), the Mahalanobis distance between x_i^p and C_q is successfully converted to the Euclidean distance

between $x_i^{'p}$ and $C_q^{'}$. Along this line, we can represent the correlation between two distributions of the k-th class, shown in Figure 3, as follows:

$$\mathcal{R}(C_k^1, C_k^2) = (PC_k^1)^T(PC_k^2) = (C_k^1)^T P^T P C_k^2 = (C_k^1)^T A C_k^2. \tag{13}$$

Essentially, the term (13) describes the inner-product between the mean vectors, C_k^1 and C_k^2, of the two distributions [1] from the k-th class, and the larger the inner-product is, the more similar the two vectors PC_k^1 and PC_k^2 are, i.e., much nearer to each other in the metric space. By integrating the correlation term (13) into the objective function of ORMML, we can develop its cross-distribution counterpart, coined as CD-ORMML, as follows:

$$min_A \ \sum_{k,i}\beta_{ki} + \lambda_1 \sum_{p,q,i}\xi_{pqi}\cdot w_{pq} - \lambda_2 \sum_k (C_k^1)^T A C_k^2 \tag{14a}$$

$$s.t. \quad d_A^2(x_i^p, \widetilde{C_q}) \geq \eta|p-q| - \xi_{pqi}, \ (p \neq q) \tag{14b}$$

$$d_A^2(x_i^k, \widetilde{C_k}) \leq \beta_{ki} \tag{14c}$$

$$A \geq 0 \tag{14d}$$

$$\beta_{ki} \geq 0 \tag{14e}$$

$$\xi_{pqi} \geq 0 \tag{14f}$$

where x_i^p is a sample from the p-th class, C_k^1 and C_k^2 denote the mean vectors of the first and second distributions of the k-th class, respectively, while $\widetilde{C_q}$ stands for the total mean vector across the distributions of the q-th class, and λ_1 and λ_2 are two nonnegative trade-off parameters to balance the within-distribution loss and the between-distribution correlation.

Similar to ORMML in (5), the formulation (14) of CD-ORMML is also a SDP problem, so we also adopt the projected gradient descent to solve it. After eliminating the slack and auxiliary variables, rewriting the objective function, and taking a derivative of (14) with respect to A, we obtain

$$\frac{\partial \mathcal{J}_{CD-ORMML}(A)}{\partial A} = \begin{cases} \sum_{k,i}(x_i^k - \widetilde{C_k})(x_i^k - \widetilde{C_k})^T - \lambda_2 C_k^1 (C_k^2)^T, \\ \quad \textbf{if } \eta|p-q| \leq tr((x_i^p - \widetilde{C_q})^T A (x_i^p - \widetilde{C_q})); \\ \sum_{k,i}(x_i^k - \widetilde{C_k})(x_i^k - \widetilde{C_k})^T - \lambda_1 \sum_{p,q,i}(x_i^p - \widetilde{C_q})(x_i^p - \widetilde{C_q})^T \cdot w_{pq} \\ \quad -\lambda_2 C_k^1 (C_k^2)^T, \\ \quad \textbf{if } \eta|p-q| > tr((x_i^p - \widetilde{C_q})^T A (x_i^p - \widetilde{C_q})). \end{cases} \tag{15}$$

Let A_t denote the value of A after the t-th iteration of projected gradient descent, then A_{t+1} can be computed by

$$A_{t+1} = A_t - \sigma \frac{\partial \mathcal{J}_{CD-ORMML}(A)}{\partial A}\Big|_{A_t}. \tag{16}$$

Similar to (10)-(11), to guarantee the PSD characteristic of A_{t+1}, we perform spectral decomposition of A_{t+1} as

$$A_{t+1} = U_{t+1}\Lambda_{t+1}U_{t+1}^T, \tag{17}$$

and threshold the eigenvalues to make them non-negative by

$$A_{t+1} = U_{t+1} max(0, \ \Lambda_{t+1}) U_{t+1}^T. \tag{18}$$

Based on (15)-(18), we iteratively update A_{t+1} until it converges or the maximal iteration number is reached. Finally, we summarize the CD-ORMML algorithm in Table 3. After obtaining the metric matrix A of CD-ORMML, we can perform cross-distribution ordinal distance estimations.

[1] For the sake of clarification, we in this work take two distributions as example, whereas it can be similarly extended to multiple distributions via accumulating strategy.

Table 3: Algorithm of CD-ORMML.

Input:	Training instances $\{x_i\}_{i=1}^N$ and ordinal labels $\{y_i\}_{i=1}^N$; Parameters λ_1, λ_2, h, and η.
Output:	metric matrix A.
1. Initialize $A_0 = I_D$;	
2. **for** $t = 1, 2, ..., T_{max}$ **do**	
3.	Compute $\frac{\partial \mathcal{J}_{CD-ORMML}(A)}{\partial A}$ based on (15);
4.	Update A_t based on (16);
5.	Project A_t onto the PSD cone based on (17) and (18);
6. **end for**	
7. $A \Leftarrow A_t$.	

In practice, we may only possess the data sampled across distributions without their concrete distribution information. In this case, in order to make the CD-ORMML applicable, we can adopt existing unsupervised clustering algorithms, such as k-means clustering [17] and hierarchical clustering [29], to obtain the within-class distributions for subsequent use of CD-ORMML.

