

# Less but Better: Generalization Enhancement of Ordinal Embedding via Distributional Margin

Ke Ma,<sup>1,2</sup> Qianqian Xu,<sup>3</sup> Zhiyong Yang,<sup>1,2</sup> Xiaochun Cao<sup>1,2\*</sup>

<sup>1</sup>State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences

<sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences

<sup>3</sup>Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences  
{make, yangzhiyong, caoxiaochun}@iie.ac.cn, xuqianqian@ict.ac.cn

## Abstract

In the absence of prior knowledge, ordinal embedding methods obtain new representation for items in a low-dimensional Euclidean space via a set of quadruple-wise comparisons. These ordinal comparisons often come from human annotators, and sufficient comparisons induce the success of classical approaches. However, collecting a large number of labeled data is known as a hard task, and most of the existing work pay little attention to the generalization ability with insufficient samples. Meanwhile, recent progress in large margin theory discloses that rather than just maximizing the minimum margin, both the margin mean and variance, which characterize the margin distribution, are more crucial to the overall generalization performance. To address the issue of insufficient training samples, we propose a margin distribution learning paradigm for ordinal embedding, entitled Distributional Margin based Ordinal Embedding (*DMOE*). Precisely, we first define the margin for ordinal embedding problem. Secondly, we formulate a concise objective function which avoids maximizing margin mean and minimizing margin variance directly but exhibits the similar effect. Moreover, an Augmented Lagrange Multiplier based algorithm is customized to seek the optimal solution of *DMOE* effectively. Experimental studies on both simulated and real-world datasets are provided to show the effectiveness of the proposed algorithm.

The problem of analyzing a set of  $n$  objects given *similarity* information is an inherent part in a broad variety of tasks in artificial intelligence (Heikinheimo and Ukkonen 2013; Wilber, Kwak, and Belongie 2014), machine learning (Jamieson and Nowak 2011; Kleindessner and Luxburg 2014; Arias-Castro 2017; Kleindessner and von Luxburg 2017), information retrieval (Liu et al. 2016), data mining (Le and Lauw 2016) and computer vision (Wilber et al. 2015). Many algorithms are based on the assumption that ‘similar’ inputs should generate ‘close’ outputs. In a numerical setting of embedding, a similarity function (or, equivalently, a dissimilarity function) quantifies how ‘similar’ objects are to others. The required input is the distance or similarity matrix of items. We calculate a set of embedded points which aims to preserve such similarities as well as possible. However, in recent years a whole new branch of the liter-

ature has emerged, which is the comparison-based embedding. Instead of evaluating similarity directly, we collect the similarity comparisons as follows:

“Is the similarity between object  $i$  and  $j$  larger than the similarity between  $l$  and  $k$ ?”

The corresponding problem is ordinal embedding. These two types of supervision information, numerical similarities and relative comparisons, are all generated by human beings. Nevertheless, the latter one provides similarity estimates on a relative scale instead of the absolute scale. The comparison-based setting is a special case of the observation that humans are better at comparing two stimuli than at identifying a single one (Stewart, Brown, and Chater 2005). Consequently, the relative comparison is a more reliable form for incorporating human knowledge with artificial intelligence tasks.

The ordinal embedding problem was firstly studied by (Shepard 1962a; 1962b; Kruskal 1964a; 1964b) in the psychometric society. In recent years, it has drawn a lot of attention (Schultz and Joachims 2003; Agarwal et al. 2007; Tamuz et al. 2011; van der Maaten and Weinberger 2012; Terada and Luxburg 2014; Amid and Ukkonen 2015; Heim et al. 2015; Jain, Jamieson, and Nowak 2016; Ma et al. 2018). One class of these typical methods is margin-based ordinal embedding which solves the problem under the classification framework. The well-known Generalized Non-Metric Multidimensional Scaling (*GNMDS*) (Agarwal et al. 2007) aims at finding a Gram matrix  $G$  such that the pairwise distances of embedded points satisfy the partial order constraints. Stochastic Triplet Embedding (*STE/TSTE*) (van der Maaten and Weinberger 2012) is proposed to jointly penalize the violated constraints and reward the satisfied constraints via logistic loss. Multi-view Triplet Embedding (*MTE*) (Amid and Ukkonen 2015) decomposes the *STE* objective function as different components and re-weights them for a better explanation. The other class of ordinal embedding methods uses the nearest neighbor graphs to model the similarity comparisons. Structure Preserving Embedding (*SPE*) (Shaw and Jebara 2009) and Local Ordinal Embedding (*LOE*) (Terada and Luxburg 2014) embed unweighted nearest neighbor graphs to Euclidean spaces with convex and non-convex objective functions. The nearest neighbor adjacency matrix can be transformed into ordinal constraints, but it is not a standard equipment in comparison-

\*The corresponding authors.

based scenarios. With this limitation, *SPE* and *LOE* are not suitable for ordinal embedding via quadruplets or triple comparisons.

A common issue of the existed ordinal embedding methods is the dependence of large samples of similarity comparisons. (Kleindessner and Luxburg 2014; Arias-Castro 2017) show the consistency of ordinal embedding problem. When the number of the objects  $n$  tends to infinity, the set of embedded points always converges to the set of original points, up to similarity transformations; the rate of convergence depends on the Hausdorff distance between the ground-truth points. Later (Jain, Jamieson, and Nowak 2016) show a finite sampling result of consistency. Learning an embedding which predicts nearly as well as the true embedding needs  $\Theta(pn \log n)$  samples, where  $p$  is the embedding dimension. There is a strong condition that the triple-wise comparisons are generated from the classical Bradley-Terry-Luce (*BTL*) model (Bradley and Terry 1952; Luce 1959) and this assumption could not be verified in the actual applications. The theoretical results suggest that only the adequateness of similarity comparisons can promise the prediction result. However, the cost of eliciting relative similarity comparisons from human beings would be prohibitive. The amenable applications for collecting the relative similarity comparisons, e.g., crowdsourcing and human computation, need passively waiting for participants and stimulate them with money to get the desired information. Without prior knowledge, the relative comparisons always involve all objects, and the number of possible comparisons could be  $\Theta(n^4)$ . The spending of data collection presents ordinal embedding methods with a dilemma: the insufficient samples would limit the potential performance; the adequate samples with prospective results would be cumbersome. Unfortunately, most of the traditional methods ignore that the generalization is the main concern in ordinal embedding task with insufficient samples.

In this paper, we propose a new method, named Distributional Margin based Ordinal Embedding (*DMOE*), which tries to achieve strong generalization performance by optimizing the margin distribution in ordinal embedding problem. Inspired by the recent results in classification (Zhang and Zhou 2014; Yang, Lu, and Zhang 2016), we define the margin of ordinal embedding and characterize the margin distribution by the first- and second-order statistics, and try to maximize the margin mean and minimize the margin variance simultaneously. For optimization, we propose an alternating direction method of multipliers (*ADMM*) for *DMOE* with semi-definite and low-rank constraints. Comprehensive experiments on the synthetic and real-world datasets show the superiority of our method to other ordinal embedding algorithms, verifying that the margin distribution is more crucial for generalization than minimum margin.

