Robust Ordinal Embedding from Contaminated Relative Comparisons

Ke Ma[1,2], Qianqian Xu[3], Xiaochun Cao[1]*

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
{make, caoxiaochun}@iie.ac.cn, xuqianqian@ict.ac.cn

Abstract

Existing ordinal embedding methods usually follow a two-stage routine: outlier detection is first employed to pick out the inconsistent comparisons; then an embedding is learned from the clean data. However, learning in a multi-stage manner is well-known to suffer from sub-optimal solutions. In this paper, we propose a unified framework to jointly identify the contaminated comparisons and derive reliable embeddings. The merits of our method are three-fold: (1) By virtue of the proposed unified framework, the sub-optimality of traditional methods is largely alleviated; (2) The proposed method is aware of global inconsistency by minimizing a corresponding cost, while traditional methods only involve local inconsistency; (3) Instead of considering the nuclear norm heuristics, we adopt an exact solution for rank equality constraint. Our studies are supported by experiments with both simulated examples and real-world data. The proposed framework provides us a promising tool for robust ordinal embedding from the contaminated comparisons.

Introduction

The solutions to many tasks involve the similarity estimation of data samples, *e.g.* clustering (von Luxburg 2007), classification (Chen et al. 2009) and representation learning (Wei et al. 2018). Traditionally, the objective similarity measurement extracts the features of data points to calculate their distances in some space, while the Euclidean distance, cosine similarity, Hamming distance and Kullback-Leibler divergence are the favorite measurements. In real-world scenarios, the situation becomes complicated since the off-the-shelf similarity functions are unreliable. It is too difficult to customize a similarity function for the specific data, or simply unavailable of the features or attributes. Consequently, the subjective method is more appropriate when the objective criteria are ambiguous. Among various subjective approaches for similarity estimation, relative comparison is expected to yield more reliable results. Instead of evaluating the similarity on an absolute scale, the relative similarity only needs individuals to answer a "yes or no" question as: *"Is the similarity between object i and j larger than the similarity between l and k?"* Then a point configuration is constructed to preserve the above constraints as much as possible. This task is known as the ordinal embedding. The problem first arises in the psychometric society (Shepard 1962a; 1962b; Kruskal 1964a; 1964b). In recent years, it has gained increasingly attention. (Agarwal et al. 2007; Tamuz et al. 2011; Jamieson and Nowak 2011; van der Maaten and Weinberger 2012; Kleindessner and Luxburg 2014; Terada and Luxburg 2014; Amid and Ukkonen 2015; Jain, Jamieson, and Nowak 2016; Ma et al. 2018)

However, the relative comparison approach leaves a cumbersome burden on participants with a large number of annotations. Due to its economical and scalable implementation, crowdsourcing platforms (*e.g.*, MTurk, Innocentive, CrowdFlower, CrowdRank, and Allourideas) are always restored to annotate these relative comparisons. Nevertheless, crowdsourced relative comparisons are not without pitfalls – the crowd is not all trustworthy (Xu et al. 2012; Chen et al. 2013). Since participants perform experiments without supervision on the Internet, when the testing time for a single participant lasts too long, the annotators may give untrustworthy feedbacks. Such unreliable labels bring great challenges to control the quality of crowsourced relative comparisons, let alone utilizing them in the subsequent tasks. Therefore, how to derive reliable embeddings from these contaminated data has become an urgent issue in the ordinal embedding research.

Traditional robust ordinal embedding methods are usually two-staged: (1) The outlier detection is employed to pick out the inconsistent comparisons. (2) Based on the cleaned data, the embedding is constructed to map items into a Euclidean space with a low dimension. Various methods have been developed in literature for outlier detection, such as majority voting (Welinder et al. 2010), M-estimator (Huber and Ronchetti 2009), Least Median of Squares (LMS) (Rousseeuw 1984), S-estimators (Rousseeuw and Yohai 1984), Least Trimmed Squares (LTS) (Rousseeuw and Leroy 2005), and Thresholding based Iterative Procedure for Outlier Detection (She and Owen 2011) etc. Among these studies, perhaps the most well-known one is majority voting. A large budget is allocated to obtain multiple annotations for each comparisons. These annotations are then aggregated so as to eliminate label noise (McFee and Lanckriet 2011). However, the effectiveness of the majority voting

*The corresponding authors.

strategy is often limited by the sparsity problem – it is typically infeasible to have many annotators for each relative comparisons. Moreover, it has been found that when pairwise local rankings are integrated into a global ranking, it is possible to detect outliers that can cause global inconsistency and yet are locally consistent, *i.e.*, supported by majority votes (Jiang et al. 2011). Worse, the existing ordinal embedding models, such as GNMDS (Agarwal et al. 2007), CKL (Tamuz et al. 2011) and STE (van der Maaten and Weinberger 2012; Jain, Jamieson, and Nowak 2016), only adopt the classification scheme via predicting the labels of the relative comparisons to construct the embeddings. The generalization of the embeddings would be damaged grievously when the labels of the training samples are wrong. As a consequence, separating as two unrelated parts, the outlier detection and the classification-based embedding would only obtain the local optimal solutions individually.