Clarification

The *cross-distribution metric learning* in CD-ORMML may at first glance make the readers mistake it for the *cross-modality metric learning* or *multi-task metric learning*. Actually, there are essential distinctions between them. To be specific, in multi-task metric learning, usually fixed numbers of two or more metric tasks are involved, each of which associates with an unshared metric matrix to be learned, and the tasks achieve to improve their respective learning by means of collaborating with the others. More importantly, each of the tasks has its own (unshared) training data as in [48], [47], [5], [25]. By contrast, throughout the learning process of CD-ORMML, only one metric matrix is involved (see the metric matrix A in the formulation (14)), meaning that our CD-ORMML is actually a single task learning. Moreover, since the number of distributions in each class in our cross-distribution metric learning scenario may be different, the multi-task based metric methods with fixed number of tasks cannot be applied here. In cross-modality metric learning, each of training/testing samples has the same number and types of modalities, and each modality correspond to a type of feature representation as in [28], [16], [42], [52], [43]. By contrast, in our cross-distribution learning of CD-ORMML, each of the training/testing samples has only one type of feature representation. That is, the cross-distribution learning scenario in CD-ORMML is essentially with single modality. Therefore, the cross-modality based metric learning is difficult to adapt to the cross-distribution scenario concerned in this work.

4. Time Complexity Analysis

As summarized in Table 2, the time complexity of ORMML algorithm mainly consists of two parts: initializing the metric matrix in line 1 and the iteration process from line 2 to line 6. Concretely, the initialization process for A_0 costs $O(D^2)$ with D being the data feature dimension. For the iteration procedure, the time complexity of computing the derivative $\frac{\partial \mathcal{J}_{ORMML}(A)}{\partial A}$ in line 3 is $O(max(ND^2,\ KND^2))$ with N being the total number of training samples and K the number of classes, updating A_t in line 4 needs $O(D^2)$, and the time complexity for line 5 is $O(2D^3)$. Therefore, the time complexity for lines 2-6 is $O(T_{max} \cdot max(ND^2 + D^2 + 2D^3,\ KND^2 + D^2 + 2D^3))$. In total, the time complexity of ORMML is $O(D^2 + T_{max} \cdot max(ND^2 + D^2 + 2D^3,\ KND^2 + D^2 + 2D^3))$.

For the CD-ORMML algorithm in Table 3, the time complexities of all the lines, except for line 3 whose time complexity is $O(max(ND^2 + D^2,\ KND^2 + D^2))$, are the same as those of ORMML in Table 2. Therefore, the total time complexity of CD-ORMML algorithm summarized in Table 3 is $O(D^2 + T_{max} \cdot max(ND^2 + 2D^2 + 2D^3,\ KND^2 + 2D^2 + 2D^3))$.

5. Experiment

In this section, we conduct experiments to make evaluations on the proposed methods.

5.1. Datasets and Settings

Prior to reporting the experimental results, we introduce the datasets used in the experiments and the settings, respectively.

Datasets: We first conduct human age estimation and head pose recognition experiments on three commonly used aging datasets, i.e., FG-NET, Morph Album I and Album II [9] and one head pose dataset UMIST, respectively. For FG-NET, it consists of 1,002 facial images taken from 82 individuals, and their ages range from 0 to 69 years old. For the Morph Album I, it contains 1,690 images from about 631 individuals mainly of African and European, and the age ranges from 15 to 68 years old. For the Morph Album II, it consists of over 55,000 images from about 5,475 white individuals as well as other ethnic ones, averagely with about 2 to 3 pictures per person aging from 16 to 77. Moreover, the head pose dataset UMIST contains about 564 pictures from 20 persons. Image examples from the four datasets are shown in Figure 4. Besides the above four face-domain related image datasets, we also conduct

Figure 4: Image examples from the FG-NET (a), the Morph Album I (b), the Morph Album II (c) and the UMIST (d).

experiments on a non-face related image dataset, i.e., the Mixed-Gambles task dataset [39], which is a functional MRI dataset.

Settings: Throughout the experiments, the trade-off parameter α in (1) and (4) is set to 10^3, and 10^8 in (5); the parameters λ_1 and λ_2 in (14) are set to 10^8 and 10^2, respectively. And we adopt the *mean absolute error* (MAE) as the performance measure, in which $MAE := \frac{1}{N}\sum_{i=1}^{N}|\widehat{l_i} - l_i|$ with l_i and $\widehat{l_i}$ denoting the ground-truth and predicted values, respectively. Moreover, without loss of generality we uniformly adopt the 5-nearest-neighbors as the predictor after the metric estimation. Besides the ordinal metric learning methods mkNN, OITML, and LDMLR reviewed in Section 2, to demonstrate the effectiveness and superiority of the proposed metric learning algorithms, we also introduce the representative metric learning methods LFDA [33], ITML [2], LMNN [44] and LDMLT [26] for experimental comparison.