## Problem Definition

Throughout the paper, scalars, vectors, matrices and sets are denoted as lowercase letters ( $x$ ), bold lower case letters ( $\mathbf{x}$ ), bold capital letters ( $\mathbf{X}$ ) and calligraphy uppercase letters ( $\mathcal{X}$ ).  $x_{ij}$  denotes the  $(i, j)$  entry of  $\mathbf{X}$ .  $[n]$  is the set of  $\{1, \dots, n\}$ .  $\mathbb{E}(\cdot)$  represents the expectation.

Suppose  $\mathcal{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_n\}$  is a set of  $n$  objects, we assume that a certain but unknown similarity function  $\zeta : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}^+$  assigns similarity value  $\zeta_{ij}$  for a pair of objects  $(\mathbf{o}_i, \mathbf{o}_j)$ . With similarity function  $\zeta$ , a quadruplet  $q = (i, j, l, k)$  defines the corresponding ordinal constraint, and these constraints lead to the ordinal embedding problem.

**Definition 1** (Ordinal Constraints). *Given a set of quadruplets*

$$\mathcal{Q} = \{q \mid q = (i, j, l, k), (i, j) \neq (l, k), \\ i \neq j, l \neq k, i, j, l, k \in [n]\} \quad (1)$$

which is a subset of  $[n]^4$ , the ordinal constraints  $\mathcal{Y}_{\mathcal{Q}} = \{y_q \mid q \in \mathcal{Q}\} \subset \{-1, +1\}^{|\mathcal{Q}|}$ , implies the similarity partial order of object pairs in  $\mathcal{O}$  as

$$y_q = \begin{cases} +1, & \text{if } \zeta_{ij} < \zeta_{lk}, \\ -1, & \text{if } \zeta_{ij} > \zeta_{lk}. \end{cases} \quad (2)$$

Our goal here is to obtain a set of embedded points  $\mathbf{X}$  which satisfy the ordinal constraints  $\mathcal{Q}$ . Without prior knowledge, embedding  $\mathcal{O}$  into a Euclidean space  $\mathbb{R}^p$  is the most common situation which assumes that the squared Euclidean distances among embedded points are inversely proportional to the unknown similarity values. Specifically, a large distance of two embedded points  $d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$  means the corresponding objects  $\mathbf{o}_i$  and  $\mathbf{o}_j$  would have small similarity value  $\zeta_{ij}$ . This assumption connects the squared Euclidean distances of  $\mathbf{X}$  and ordinal constraints  $\mathcal{Q}$ . We further give the formal definition of ordinal embedding.

**Definition 2** (Ordinal Embedding). *Suppose  $\mathcal{Q}$  is a collection of quadruplets which are drawn independently and uniformly at random and  $\mathcal{Y}_{\mathcal{Q}}$  is the correspondence ordinal constraints of object set  $\mathcal{O}$ . Let  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathbb{R}^{p \times n}$  is the desired embedding in the Euclidean space  $\mathbb{R}^p$  where  $p \ll n$  and  $\mathbf{D} = \{d_{ij}^2\} \in \mathbb{R}^{n \times n}$  is the squared Euclidean distance matrix of embedding  $\mathbf{X}$ . Ordinal embedding is the problem of obtaining  $\mathbf{X}$  with ordinal constraints  $\mathcal{Y}_{\mathcal{Q}}$  on  $\mathbf{D}$  such that*

$$\text{sign}(y_q \cdot \Delta_q \mathbf{D}) > 0, \forall y_q \in \mathcal{Y}_{\mathcal{Q}},$$

where

$$\Delta_q \mathbf{D} = d_{ij}^2 - d_{lk}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 - \|\mathbf{x}_l - \mathbf{x}_k\|_2^2.$$

Note that one cannot consistently estimate the underlying embedding  $\mathbf{X}$  with only ordinal supervision and without direct observations. In the case when no direct measurements are available, say the metric information of  $\mathcal{O}$  as the input, the underlying embedding  $\mathbf{X}$  is only identifiable up to certain monotonic transformations, e.g., rotation, reflection, translation, and scaling. Therefore, the sign consistency is adopted as the goal of ordinal embedding.

By the above definition, the margin of instance  $(\mathbf{X}_q, y_q)$  can be naturally defined as

$$\gamma_q = y_q \cdot \Delta_q \mathbf{D}, \quad (3)$$

where  $\mathbf{X}_q = \{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l, \mathbf{x}_k\}$ ,  $\forall q \in \mathcal{Q}$ .

Despite the close relationship between  $\mathbf{D}$  and  $\mathbf{X}$ ,  $\Delta_q \mathbf{D}$  is a nonlinear function of  $\mathbf{X}$  and it always leads to a non-convex optimization problem. Here we introduce the Gram

matrix of  $\mathbf{X}$  and construct a margin function as a linear function of Gram matrix. Firstly, a map is established to connect the distance matrix  $\mathbf{D}$  and the Gram matrix  $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$ :

$$\begin{aligned} d_{ij} &= g_{ii} - 2g_{ij} + g_{jj} \\ \mathbf{D} &= \text{diag}(\mathbf{G}) \cdot \mathbf{1}^\top - 2\mathbf{G} + \mathbf{1} \cdot \text{diag}(\mathbf{G})^\top, \end{aligned}$$

where  $\text{diag}(\mathbf{G})$  is the column vector composed of the diagonal entries of  $\mathbf{G}$  and  $\mathbf{1}$  is the  $n$ -dimension vector with all entries being 1. With a little abuse of  $\Delta_q$ , the margin of instance  $(\mathbf{X}_q, y_q)$  can be written as

$$\begin{aligned} \gamma_q &= y_q \cdot \Delta_q \mathbf{D} = y_q (d_{ij} - d_{lk}) \\ &= y_q (g_{ii} - 2g_{ij} + g_{jj} - g_{ll} + 2g_{lk} - g_{kk}) \quad (4) \\ &:= y_q \cdot \Delta_q \mathbf{G}, \quad \forall q \in \mathcal{Q}. \end{aligned}$$

By the definition of (4), ordinal embedding can be formulated as the following convex optimization problem.

**Definition 3** (The Margin-based Ordinal Embedding). *Let  $l: \mathbb{R}^+ \rightarrow \mathbb{R}$  be a loss function which satisfies*

$$l(x) : \begin{cases} > 0, & \text{if } x > 0, \\ \leq 0, & \text{if } x < 0. \end{cases} \quad (5)$$

Given the ordinal constraints  $\mathcal{Y}_{\mathcal{Q}}$ , the ordinal embedding problem can be formulated as a semi-definite programming of Gram matrix  $\mathbf{G}$ :

$$\begin{aligned} \min_{\mathbf{G}} \quad & L(\mathbf{G}, \mathcal{Y}_{\mathcal{Q}}) \\ \text{s.t.} \quad & \mathbf{G} \succeq 0, \quad \text{rank}(\mathbf{G}) \leq p, \end{aligned} \quad (6)$$

where  $L(\mathbf{G}, \mathcal{Y}_{\mathcal{Q}}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} l(\gamma_q)$ .