In this paper we propose a unified approach to detect outliers in contaminated data and derive robust ordinal embedding simultaneously. Specifically, instead of detecting outliers locally and independently for each comparison, our method works globally in a sense that the global inconsistency is explicitly penalized by the loss function. This enables us to identify those outliers which would be locally consistent with the majority results but in fact lead to significant global ranking inconsistency. The proposed model considers a partially penalized LASSO problem in the semi-definite programming. An efficient algorithm is proposed to obtain the embedding with exact desired dimension. The experiments are carried out on synthetic and real-world datasets. The results demonstrate that our method outperforms the state-of-the-art alternatives.

Robust Ordinal Embedding

Preliminaries

Let $\mathcal{O} = \{\boldsymbol{o}_1, \dots, \boldsymbol{o}_n\}$ be a set of objects which need to obtain the embedding. There exists an unknown but fixed similarity function $\zeta : \mathcal{O}^2 \to \mathbb{R}^+$ which assigns the similarity value ζ_{ij} to a pair of objects $(\boldsymbol{o}_i, \boldsymbol{o}_j)$. In this sense, the ranking of $\{\zeta_{ij}\},\ i, j \in [n]$ will produce a total order. However, without any prior knowledge, $\mathcal{O}$ and $\{\zeta_{ij}\}$ are both unknown. Therefore, we establish the form of side-information that will drive our ordinal embedding algorithm, that is, the relative similarity measurements collected from human labelers.

Given a set of objects $\mathcal{O}$ and a set of annotators $\mathcal{U}$, a collection of relative similarity measurements can be written as

$$\mathcal{C}_{\mathcal{U}} = \left\{ (i,j,l,k)_u \,\middle|\, \begin{array}{c} i,\ j,\ l,\ j \in [n],\ u \in \mathcal{U} \\ i \neq j,\ l \neq k,\ (i,j) \neq (l,k) \end{array} \right\}, \quad (1)$$

where a tuple $(i,j,l,k)_u$ is interpreted as "worker u annotates that i and j are more similar than l and k". (This measurement subsumes the triple-wise comparison situation when $i = l$.) The goal of ordinal embedding is to find an embedding $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \in \mathbb{R}^{p \times n}$ such that

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) < d(\boldsymbol{x}_l, \boldsymbol{x}_k),\ \forall\ (i,j,l,k)_u \in \mathcal{C}_{\mathcal{U}}, \quad (2)$$

where $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ is a distance function of Euclidean space $\mathbb{R}^p$. As both $\mathcal{O}$ and ζ lose the explicit form in ordinal embedding problem, the squared Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 := d_{ij}$ is always adopted in (2). We denote $\boldsymbol{D} = \{d_{ij}\}$ as the distance matrix of $\boldsymbol{X}$.

Despite the distance matrix $\boldsymbol{D}$ is directly related to the embedding $\boldsymbol{X}$, the squared Euclidean distance is a nonlinear function of $\boldsymbol{X}$. Here we introduce the Gram matrix of $\boldsymbol{X}$ and conduct the distance as a linear function of Gram matrix. It is known that there is a map between the distance matrix $\boldsymbol{D}$ and the Gram matrix $\boldsymbol{G} = \{g_{ij}\} = \boldsymbol{X}^\top \boldsymbol{X}$ as

$$d_{ij} = g_{ii} - 2g_{ij} + g_{jj}, \quad (3a)$$

$$\boldsymbol{D} = \mathit{diag}(\boldsymbol{G}) \cdot \mathbf{1}^\top - 2\boldsymbol{G} + \mathbf{1} \cdot \mathit{diag}(\boldsymbol{G})^\top, \quad (3b)$$

where $\mathit{diag}(\boldsymbol{G})$ is the column vector composed of the diagonal entries of $\boldsymbol{G}$ and $\mathbf{1}$ is the n-dimension vector whose all entries equal to 1.