5.2. Evaluation Irrespective of Data Distributions

In this subsection, we first conduct evaluations on the four face-related image datasets, irrespective of the data distributions. More specifically, to generate matching quantities of training data for each of the classes, we choose an age range from 0 to 19 years old from FG-NET to constitute four age groups, and from 15 to 39 years old to form five age groups from Morph Album I by dividing the ranges evenly. For the Morph Album II, we choose all the samples from the white race to constitute totally five age groups, i.e., 16-25, 26-35, 36-45, 46-55 and 56-77 years old. Moreover, to generate ordinal head pose classes with matching data, we divide the UMIST dataset into six angles of head pose, from frontal view to profile. On the generated datasets, we extract AAM features from the FG-NET and Morph Album I, BIF representations from the Morph Album II and raw pixel values to represent the head pose on

the UMIST. On the four datasets, we uniformly extract 95% energy of the feature representations for experiment, and report the experimental results, averaged over 10 random runs, in Figure 5. From it, we can find that,

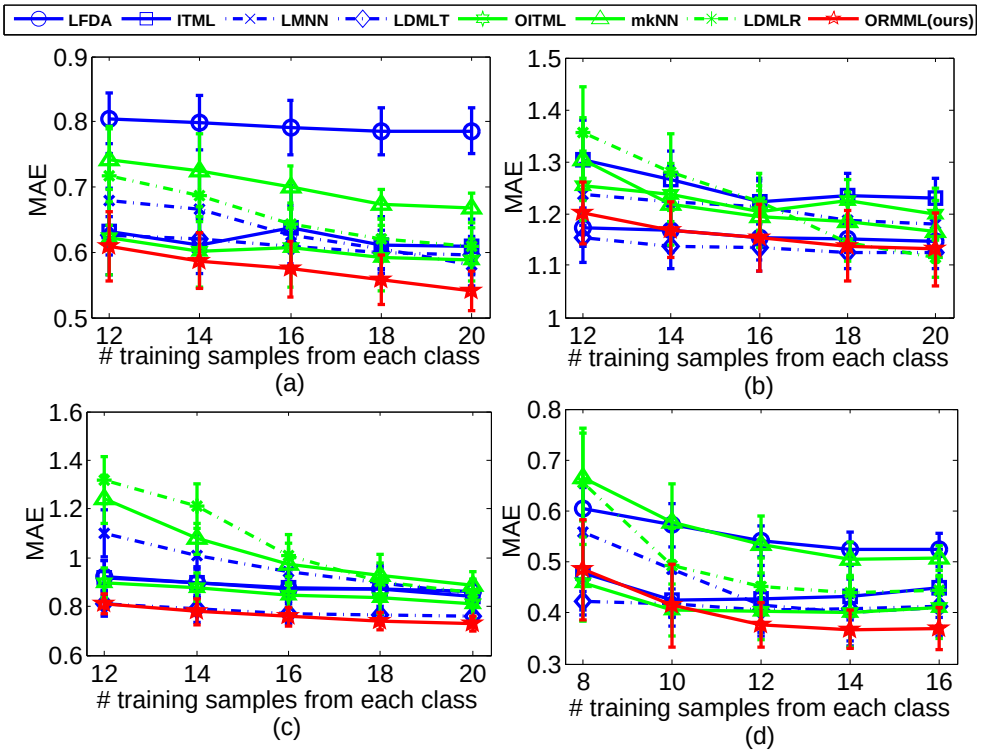


Figure 5: Comparison of estimations in different metric spaces on the FG-NET (a), Morph Album I (b), Morph Album II (c) and UMIST (d).

- Generally, the MAEs yielded by the weighting-based ordinal metric learning methods (i.e., the mkNN and LDMLR, plotted in green color) are higher than those by most of the non-ordinal metric learning methods (plotted in blue color). It shows that the strategy of weighting-based ordinal metric learning is not powerful enough to learn an effective ordinal metric to preserve the ordinal relationship between the data classes.

- The MAEs of OITML are consistently lower than those of its non-ordinal version, i.e., ITML. It shows that incorporating orinality weights can improve the performance of metric learning. On the other hand, the estimation accuracies of OITML are still not so desirable, even worse than some of the non-ordinal methods such as the LDMLT, which witnesses that inducing the metric matrix to learn following a metric relationship predefined in the original Euclidean space is likely to mislead the metric learning.

- In most cases, the proposed ORMML yielded the lowest MAEs, which demonstrates the effectiveness and superiority of our metric learning strategy.

Besides the experiments on the four face-related image datasets, we also conduct experiments on a non-face image datset, i.e., the Mixed-gambles task dataset, which contains a set of images with 16 ordinal levels. Similar to [19], we also adopt the GLM regression coefficients as the feature representation, and report the averaged experimental results over 10 runs in Figure 6, which demonstrates similar rules as in Figure 5. It shows that besides the face-domain image datasets, the proposed ORMML also works and demonstrates its superiority in performance on non-face image data.