We note that (6) is a semi-definite programming (SDP) and  $\mathbf{G} \succeq 0$  comes from the fact that  $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$  is a positive semi-definite matrix. Furthermore, the desired embedding dimension  $p$  is a parameter of the ordinal embedding. It is well known that there exists a perfect embedding  $\mathbf{X}$  estimated by any label set  $\mathcal{Y}$  on the Euclidean distances in  $\mathbb{R}^{n-2}$ , even for the noisy constraints (Borg and Groenen 2003). However, the low-dimensional setting where  $p \ll n$  is the main task of this work. The smallest  $p$  for noisy ordinal constraints  $\mathcal{Y}_{\mathcal{Q}}$  is a future direction which worths pursuing. The choice of  $p$  in the experiment section depends on the potential applications.

For example, the Generalized Non-metric Multidimensional Scaling (GNMDS) follows the SVM formulation to obtain the  $\mathbf{G}$  by solving

$$\begin{aligned} \min_{\mathbf{G}, \boldsymbol{\xi}} \quad & \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \xi_q \\ \text{s.t.} \quad & \gamma_q \geq \gamma_0 - \xi_q, \quad \xi_q \geq 0, \quad \forall q \in \mathcal{Q}, \\ & \mathbf{G} \succeq 0, \quad \text{rank}(\mathbf{G}) \leq p, \end{aligned} \quad (7)$$

where  $\gamma_0$  is a relaxed minimum margin and  $\boldsymbol{\xi} = \{\xi_q\}_{q \in \mathcal{Q}}$  is the slack variable.

## Distributional Margin based Embedding

The relaxed minimum margin  $\gamma_0$  in GNMDS indeed characterizes the top minimum margins of all instance  $\{\mathbf{X}_q, y_q\}_{q \in \mathcal{Q}}$ . In margin theory of classification, it is known that maximizing the minimum margin of training examples is not sufficient to achieve fulfilling generalization performance (Reyzin and Schapire 2006). The margin distribution of training examples, rather than the minimum margin, is more crucial to generalization performance in classification (Gao and Zhou 2013; Zhang and Zhou 2016).

### Formulation

The two most usual statistics for characterizing the margin distribution are the first- and second-order statistics, that is, the mean and the variance of margin. According to (4), the margin mean of training samples  $\{(\mathbf{X}_q, y_q)\}$  is

$$\bar{\gamma} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \gamma_q \quad (8)$$

and the margin variance is

$$\hat{\gamma} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} (\gamma_q - \bar{\gamma})^2. \quad (9)$$

Intuitively, we attempt to maximize the margin mean and minimize the margin variance simultaneously in ordinal embedding problem (6).

First, there is a straightforward idea to achieve our goal as considering the margin mean (8) and the margin variance (9) in (6) explicitly. Although (6) can adopt different loss functions, we will focus on SVM formulation (7) because the hinge loss is a natural form of margin. Considering the margin distribution, the optimization problem (7) can be formulated as

$$\begin{aligned} \min_{\mathbf{G}, \boldsymbol{\xi}} \quad & \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \xi_q - \lambda_1 \bar{\gamma} + \lambda_2 \hat{\gamma} \\ \text{s.t.} \quad & \gamma_q \geq \gamma_0 - \xi_q, \quad \xi_q \geq 0, \quad \forall q \in \mathcal{Q}, \\ & \mathbf{G} \succeq 0, \quad \text{rank}(\mathbf{G}) \leq p, \end{aligned} \quad (10)$$

where  $\lambda_1$  and  $\lambda_2$  are the trade-off parameters for balancing the impacts of  $\bar{\gamma}$  and  $\hat{\gamma}$ . It is apparent that GNMDS (7) is a degenerate case of (10) when  $\lambda_1$  and  $\lambda_2$  equal to 0. However, there exists an obvious drawback of (10) with directly optimizing the margin distribution: tuning the parameters,  $\lambda_1$  and  $\lambda_2$ , is an obstacles of solving (10) efficiently. Therefore, a new lightweight formulation is proposed to optimize margin distribution implicitly.

Recall that SVM fixes the minimum margin as 1 by scaling the margin with the norm of linear predictor. Following the similar way, we can scale the margin of  $(\mathbf{X}_q, y_q)$  in ordinal embedding (4) and set the margin mean as a constant. This would not result in a sub-optimal solution because the ordinal constraints  $\mathcal{Y}_{\mathcal{Q}}$  can only determine an embedding  $\mathbf{X}$  up to the monotonic transformations. Without loss of generality, the mean of  $\gamma_{\mathcal{Q}} = \{\gamma_q | q \in \mathcal{Q}\}$  can be set as a constant and an equality constraint is conducted

$$\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \gamma_q = \tilde{\gamma}_0. \quad (11)$$

On the other hand, we want to minimize the variance of  $\gamma_{\mathcal{Q}}$ . By (11), the deviation of  $\gamma_q$  to the margin mean  $\tilde{\gamma}_0$  is  $|\gamma_q - \tilde{\gamma}_0|$ , and we force the deviation to be smaller than  $\varepsilon_q \geq 0$  as

$$|\gamma_q - \tilde{\gamma}_0| \leq \varepsilon_q, \forall q \in \mathcal{Q}. \quad (12)$$

Thus, minimizing  $\varepsilon_q$  is equivalent to minimize the margin variance (9). Meanwhile, (12) implies the margin mean constraint (11).

Constraints like (12) in optimization problems always involve two inequality,  $\forall q \in \mathcal{Q}$ ,

$$\gamma_q \leq \tilde{\gamma}_0 + \varepsilon_q, \quad (13a)$$

$$\gamma_q \geq \tilde{\gamma}_0 - \varepsilon_q. \quad (13b)$$

Note that the soft-margin constraint

$$\gamma_q \geq \tilde{\gamma}_0 - \xi_q \quad (14)$$

plays the same role as (13b). Replacing (13b) with (14) and adding (13) into (10), we arrive at the following formulation

$$\begin{aligned} \min_{\mathbf{G}, \xi, \varepsilon} \quad & \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \xi_q + \nu \cdot \varepsilon_q \\ \text{s.t.} \quad & \gamma_q \geq \tilde{\gamma}_0 - \xi_q, \gamma_q \leq \tilde{\gamma}_0 + \varepsilon_q, \\ & \xi_q \geq 0, \varepsilon_q \geq 0, \forall q \in \mathcal{Q}, \\ & \mathbf{G} \succeq 0, \text{rank}(\mathbf{G}) \leq p. \end{aligned} \quad (15)$$

This optimization problem corresponds to dealing with such a loss function,  $\forall q \in \mathcal{Q}$

$$\ell_{\tilde{\gamma}_0, \nu}(\gamma_q) = \max(\tilde{\gamma}_0 - \gamma_q, 0) + \nu \cdot \max(\gamma_q - \tilde{\gamma}_0, 0). \quad (16)$$

The trading-off parameter  $\nu$  in (15) can capture the asymmetry between the sign correctness and the dispersion of  $\{\gamma_q\}_{q \in \mathcal{Q}}$ . When  $\nu = 1$  and ignoring the semi-definite and rank constraints, (15) is similar to the support vector regression (SVR) (Drucker et al. 1997). In SVR, the  $\tau$ -insensitive loss

$$|\zeta|_{\tau} = \begin{cases} 0, & \text{if } |\zeta| \leq \tau, \\ |\zeta| - \tau, & \text{otherwise} \end{cases} \quad (17)$$

produces two slack variables  $\xi$  and  $\xi^*$  for each training example to guard against outliers. As  $\nu = 1$ , the loss function (16) is explicitly the same loss function adopted by SVR as  $\tau = 0$ . In our formulation (15),  $\xi_q$  and  $\varepsilon_q$  conduct the similar constraints like SVR but have totally different meanings. All the training examples are used to learn the margin distribution in (15), but the optimal solution of SVR is only spanned by the support vectors which is sparse in the training data. Figure 1 depicts the situations of the learned margin distribution graphically. Some theoretical results are provided at the end of this section.

**Theorem 1.** *Suppose that*

$$\mathcal{G}_{\mu} = \{\mathbf{G} \in \mathbb{S}_+^n : \|\mathbf{G}\|_{\infty} \leq \mu, \|\mathbf{G}\|_* \leq \lambda\}$$

*and the true Gram matrix  $\mathbf{G}^* \in \mathcal{G}_{\mu}$ . Let  $\hat{\mathbf{G}}$  be a solution of (15). With probability at least  $1 - \delta$ , it holds that*

$$\begin{aligned} & R(\hat{\mathbf{G}}) - R(\mathbf{G}^*) \\ & \leq 4\mu(1 + \nu) \sqrt{\frac{32np \log n}{|\mathcal{Q}|}} + 8\mu(1 + \nu) \sqrt{\frac{2 \log \frac{2}{\delta}}{|\mathcal{Q}|}}, \end{aligned} \quad (18)$$

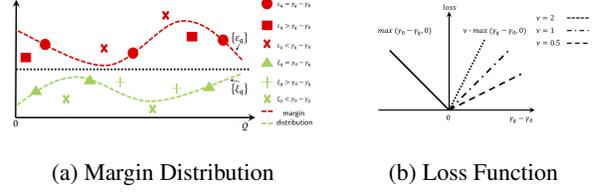


Figure 1: (a) The Margin Distribution obtained by solving (15). (b) The loss function (16) with different  $\nu$ .

where  $R(\cdot)$  is the risk, as for any  $\mathbf{G} \in \mathcal{G}_{\mu}$

$$R(\mathbf{G}) = \mathbb{E}[\ell(\gamma_q)] = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} p_q \ell(\gamma_q) + (1 - p_q) \ell(-\gamma_q),$$

$p_q = \mathbb{P}(y_q = 1)$ . Here the expectation respects to both the uniformly random selection of the quadruplet  $q$  and its label  $y_q$ .

Theorem 1 says that  $|\mathcal{Q}|$  must scale like  $\Theta(pn \log n)$  which leads to the bounded error  $R(\hat{\mathbf{G}}) - R(\mathbf{G}^*)$ . This result is consistent with known finite sample bounds (Jamieson and Nowak 2011). The details are provided in the supplementary materials. The generalization bound of margin distribution is a future direction.

## Optimization

Consider that  $\Delta_q$  is a linear operator on  $\mathbb{S}_+^n$ ,  $\Delta_q$  has its symmetric  $n \times n$  matrix form in  $\mathbb{S}_+^n$ , the positive semi-definite cone of  $n \times n$  symmetric matrix. Given a ordinal constraints  $q = (i, j, l, k)$ ,  $\mathbf{K}_q$  is the matrix form of  $\Delta_q$  where

$$\mathbf{K}_q = \begin{matrix} & i & j & l & k \\ \begin{matrix} i \\ j \\ l \\ k \end{matrix} & \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \end{matrix} \quad (19)$$

and

$$\Delta_q \mathbf{G} = \langle \mathbf{K}_q, \mathbf{G} \rangle = \text{tr}(\mathbf{K}_q \mathbf{G}) = \text{vec}(\mathbf{K}_q)^\top \text{vec}(\mathbf{G}). \quad (20)$$

With the trick that

$$\tilde{\gamma}_0 - \gamma_q = y_q y_q \tilde{\gamma}_0 - y_q \Delta_q \mathbf{G} = y_q (y_q \tilde{\gamma}_0 - \Delta_q \mathbf{G}), \quad (21)$$

we note

$$\begin{aligned} \mathbf{e}_{\mathcal{Q}} &= [e_1, \dots, e_{|\mathcal{Q}|}]^\top \\ &= \mathbf{y}_{\mathcal{Q}} \odot \mathbf{\Gamma}_0 - \begin{pmatrix} \vdots \\ \text{vec}(\mathbf{K}_q)^\top \\ \vdots \end{pmatrix}_{|\mathcal{Q}| \times n^2} \cdot \text{vec}(\mathbf{G}) \quad (22) \\ &\triangleq \mathbf{y}_{\mathcal{Q}} \odot \mathbf{\Gamma}_0 - \mathcal{K} \cdot \text{vec}(\mathbf{G}), \end{aligned}$$

where  $\mathbf{y}_{\mathcal{Q}} \in \{-1, +1\}^{|\mathcal{Q}|} = [y_1, \dots, y_{|\mathcal{Q}|}]^\top$ ,  $\mathbf{\Gamma}_0$  is a  $|\mathcal{Q}|$ -dimension vector with all entries are  $\tilde{\gamma}_0$  and  $\odot$  is the Hadamard product. Furthermore, we introduce the redundant variables to make the objective separable which can be

solved by the ALM framework efficiently. The optimization is converted into:

$$\begin{aligned} \min_{\mathcal{T}} \quad & \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_1)_+\|_p^p + \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_2)_+\|_p^p + \lambda \|\mathbf{G}_1\|_* \\ \text{s.t.} \quad & \mathbf{G} = \mathbf{G}_1, \mathbf{G} = \mathbf{G}_2, \mathbf{G}_2 \succeq 0, \\ & \mathbf{e}_1 = \mathbf{e}_{\mathcal{Q}}, \mathbf{e}_2 = -\mathbf{e}_{\mathcal{Q}}, \end{aligned} \quad (23)$$