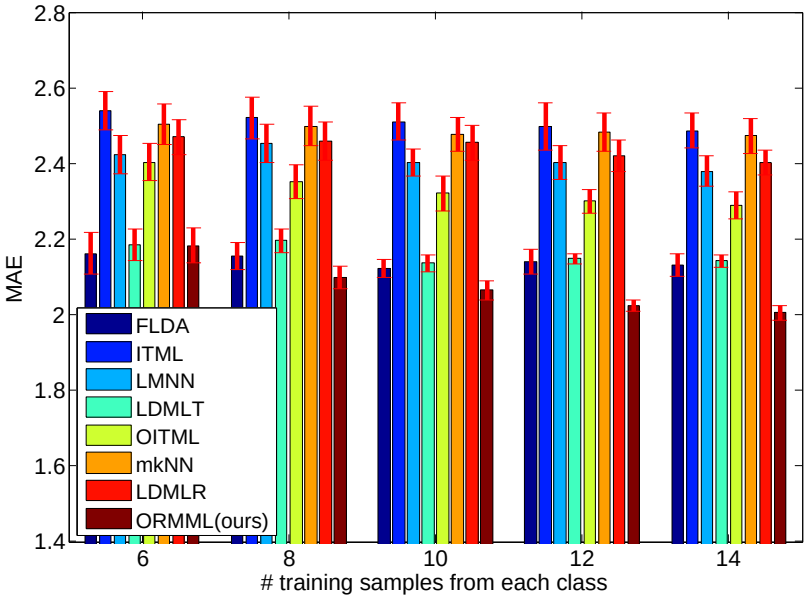
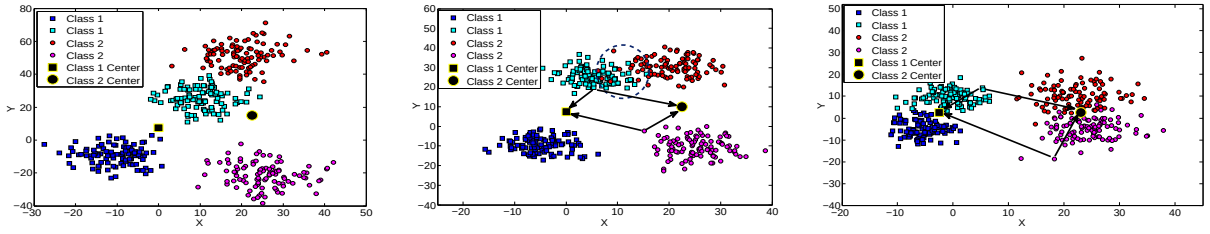


Figure 6: Comparison of estimations on the Mixed-gambles dataset.

5.3. Evaluation Across Data Distributions

5.3.1. On Synthetic Dataset

Prior to experiments on real-world cross-distribution data, we first experiment on a synthetic dataset and demonstrate in Figure 7 the superiority of CD-ORMML over ORMML in capturing the correlations of distributions within



(a) In original space (b) In ORMML space (c) In CD-ORMML space

Figure 7: Distributions of a synthetic dataset in the original (a), ORMML (b), and CD-ORMML (c) distance metric space. Different shapes represent the data classes, while different colors indicate the distributions.

the classes. From it, we can find that in the metric space of ORMML, there is a crossover between the distributions from different classes (see the dotted ellipse region in Figure 7(b)). As a comparison, in the CD-ORMML space the distributions of different classes are separated away with clear margin and the distributions belonging to the same classes are pulled much closer (see Figure 7(c)). It means that compared with ORMML, CD-ORMML can preserve more correlations between the distributions within each class and thus facilitates the subsequent estimations.

5.3.2. On Real-World Dataset

Besides the above synthetic experiment, we also conduct experiments on real cross-distribution data. To this end, we collect four types of race groups, i.e., White Male (WM), White Female (WF), Black Male (BM) and Black Female (BF), from the Morph Album II to simulate the cross-distribution scenario. Specifically, we randomly choose 8,057 WMs, 2,610 WFs, 8,057 BMs and 2,610 BFs, each of the races aging in the same age range as the Morph

Album II (i.e., 16 to 77 years old), and represent them by extracting about 95% principle components of BIFs as feature representation. Then, we divide each of the four races into five age groups, i.e., 16-25, 26-35, 36-45, 46-55 and 56-77 years old. Besides, to evaluate the applicability of CD-ORMML in handling such a scenario where the data are sampled across distributions but the concrete distribution information is unknown, we first adopt the widely-used k-means method to cluster the distributions and then perform CD-ORMML on the clustered data. For the sake of distinction, we call the CD-ORMML with such clustering *CD-ORMML (Distribution-Unknown)*, and meanwhile we call the CD-ORMML as *CD-ORMML (Distribution-Known)* which is directly performed on training data with distributions given. Finally, on the generated cross-distribution datasets, we conduct cross-distribution age estimation and report the results in Table 4. From it, we find that,

- By taking into account the correlation information between the distributions of each class, the CD-ORMML, no matter the data distribution information is given or not, reduce the estimation errors from 1.02 by ORMML to about 0.94, with about 8% MAE reduction. It confirms that introducing the distributions correlation into learning can further improve the prediction accuracy of ORMML. In addition, the CD-ORMML yields the lowest average estimation error (with the lowest MAEs in most cases) among all the compared methods. It demonstrates the superiority of the CD-ORMML in handling such ordinal problems with cross-distribution data.

- Even the data distribution information is not provided, through employing the existing clustering approach to cluster the distributions, the *CD-ORMML (Distribution-Unknown)* still shares the best average estimation accuracy with the *CD-ORMML (Distribution-Known)* which is performed directly on the given distributions. It demonstrates the effectiveness and superiority of the proposed methods in handling the cross-distribution data scenarios.