where  $\mathcal{T} = \{\mathbf{G}, \mathbf{G}_1, \mathbf{G}_2, \mathbf{e}_1, \mathbf{e}_2\}$  is the set of all the parameters to be solved and  $\|\cdot\|_*$  is the nuclear norm which is the convex surrogate of matrix rank constraints. It is worth mentioning that (23) is a convex optimization problem as the feasible set of each constraint is a convex set and the objective function is convex. The Lagrange function of (23) can be written in the following form:

$$\begin{aligned} \mathcal{L}(\mathcal{T}) = & \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_1)_+\| + \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_2)_+\| \\ & + \Phi(\mathbf{z}_1, \mathbf{e}_1 - \mathbf{e}_{\mathcal{Q}}) + \Phi(\mathbf{z}_2, \mathbf{e}_2 + \mathbf{e}_{\mathcal{Q}}) \\ & + \lambda \|\mathbf{G}_1\|_* + \Phi(\mathbf{Z}_3, \mathbf{G} - \mathbf{G}_1) \\ & + \delta(\mathbf{G}_2 \succeq 0) + \Phi(\mathbf{Z}_4, \mathbf{G} - \mathbf{G}_2), \end{aligned} \quad (24)$$

with

$$\Phi(\mathbf{u}, \mathbf{v}) = \frac{\mu}{2} \|\mathbf{v}\|^2 + \langle \mathbf{u}, \mathbf{v} \rangle, \mu > 0.$$

$\|\cdot\|$  is  $\ell_2$  norm for vector and the Frobenius norm for matrix. In addition,  $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{|\mathcal{Q}|}$  and  $\mathbf{Z}_3, \mathbf{Z}_4 \in \mathbb{R}^{n \times n}$  are Lagrange multipliers.  $\delta$  is the Dirac delta function whose function value would be infinity if the condition is not satisfied. Below are the solutions to each sub-problem.

**$\mathbf{e}_1$  sub-problem.** With the variables unrelated to  $\mathbf{e}_1$  fixed, we have the sub-problem of  $\mathbf{e}_1$ :

$$\mathbf{e}_1^{(t+1)} = \arg \min_{\mathbf{e}_1} \frac{1}{\mu^{(t)}} \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_1)_+\|_p^p + \frac{1}{2} \|\mathbf{e}_1 - \mathbf{s}_1^{(t)}\|_2^2, \quad (25)$$

where

$$\mathbf{s}_1^{(t)} = \mathbf{e}_{\mathcal{Q}}^{(t)} - \frac{\mathbf{z}_1^{(t)}}{\mu^{(t)}}.$$

It's worth noting that  $(\cdot)_+$  is a piece-wise linear function. Thus, to seek the minimum of each element in  $\mathbf{e}_1$ , we just need to pick the smaller value between  $y_q e_q^1$  and 0. The solution of (25) is

$$\mathbf{e}_1^{(t+1)} = \Omega \left( \mathcal{S}_{\frac{1}{\mu^{(t)}}}[\mathbf{s}_1^{(t)}] \right) + \bar{\Omega} \left( \mathbf{s}_1^{(t)} \right), \quad (26)$$

where  $\Omega: \mathbb{R}^{\mathcal{Q}} \rightarrow \mathbb{R}^{\mathcal{Q}}$  is an indicator function as  $[\Omega(\mathbf{w})]_q = \mathbb{I}(w_q > 0) \cdot w_q$ ,  $\mathbf{w} \in \mathbb{R}^{\mathcal{Q}}$  and  $\bar{\Omega}$  is the complementary support of  $\Omega$ . The definition of shrinkage operator on scalars is  $\mathcal{S}_{\tau>0}[u] = \text{sign}(u)(|u| - \tau)_+$  and it is an element-wise operator for vector and matrix.

**$\mathbf{e}_2$  sub-problem.** Similarly, picking out the terms related to  $\mathbf{e}_2$  gives the following sub-problem:

$$\mathbf{e}_2^{(t+1)} = \arg \min_{\mathbf{e}_2} \frac{1}{\mu^{(t)}} \|(\mathbf{y}_{\mathcal{Q}} \odot \mathbf{e}_2)_+\|_p^p + \frac{1}{2} \|\mathbf{e}_2 - \mathbf{s}_2^{(t)}\|_2^2, \quad (27)$$

where

$$\mathbf{s}_2^{(t)} = -\mathbf{e}_{\mathcal{Q}}^{(t)} - \frac{\mathbf{z}_2^{(t)}}{\mu^{(t)}},$$

and the solution of  $\mathbf{e}_2$  sub-problem is just replaced the  $\mathbf{s}_2^{(t)}$  with  $\mathbf{s}_1^{(t)}$  in (26).

**$\mathbf{G}$  sub-problem.** Dropping the terms independent on  $\mathbf{G}$  leads to the following problem:

$$\arg \min_{\mathbf{G}} \left( \begin{aligned} & \Phi(\mathbf{z}_1^{(t)}, \mathbf{e}_1^{(t)} - \mathbf{e}_{\mathcal{Q}}^{(t)}) + \Phi(\mathbf{z}_2^{(t)}, \mathbf{e}_2^{(t)} + \mathbf{e}_{\mathcal{Q}}^{(t)}) \\ & + \Phi(\mathbf{Z}_3^{(t)}, \mathbf{G} - \mathbf{G}_1^{(t)}) + \Phi(\mathbf{Z}_4^{(t)}, \mathbf{G} - \mathbf{G}_2^{(t)}) \end{aligned} \right). \quad (28)$$

We have

$$\begin{aligned} & 2(\mathcal{K}^\top \mathcal{K} + \mathbf{I}) \cdot \text{vec}(\mathbf{G})^{(t+1)} \\ & = \mathcal{K}^\top \left( \mathbf{e}_1^{(t)} + \frac{1}{\mu^{(t)}} \mathbf{z}_1^{(t)} - \mathbf{e}_2^{(t)} - \frac{1}{\mu^{(t)}} \mathbf{z}_2^{(t)} \right) \\ & + \text{vec} \left( \mathbf{G}_1^{(t)} - \frac{1}{\mu^{(t)}} \mathbf{Z}_3^{(t)} + \mathbf{G}_2^{(t)} - \frac{1}{\mu^{(t)}} \mathbf{Z}_4^{(t)} \right), \end{aligned} \quad (29)$$

and note the right-hand side as  $\mathbf{w}$ , we have

$$\text{vec}(\mathbf{G})^{(t+1)} = \frac{1}{2} (\mathcal{K}^\top \mathcal{K} + \mathbf{I})^{-1} \mathbf{w},$$

and  $\mathbf{G}^{(t+1)}$  is the matrix form of  $\text{vec}(\mathbf{G})^{(t+1)}$ .

**$\mathbf{G}_1$  sub-problem.** There are two terms in (24) involving  $\mathbf{G}_1$ . The associated optimization problem of  $\mathbf{G}_1$  is

$$\mathbf{G}_1^{(t+1)} = \arg \min_{\mathbf{G}_1} \lambda \|\mathbf{G}_1\|_* + \Phi(\mathbf{Z}_3, \mathbf{G} - \mathbf{G}_1), \quad (30)$$

and solving this problem yields

$$\mathbf{G}_1^{(t+1)} = \mathcal{U} \mathcal{S}_{\frac{\lambda}{\mu^{(t)}}}(\boldsymbol{\Sigma}) \mathbf{V}^\top, \quad (31)$$

where

$$\mathbf{G}_1^{(t)} + \frac{1}{\mu^{(t)}} \mathbf{Z}_3^{(t)} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top$$

and  $\mathcal{S}_\tau(\cdot)$  is the shrinkage operator.

**$\mathbf{G}_2$  sub-problem.** Considering the potential asymmetric of  $\mathbf{G}^{(t)}$ , we claim that  $\mathbf{G}_2$  is the nearest symmetric positive semi-definite matrix of  $\mathbf{G}^{(t)}$  in Frobenius norm (Higham 1988). By the following theorem, we show the explicit solution of  $\mathbf{G}_2^{(t)}$ .

**Theorem 2.** Suppose that  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , and let  $\mathbf{B} = (\mathbf{A} + \mathbf{A}^\top)/2$ ,  $\mathbf{C} = (\mathbf{A} - \mathbf{A}^\top)/2$  be the symmetric and skew-symmetric parts of  $\mathbf{A}$  respectively. If we do polar decomposition of  $\mathbf{B}$  as  $\mathbf{B} = \mathbf{U} \mathbf{H}$  where  $\mathbf{U}$  is orthogonal matrix  $\mathbf{U} \mathbf{U}^\top = \mathbf{I}$  and  $\mathbf{H}$  is positive semi-definite matrix,  $\mathbf{X}_F = (\mathbf{B} + \mathbf{H})/2$  is the unique approximation of  $\mathbf{A}$  in the Frobenius norm with positive semi-definite constraint, and the distance  $\rho_F(\mathbf{A})$  in the Frobenius norm from  $\mathbf{A}$  to  $\mathbb{S}_+^n$  is

$$\rho_F^2(\mathbf{A}) = \sum_{\sigma_i(\mathbf{B}) < 0} \sigma_i^2(\mathbf{B}) + \|\mathbf{C}\|_F^2,$$

where  $\sigma_i(\mathbf{B})$ ,  $i = 1, \dots, n$  is the eigenvalue of  $\mathbf{B}$ .

Consequently, the explicit solution of  $\mathbf{G}_2^{(t)}$  is

$$\begin{aligned} \mathbf{G}_2^{(t)} = & \frac{1}{4} \left( \mathbf{G}^{(t)} + \frac{1}{\mu} \mathbf{Z}_4 + \mathbf{G}^{(t)\top} + \frac{1}{\mu} \mathbf{Z}_4^\top \right) \\ & + \frac{1}{4} \sqrt{\left( \mathbf{G}^{(t)} + \frac{1}{\mu} \mathbf{Z}_4 \right)^\top \left( \mathbf{G}^{(t)} + \frac{1}{\mu} \mathbf{Z}_4 \right)}, \end{aligned} \quad (32)$$

where  $\sqrt{A}$  is the square root of  $A \in \mathbb{S}_+^n$ ,  $A = VSV^{-1}$  and  $\sqrt{A} = VS^{\frac{1}{2}}V^{-1}$ ,  $S$  is a diagonal matrix and  $S^{\frac{1}{2}}$  is element-wise square root of  $S$ .

---

**Algorithm 1:** The ADMM method for solving (23)

---

**Input:**  $\mathcal{K}, \mathbf{y}_Q, \gamma_0, \nu, \lambda$

**Output:**  $\mathbf{G}$

1 Initialize  $\mathbf{G}_0, \mu, \mathbf{z}_1^{(0)} = \mathbf{z}_2^{(0)} = \mathbf{0}_{|Q|}, t := 1,$   
 $\mathbf{Z}_1^{(0)} = \mathbf{Z}_2^{(0)} = \mathbf{0}_{n \times n}$ , calculate  $\mathbf{e}_Q$  via (22);

2 **while** *Not Converged* **do**

3     update  $\mathbf{e}_1, \mathbf{e}_2$  by solving (25) and (27);

4     update  $\mathbf{G}$  by solving (28);

5     update  $\mathbf{G}_1, \mathbf{G}_2$  via (31) and (32);

6     calculate  $\mathbf{e}_Q$  via (22);

7     update multipliers and  $\mu$  as

$$\begin{aligned} \mathbf{z}_1^{(t+1)} &= \mathbf{z}_1^{(t)} + \mu^{(t)}(\mathbf{e}_1^{(t)} - \mathbf{e}_Q) \\ \mathbf{z}_2^{(t+1)} &= \mathbf{z}_2^{(t)} + \mu^{(t)}(\mathbf{e}_2^{(t)} + \mathbf{e}_Q) \\ \mathbf{Z}_1^{(t+1)} &= \mathbf{Z}_1^{(t)} + \mu^{(t)}(\mathbf{G}^{(t)} - \mathbf{G}_1^{(t)}) \\ \mathbf{Z}_2^{(t+1)} &= \mathbf{Z}_2^{(t)} + \mu^{(t)}(\mathbf{G}^{(t)} - \mathbf{G}_2^{(t)}) \\ \mu^{(t+1)} &= \rho\mu^{(t)}, \rho > 1 \end{aligned} \quad (33)$$

$t := t + 1;$

8 **end**

---

For clarity, the procedure of solving (23) is outlined in Algorithm 1. The algorithm would not be terminated until the change of objective value in two successive iterations is smaller than a threshold (in the experiments, 0.001 is the default setting).

## Empirical Study

In this section, we show the results of simulations and real-world data experiments to demonstrate the effectiveness of the proposed algorithms. As the existed margin-based ordinal embedding methods, such as GNMDS, STE and TSTE, just use triple-wise comparisons as the ordinal constraints, we treat triple-wise comparisons as the input of the proposed algorithm for fair competition. The triple-wise comparisons  $\mathcal{T} = \{(i, j, k)\}$  is a special case of quadruplets which means  $l = i$  in  $q = (i, j, l, k) \in Q$ . The  $\Delta_p$  of triplet  $t = (i, j, k)$  is also a symmetric  $n \times n$  matrix indicated by  $(i, j, k)$  as

$$\mathbf{K}_t = \begin{matrix} & i & j & k \\ \begin{matrix} i \\ j \\ k \end{matrix} & \begin{pmatrix} 0 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \end{matrix}. \quad (34)$$

Replacing  $\mathbf{K}_q$  with  $\mathbf{K}_t$  in those sub-problems in Algorithm 1, the proposed *DMOE* method could handle the triplets set  $\mathcal{T}$  as the ordinal constraints. The reproducible code can be found here<sup>1</sup>.