5.4. Influence of Margin Scale on Metric Performance

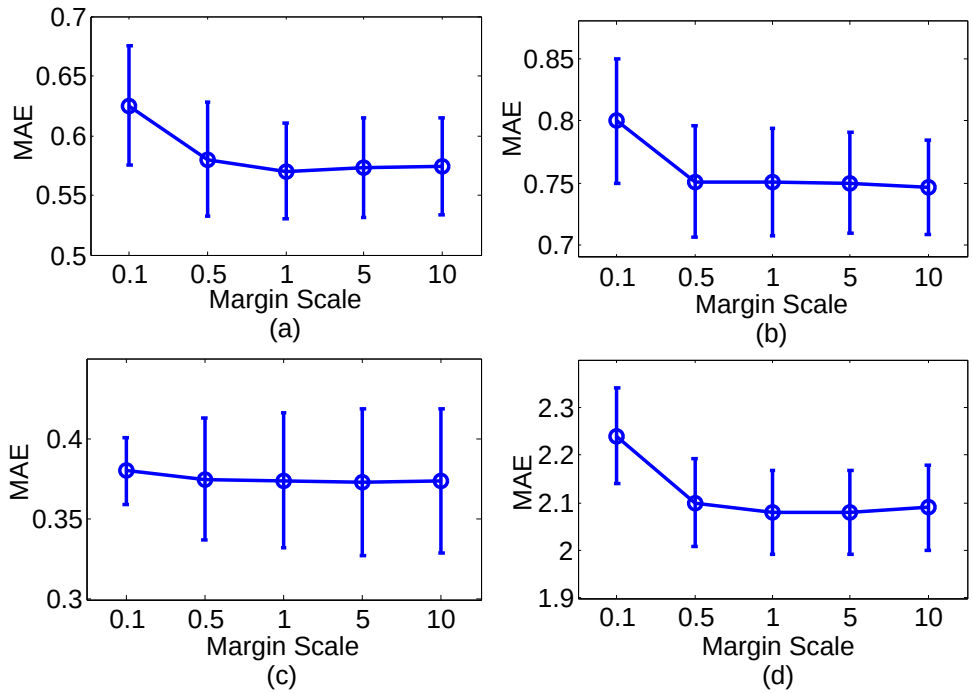


Figure 8: Influence of margin scale η to ORMML on FG-NET (a), Morph Album II (b), UMIST (c) and Mixed-gambles (d).

From (5b) and (14b), it can be found that the margins separating two classes are scaled by the hyper-parameter η. To explore its influence on the metric and without loss of generality, we conduct experiments by randomly selecting 16 samples for each age group from the FG-NET, Morph Album II, 12 samples from the UMIST, and 10 samples from the Mixed-gambles, respectively, and plot the results in Figure 8. From the four subfigures in Figure 8, we can see that when $\eta \geq 1$, the MAE generally begins to level off with increasing value of η. And the influence of η on CD-ORMML is quite similar to this. Therefore, it indicates that both ORMML and CD-ORMML are insensitive to the margin scale and thus we can set η to a fixed value, e.g., 1 in the experiments.

5.5. Convergence Evaluation

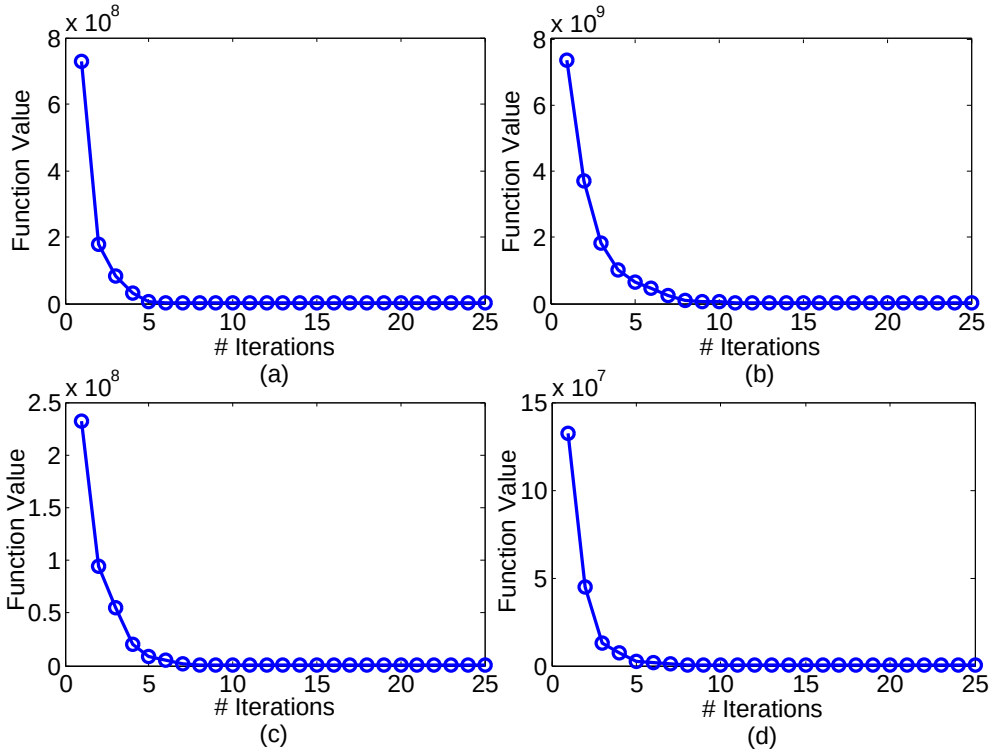


Figure 9: Convergence of ORMML algorithm on FG-NET (a), Morph Album II (b), UMIST (c) and Mixed-gambles (d).