<sup>1</sup><https://github.com/alphaprime/DMOE>

## Simulation

**Settings.** The synthesized dataset consists of 100 points  $\{\mathbf{x}_i\}_{i=1}^{100} \subset \mathbb{R}^{10}$ , where  $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{20}\mathbf{I})$ ,  $\mathbf{I} \in \mathbb{R}^{10 \times 10}$  is the identity matrix. The possible similarity triple comparisons are generated based on the Euclidean distances between  $\{\mathbf{x}_i\}$ . We randomly sample  $|\mathcal{T}| = \{200, 500, 1,000, 10,000\}$  triplets as the training set and the test set is the rest of all triplets. The embedding dimension is fixed to 10.

**Evaluation Metrics.** We employ the generalization error to evaluate generalization ability of various algorithms. As the learned Gram matrix  $\mathbf{G}$  from partial triple comparisons set  $\mathcal{T} \subset [n]^3$  may be generalized to unknown triplets, the percentage of held-out triplets which is not satisfied in the  $\mathbf{G}$  is the generalization error of the learned embedding.

**Competitors.** We compare the proposed algorithm with three well-known ordinal embedding methods: GNMDS (Agarwal et al. 2007), STE and TSTE (van der Maaten and Weinberger 2012). Note that we adopt the optimization strategy proposed by (Jain, Jamieson, and Nowak 2016), which performs gradient descent with line search, and projects the Gram matrix onto the subspace spanned by the top  $p$  eigenvalues at each step (i.e. setting the smallest  $n-p$  eigenvalues to 0). We call the three competitors: GNMDS- $p$ , STE- $p$  and TSTE- $p$ , correspondingly. The optimization problem of GNMDS is (7). STE replaces the hinge loss by logistic loss in (7) and adopts Gaussian kernel to predict the label:

$$l_{\text{ste}}(\mathbf{G}^{(t)}, y_p) = \log(1 + \exp(\gamma_p^{(t)})),$$

where  $\gamma_p^{(t)} = \Delta_p \mathbf{G}^{(t)}$ . TSTE employs the heavy-tailed Student-t kernel:

$$l_{\text{tste}}(\mathbf{G}^{(t)}, y_p) = \log \left( 1 + \left( \gamma_p^{(t)} \right)^{-\frac{\alpha+1}{2}} \right).$$

The regularization parameters of the competitors are tuned for the best performance under the different settings.

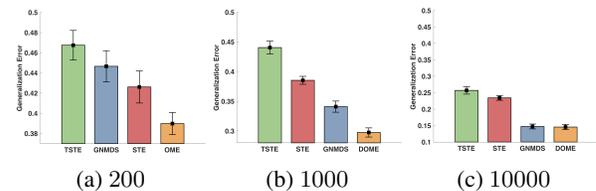


Figure 2: Generalization errors of DMOE, GNMDS- $p$ , STE- $p$  and TSTE- $p$  on the synthetic dataset.

**Results.** From Figure 2 and Table 1, the following phenomena can be observed. First of all, the generalization ability of all methods would be improved when the number of training samples increases. The decrease of standard derivation also improves the stability. Moreover, the proposed algorithm shows better generalization performance than the traditional methods in all four settings. Compared with GNMDS- $p$ /STE- $p$ /TSTE- $p$  which need more training samples, our method can achieve better results with fewer training samples. This is our main motivation to optimize

Table 1: Performance Comparison on synthetic dataset with 200, 1000 and 10000 samples as training data, respectively

algorithm	(a) 200 samples				(b) 1000 samples				(c) 10000 samples			
	min	median	max	std	min	median	max	std	min	median	max	std
GNMDS- $p$	0.419	0.447	0.476	0.016	0.318	0.341	0.359	0.009	0.143	0.147	0.154	<b>0.007</b>
STE- $p$	0.397	0.426	0.461	0.016	0.375	0.385	0.401	<b>0.007</b>	0.219	0.234	0.251	<b>0.007</b>
TSTE- $p$	0.440	0.468	0.498	0.014	0.426	0.441	0.466	0.011	0.238	0.257	0.271	0.011
DMOE	<b>0.372</b>	<b>0.390</b>	<b>0.410</b>	<b>0.011</b>	<b>0.281</b>	<b>0.298</b>	<b>0.305</b>	0.008	<b>0.142</b>	<b>0.146</b>	<b>0.151</b>	0.008

Table 2: Performance Comparison on music artists dataset with 200, 500, 1000 and 5000 samples as training data.

algorithm	(a) 200 samples				(b) 1000 samples				(c) 5000 samples			
	min	median	max	std	min	median	max	std	min	median	max	std
GNMDS- $p$	0.391	0.403	0.416	0.007	0.307	0.317	0.332	0.007	0.225	0.239	0.257	<b>0.006</b>
STE- $p$	0.444	0.455	0.475	0.008	0.397	0.415	0.429	0.007	0.252	0.275	0.294	0.011
TSTE- $p$	0.416	0.436	0.458	0.011	0.377	0.389	0.406	0.007	0.243	0.259	0.297	0.013
DMOE	<b>0.372</b>	<b>0.385</b>	<b>0.400</b>	<b>0.007</b>	<b>0.281</b>	<b>0.291</b>	<b>0.307</b>	0.007	<b>0.216</b>	<b>0.227</b>	<b>0.244</b>	0.007

the margin distribution instead of maximizing the minimum margin like the classic methods. Third, the results of GNMDS- $p$  verifies that only maximizing the minimum margin would not necessarily lead to better generalization performances as the STE- $p$  is better than GNMDS when train samples are few.

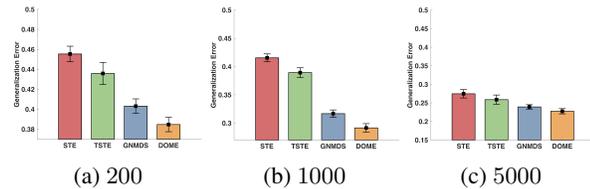


Figure 3: Generalization errors of DMOE, GNMDS, STE and TSTE on the music artists dataset.

## Music Artist Data

**Settings.** The music artist data is collected by (Ellis et al. 2002) via a web-based survey in which 1,032 users provided 213,472 triplets on the similarity of 412 music artists. We use the data pre-processed by (van der Maaten and Weinberger 2012) which includes only 9,107 triplets for  $n = 400$  artists. The size of training samples is variant from 200 to 5,000 and the rest of triplets are treated as test set. The desired dimension of embedding is  $d = 9$  as these music artists can be classified by genre into 9 categories.