We also validate the convergence of ORMML and CD-ORMML algorithms, summarized in Tables 2 and 3, respectively. Specifically, we take the same experimental settings as in Section 5.4, and demonstrate the results in Figure 9. It can be observed that after about 25 iterations the ORMML algorithm converges to the global optimum. And the convergence speed of CD-ORMML algorithm in the experiments is similar to ORMML.

6. Conclusion

In this work, we proposed a novel type of ordinal metric learning method, coined as ORMML, by seeking a sequence of margins to make the ordinal data distribute orderly in the metric space. Then, to cope with more realistic cross-distribution data scenarios, we developed a cross-distribution variant of ORMML, named CD-ORMML, by maximizing the correlation between the within-class distributions and formulating it as a regularization term in the objective function. Finally, extensive experiments demonstrated the effectiveness and superiority of the proposed methods on ordinal image data estimation, no matter the data are sampled across distributions or not. Although the ordinal characteristic of the data is well considered in the proposed metric learning methods, some more characteristics

behind the data are not considered yet, such as the spatial structure information of images and the type of sample representation space. Therefore, in the future we will extend our methods to tensor learning [10], [24], by representing the samples in advanced representation space like Banach space [20].

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grants 61472186, 61402215 and 61300154, the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20133218110032, the Funding of Jiangsu Innovation Program for Graduate Education under Grant *CXLX*13_159, and the Fundamental Research Funds for the Central Universities and Jiangsu *Qing-Lan Project*.

References

[1] Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D., 2005. Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research 6, 937–965.

[2] Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S., 2007. Information-theoretic metric learning, in: Proceedings of the 24th International Conference on Machine Learning, ACM. pp. 209–216.

[3] Denoeux, T., 1995. A k-nearest neighbor classification rule based on dempster-shafer theory. Systems, Man and Cybernetics, IEEE Transactions on 25, 804–813.

[4] Ding, C., He, X., 2004. K-means clustering via principal component analysis, in: Proceedings of the 21th International Conference on Machine Learning, ACM. p. 29.

[5] Fang, C., Rockmore, D.N., 2015. Multi-task metric learning on network data, in: Advances in Knowledge Discovery and Data Mining. Springer, pp. 317–329.

[6] Fouad, S., Tino, P., 2013. Ordinal-based metric learning for learning using privileged information, in: Neural Networks (IJCNN), The 2013 International Joint Conference on, pp. 1–8.

[7] Globerson, A., Roweis, S.T., 2005. Metric learning by collapsing classes, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 451–458.

[8] Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R., 2004. Neighbourhood components analysis, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 513–520.

[9] Guo, G., Mu, G., Fu, Y., Huang, T.S., 2009. Human age estimation using bio-inspired features, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 112–119.

[10] He, Q., Wu, C., 2011. Separating theorem of samples in banach space for support vector machine learning. International Journal of Machine Learning and Cybernetics 2, 49–54.

[11] Hearst, M.A., Dumais, S., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines. Intelligent Systems and their Applications, IEEE Transactions on 13, 18–28.

[12] Hong, Y., Li, Q., Jiang, J., Tu, Z., 2011. Learning a mixture of sparse distance metrics for classification and dimensionality reduction, in: Proceedings of the 2011 IEEE International Conference on Computer Vision, IEEE. pp. 906–913.

[13] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2000. Introduction to the logistic regression model. Wiley Online Library.

[14] Jacobs, R.A., 1988. Increased rates of convergence through learning rate adaptation. Neural Networks 1, 295–307.

[15] Jiang, N., Liu, W., Wu, Y., 2012. Order determination and sparsity-regularized metric learning adaptive visual tracking, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1956–1963.

[16] Jin, Y., Lu, J., Ruan, Q., 2015. Coupled discriminative feature learning for heterogeneous face recognition. Information Forensics and Security, IEEE Transactions on 10, 640–652.

[17] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 881–892.

[18] Lee, J.E., Jin, R., Jain, A.K., 2008. Rank-based distance metric learning: An application to image retrieval, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–8.

[19] Li, C., Liu, Q., Liu, J., Lu, H., 2015. Ordinal distance metric learning for image ranking. Neural Networks and Learning Systems, IEEE Transactions on 26, 1551–1559.

[20] Li, J., Han, G., Wen, J., Gao, X., 2011. Robust tensor subspace learning for anomaly detection. International Journal of Machine Learning and Cybernetics 2, 89–98.

[21] Liao, S., Hu, Y., Zhu, X., Li, S.Z., 2015. Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206.

[22] Lin, C.J., 2007. Projected gradient methods for nonnegative matrix factorization. Neural Computation 19, 2756–2779.

[23] Lu, J., Hu, J., Zhou, X., Shang, Y., Tan, Y.P., Wang, G., 2012. Neighborhood repulsed metric learning for kinship verification, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2594–2601.

[24] Luo, Y., Tao, D., Wen, Y., Ramamohanarao, K., Xu, C., 2015. Tensor canonical correlation analysis for multi-view dimension reduction. arXiv preprint arXiv:1502.02330 .

[25] Martinel, N., Micheloni, C., Foresti, G.L., 2015. Kernelized saliency-based person re-identification through multiple metric learning. Image Processing, IEEE Transactions on 24, 5645–5658.

[26] Mei, J., Liu, M., Karimi, H.R., Gao, H., 2014. Logdet divergence-based metric learning with triplet constraints and its applications. Image Processing, IEEE Transactions on 23, 4920–4931.

[27] Mensink, T., Verbeek, J., Perronnin, F., Csurka, G., 2012. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost, in: Computer Vision–ECCV 2012. Springer, pp. 488–501.
[28] Mignon, A., Jurie, F., 2012. Cmml: A new metric learning approach for cross modal matching, in: Asian Conference on Computer Vision, pp. 14–pages.
[29] Navarro, J.F., Frenk, C.S., White, S.D., 1997. A universal density profile from hierarchical clustering. The Astrophysical Journal 490, 493.
[30] Niu, G., Dai, B., Yamada, M., Sugiyama, M., 2014. Information-theoretic semi-supervised metric learning via entropy regularization. Neural Computation 26, 1717–1762.
[31] Ochoa, X., Duval, E., 2008. Relevance ranking metrics for learning objects. Learning Technologies, IEEE Transactions on 1, 34–48.
[32] Parameswaran, S., Weinberger, K.Q., 2010. Large margin multi-task metric learning, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 1867–1875.
[33] Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B., 2013. Local fisher discriminant analysis for pedestrian re-identification, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3318–3325.
[34] Qi, G.J., Tang, J., Zha, Z.J., Chua, T.S., Zhang, H.J., 2009. An efficient sparse metric learning in high-dimensional space via l 1-penalized log-determinant regularization, in: Proceedings of the 26th International Conference on Machine Learning, ACM. pp. 841–848.
[35] Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. Systems, Man and Cybernetics, IEEE Transactions on 21, 660–674.
[36] Schultz, M., Joachims, T., 2004. Learning a distance metric from relative comparisons. Proceedings of the Advances in Neural Information Processing Systems , 41.
[37] Tao, D., Jin, L., Wang, Y., Li, X., 2015. Person reidentification by minimum classification error-based kiss metric learning. Cybernetics, IEEE Transactions on 45, 242–252.
[38] Tarlow, D., Swersky, K., Charlin, L., Sutskever, I., Zemel, R., 2013. Stochastic k-neighborhood selection for supervised and unsupervised learning, in: Proceedings of the 30th International Conference on Machine Learning, pp. 199–207.
[39] Tom, S.M., Fox, C.R., Christopher, T., Poldrack, R.A., 2007. The neural basis of loss aversion in decision-making under risk. Science 315, 515–518.
[40] Wan, S., Aggarwal, J., 2013. Spontaneous facial expression recognition: A robust metric learning approach. Pattern Recognition 47, 1859–1868.
[41] Wang, J., Gao, X., Wang, Q., Li, Y., 2012. Prodis-contshc: learning protein dissimilarity measures and hierarchical context coherently for protein-protein comparison in protein database retrieval. BMC bioinformatics 13, S2.
[42] Wang, Y., Lin, X., Wu, L., Zhang, W., Zhang, Q., 2015a. Lbmch: Learning bridging mapping for cross-modal hashing, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM. pp. 999–1002.
[43] Wang, Y., Zhang, W., Wu, L., Lin, X., Zhao, X., 2015b. Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. Neural Networks and Learning Systems, IEEE Transactions on .
[44] Weinberger, K.Q., Blitzer, J., Saul, L.K., 2005. Distance metric learning for large margin nearest neighbor classification, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 1473–1480.
[45] Xiao, B., Yang, X., Xu, Y., Zha, H., 2009. Learning distance metric for regression by semidefinite programming with application to human age estimation, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM. pp. 451–460.
[46] Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y., 2002. Distance metric learning with application to clustering with side-information, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 505–512.
[47] Yang, P., Huang, K., 2013. Geometry preserving multi-task metric learning. Machine learning 92, 133–175.
[48] Yang, P., Huang, K., Liu, C.L., 2011. Multi-task low-rank metric learning based on common subspace, in: Neural Information Processing, Springer. pp. 151–159.
[49] Yeung, D.Y., Chang, H., 2006. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints. Pattern Recognition 39, 1007–1010.
[50] Yu, J., Tao, D., Li, J., Cheng, J., 2014. Semantic preserving distance metric learning and applications. Information Sciences 281, 674–686.
[51] Yu, J., Wang, M., Tao, D., 2012. Semisupervised multiview distance metric learning for cartoon synthesis. Image Processing, IEEE Transactions on 21, 4636–4648.
[52] Zhen, Y., Rai, P., Zha, H., Carin, L., 2015. Cross-modal similarity learning via pairs, preferences, and active supervision, in: Twenty-Ninth AAAI Conference on Artificial Intelligence.
[53] Zheng, W.S., Gong, S., Xiang, T., 2013. Reidentification by relative distance comparison. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 653–668.
[54] Zhu, P., Hu, Q., Zuo, W., Yang, M., 2014. Multi-granularity distance metric learning via neighborhood granule margin maximization. Information Sciences 282, 321–331.