**Results.** According to the experimental results, Figure 3 and Table 2, we have the following observations. *DMOE* still shows better prediction result than GNMDS- $p$ /STE- $p$ /TSTE- $p$  with the same number of noisy training samples. To achieve the same generalization error, *DMOE* needs the smallest number of training samples and STE- $p$ /TSTE- $p$  need five times more than *DMOE*. This real-world data experiment verifies the proposed method, *DMOE*, has strong generalization for ordinal embedding with small training

samples. Although this dataset contains noise triplets and it is well-known that the calculation of mean and the variance is sensitive, the proposed method show the same magnitude of standard deviation and its results are not damaged by the potential wrong training samples. The robustness is still an open problem in ordinal embedding, and this is our future work.

## Conclusion

The classical ordinal embedding algorithms always need a large number of labeled data to predict unknown similarity relationship among items from learned embedded points. As collecting high-quality, large-scale labeled data from human is a hard task, generalization ability is the main challenge when we could only access small numbers of relative comparisons. Incorporating margin distribution learning paradigm gives birth to a novel algorithm for ordinal embedding, namely *DMOE*. Comprehensive experiments on synthetic dataset and real-world dataset validate the superiority of our method to traditional methods which need more training data to achieve the same generalization.

## Acknowledgment

The research of Ke Ma and Xiaochun Cao is supported by the National Key R&D Program of China (Grant No. 2016YFB0800603), the Key Program of the Chinese Academy of Sciences (No. QYZDB-SSW-JSC003) and the National Natural Science Foundation of China (No.U1636214, U1605252, 61733007). The research of Qianqian Xu is supported in part by the National Natural Science Foundation of China (No.61672514, 61390514, 61572042), the Beijing Natural Science Foundation (4182079), the Youth Innovation Promotion Association CAS, and the CCF-Tencent Open Research Fund.

## References

- Agarwal, S.; Wills, J.; Cayton, L.; Lanckriet, G. R.; Kriegman, D. J.; and Belongie, S. 2007. Generalized non-metric multidimensional scaling. *International Conference on Artificial Intelligence and Statistics* 11–18.
- Amid, E., and Ukkonen, A. 2015. Multiview triplet embedding: Learning attributes in multiple maps. *International Conference on Machine Learning* 1472–1480.
- Arias-Castro, E. 2017. Some theory for ordinal embedding. *Bernoulli* 23(3):1663–1693.
- Borg, I., and Groenen, P. 2003. Modern multidimensional scaling: theory and applications. *Journal of Educational Measurement* 40(3):277–280.
- Bradley, R. A., and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4):324–345.
- Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A. J.; and Vapnik, V. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems*, 155–161.
- Ellis, D. P.; Whitman, B.; Berenzweig, A.; and Lawrence, S. 2002. The quest for ground truth in musical artist similarity. *International Society for Music Information Retrieval Conference*.
- Gao, W., and Zhou, Z. 2013. On the doubt about margin explanation of boosting. *Artificial Intelligence* 203:1–18.
- Heikinheimo, H., and Ukkonen, A. 2013. The crowdmedian algorithm. *AAAI Conference on Human Computation and Crowdsourcing* 69–77.
- Heim, E.; Berger, M.; Seversky, L. M.; and Hauskrecht, M. 2015. Efficient online relative comparison kernel learning. *SIAM International Conference on Data Mining (SDM)* 271–279.
- Higham, N. J. 1988. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications* 103:103 – 118.
- Jain, L.; Jamieson, K. G.; and Nowak, R. D. 2016. Finite sample prediction and recovery bounds for ordinal embedding. In *Annual Conference on Neural Information Processing Systems*, 2703–2711.
- Jamieson, K., and Nowak, R. 2011. Low-dimensional embedding using adaptively selected ordinal data. *Annual Allerton Conference on Communication, Control, and Computing* 1077–1084.
- Kleindessner, M., and Luxburg, U. 2014. Uniqueness of ordinal embedding. *Conference on Learning Theory* 40–67.
- Kleindessner, M., and von Luxburg, U. 2017. Kernel functions based on triplet comparisons. In *Annual Conference on Neural Information Processing Systems*, 6810–6820.
- Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1):1–27.
- Kruskal, J. B. 1964b. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29(2):115–129.
- Le, D. D., and Lauw, H. W. 2016. Euclidean co-embedding of ordinal data for multi-type visualization. In *SIAM International Conference on Data Mining*, 396–404.
- Liu, H.; Ji, R.; Wu, Y.; and Liu, W. 2016. Towards optimal binary code learning via ordinal embedding. In *AAAI Conference on Artificial Intelligence*, 1258–1265.
- Luce, R. D. 1959. *Individual Choice Behavior*. John Wiley.
- Ma, K.; Zeng, J.; Xiong, J.; Xu, Q.; Cao, X.; Liu, W.; and Yao, Y. 2018. Stochastic non-convex ordinal embedding with stabilized barzilai-borwein step size. In *AAAI Conference on Artificial Intelligence*, 3738–3745.
- Reyzin, L., and Schapire, R. E. 2006. How boosting the margin can also boost classifier complexity. In *International Conference Machine Learning*, 753–760.
- Schultz, M., and Joachims, T. 2003. Learning a distance metric from relative comparisons. In *Annual Conference on Neural Information Processing Systems*, 41–48.
- Shaw, B., and Jebara, T. 2009. Structure preserving embedding. *International Conference on Machine Learning* 937–944.
- Shepard, R. N. 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika* 27(2):125–140.
- Shepard, R. N. 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika* 27(3):219–246.
- Stewart, N.; Brown, G. D.; and Chater, N. 2005. Absolute identification by relative judgment. *Psychological Review* 112(4):881.
- Tamuz, O.; Liu, C.; Shamir, O.; Kalai, A.; and Belongie, S. 2011. Adaptively learning the crowd kernel. *International Conference on Machine Learning* 673–680.
- Terada, Y., and Luxburg, U. 2014. Local ordinal embedding. *International Conference on Machine Learning* 847–855.
- van der Maaten, L., and Weinberger, K. 2012. Stochastic triplet embedding. *IEEE International Workshop on Machine Learning for Signal Processing* 1–6.
- Wilber, M.; Kwak, I.; Kriegman, D.; and Belongie, S. 2015. Learning concept embeddings with combined human-machine expertise. *IEEE International Conference on Computer Vision* 981–989.
- Wilber, M.; Kwak, S.; and Belongie, S. 2014. Cost-effective hits for relative similarity comparisons. *AAAI Conference on Human Computation and Crowdsourcing* 227–233.
- Yang, Z.; Lu, J.; and Zhang, T. 2016. Extreme large margin distribution machine and its applications for biomedical datasets. In *IEEE International Conference on Bioinformatics and Biomedicine*, 1549–1554.
- Zhang, T., and Zhou, Z. 2014. Large margin distribution machine. In *ACM International Conference on Knowledge Discovery and Data Mining*, 313–322.
- Zhang, T., and Zhou, Z. 2016. Optimal margin distribution machine. *CoRR* abs/1604.03348.