Table 4: Comparison of cross-distribution age estimations (MAE±STD) generated in different metric spaces. Along each row, the results in bold are the best ones. Here "Dis.1" and "Dis.2" denote the two data distributions within each age group, respectively.

Dis.1	Dis.2	#Training samples (each age group, each distribution)	LFDA	ITML	LMNN	LDMLT	OITML	mkNN	LDMLR	ORMML	CD-ORMML (Distribution-Unknown)	CD-ORMML (Distribution-Known)
WM	WF	6	1.00±0.08	1.16±0.09	1.26±0.05	**0.92±0.03**	1.08±0.07	1.13±0.06	1.41±0.10	0.98±0.05	**0.93±0.04**	**0.93±0.05**
WM	WF	8	0.94±0.04	1.13±0.07	1.21±0.08	**0.91±0.03**	1.02±0.05	1.16±0.04	1.46±0.22	0.97±0.06	**0.91±0.03**	**0.90±0.03**
WM	WF	10	0.95±0.05	1.04±0.07	1.12±0.06	0.90±0.05	0.97±0.06	1.11±0.04	1.26±0.09	0.94±0.07	**0.89±0.04**	**0.88±0.04**
WM	BM	6	1.04±0.08	1.15±0.08	1.25±0.07	**0.95±0.05**	1.03±0.05	1.15±0.05	1.35±0.09	1.02±0.06	**0.95±0.04**	**0.94±0.05**
WM	BM	8	0.96±0.03	1.15±0.07	1.22±0.09	**0.93±0.03**	1.02±0.06	1.18±0.03	1.50±0.11	1.01±0.07	**0.94±0.04**	**0.93±0.04**
WM	BM	10	0.96±0.03	1.08±0.06	1.16±0.06	**0.91±0.04**	0.98±0.05	1.16±0.05	1.23±0.05	1.01±0.07	0.93±0.04	0.93±0.04
WM	BF	6	1.02±0.13	1.17±0.09	1.25±0.09	**0.90±0.03**	1.02±0.08	1.12±0.06	1.34±0.10	0.98±0.07	0.94±0.06	0.94±0.05
WM	BF	8	0.92±0.05	1.09±0.09	1.21±0.08	**0.87±0.05**	1.01±0.06	1.15±0.04	1.47±0.17	0.97±0.09	0.92±0.06	0.92±0.04
WM	BF	10	0.90±0.04	1.06±0.08	1.12±0.08	**0.85±0.04**	0.95±0.06	1.13±0.06	1.23±0.11	0.92±0.08	0.88±0.04	0.89±0.05
WF	BM	6	1.08±0.06	1.28±0.11	1.37±0.12	1.04±0.07	1.08±0.12	1.21±0.06	1.48±0.12	1.11±0.08	**1.02±0.08**	**1.00±0.08**
WF	BM	8	1.05±0.07	1.21±0.07	1.33±0.08	1.03±0.08	1.03±0.05	1.21±0.03	1.43±0.20	1.08±0.10	1.00±0.05	**0.98±0.06**
WF	BM	10	1.06±0.07	1.20±0.09	1.27±0.03	1.04±0.08	1.03±0.07	1.19±0.05	1.36±0.09	1.06±0.07	**0.99±0.05**	**0.98±0.06**
WF	BF	6	1.10±0.06	1.23±0.09	1.27±0.10	**0.99±0.05**	1.01±0.06	1.11±0.06	1.37±0.11	1.11±0.05	1.02±0.05	1.02±0.06
WF	BF	8	1.04±0.04	1.14±0.08	1.23±0.10	0.99±0.05	0.98±0.09	1.15±0.03	1.38±0.17	1.05±0.06	1.01±0.05	**0.97±0.06**
WF	BF	10	1.04±0.05	1.11±0.07	1.17±0.08	0.99±0.05	0.96±0.06	1.17±0.05	1.26±0.11	1.03±0.05	**0.95±0.05**	0.97±0.05
BM	BF	6	1.11±0.05	1.11±0.10	1.24±0.07	1.00±0.05	0.97±0.08	1.21±0.06	1.34±0.13	1.05±0.06	**0.96±0.06**	**0.95±0.06**
BM	BF	8	1.05±0.05	1.13±0.10	1.25±0.08	0.96±0.05	0.95±0.07	1.26±0.04	1.40±0.13	1.03±0.08	**0.93±0.06**	**0.93±0.06**
BM	BF	10	1.03±0.04	1.09±0.09	1.17±0.08	0.95±0.04	0.95±0.08	1.21±0.04	1.27±0.12	0.99±0.08	**0.91±0.05**	0.93±0.08
Average Performance (Ranking)			1.01±0.06(5)	1.14±0.08(7)	1.23±0.08(9)	0.95±0.05(3)	1.00±0.07(4)	1.17±0.05(8)	1.36±0.12(10)	1.02±0.07(6)	**0.94±0.05(1)**	**0.94±0.05(1)**