To dissect the underlying geometrical structure of $\mathcal{C}_{\mathcal{U}}$, we represent $\mathcal{C}_{\mathcal{U}}$ as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ over $\mathcal{O}^2$ and note c_u as a ordered tuple $(i,j,l,k)_u$. Each vertex $v_{ij} \in \mathcal{V} \subseteq \mathcal{O}^2$ in the graph $\mathcal{G}$ corresponds to a pair $(\boldsymbol{o}_i, \boldsymbol{o}_j) \in \mathcal{O}^2$, and an edge $e_c^u \in \mathcal{E}$ from v_{ij} to v_{lk} corresponds to a relative similarity measurement labeled by worker $u \in \mathcal{U}$. The measurement between $(\boldsymbol{o}_i, \boldsymbol{o}_j)$ and $(\boldsymbol{o}_l, \boldsymbol{o}_k)$ will be labeled by different annotators and their answers to the same question would be inconsistent. It leads to the multiple edges with different directions between v_{ij} and v_{lk}. Let $\boldsymbol{e}_c^{\mathcal{U}} = \{e_c^u, u \in \mathcal{U},\ c \in \mathcal{C}_{\mathcal{U}}\}$ be the multiple edge with each single edge e_c^u has the same direction, and we assign an indicator y_c^u on e_c^u as $y_c^u = 1$ if e_c^u existed. It means that worker u measures $(\boldsymbol{o}_i, \boldsymbol{o}_j)$ and $(\boldsymbol{o}_l, \boldsymbol{o}_k)$, then she/he gives an answer which supposes that $\zeta_{ij} > \zeta_{lk}$. Notice that y_c^u is skew-symmetric, for each $u \in \mathcal{U}$, *i.e.*, $y_c^u = -y_{\bar{c}}^u$ where $c = (i,j,l,k)$ and $\bar{c} = (l,k,i,j)$. Furthermore, all the comparisons, from $(\boldsymbol{o}_i, \boldsymbol{o}_j)$ to $(\boldsymbol{o}_l, \boldsymbol{o}_k)$, are then aggregated over all annotators who have cast a vote on the two pairs. The results are represented as the weight of $\boldsymbol{e}_c^{\mathcal{U}}$, the total number of annotations on c,

$$w_c = \sum_{u \in \mathcal{U}} [y_c^u = 1],\ c \in \mathcal{C}_{\mathcal{U}}, \quad (4)$$

where $[\cdot]$ indicates the Iverson's bracket notation. We denote the whole edge weight $\boldsymbol{w} = \{w_c\} \in \mathbb{R}^{|\mathcal{E}|}$.

Existing Problems of Traditional Methods

Interpreting $\mathcal{C}_{\mathcal{U}}$ as the comparison graph $\mathcal{G}$ with multiple-edge $\{\boldsymbol{e}_c^{\mathcal{U}}\}$ will help us to infer global structure properties of $\mathcal{C}_{\mathcal{U}}$. In an ideal case, we know the similarity function ζ explicitly, and the global ranking of $\{\zeta_{ij}\}$ will leads the comparisons on $\{(\boldsymbol{o}_i, \boldsymbol{o}_j)\} \in \mathcal{O}^2$ to be a partial order. Two facts will become immediately apparent: 1) $\mathcal{G}$ is acyclic, and 2) the votes received on each edge are unanimous, *e.g.* $w_c \geq 1$ and $w_{\bar{c}} = 0$. However, it is always difficult to design such a function ζ to measure the similarity of objects in $\mathcal{O}$. That's why we need the wisdom of a crowd and measure the relative similarity by the comparisons $\mathcal{C}_{\mathcal{U}}$ from human beings. There always exists disagreement in $\mathcal{C}_{\mathcal{U}}$ as both $w_c > 0$ and $w_{\bar{c}} > 0$ would appear. Assuming these opposite annotations cannot be true concurrently, one of them will be the outlier

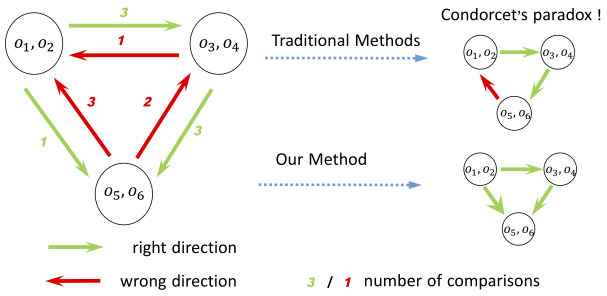


Figure 1: Traditional vs. Our methods in outlier detection. Green arrows/edges indicate correct annotations, while red arrows represent the outliers. The numbers indicate the number of votes received by each edge.

which should not be considered in the ordinal embedding problem. The traditional methods require a pretreatment of $\mathcal{C}_\mathcal{U}$. It is known as the *majority voting* method which treats one direction with larger weight as the majority and the other as the minority, then the latter one will be pruned. Obviously, *majority voting* is a "*local*" outlier detection method and it ignores the potential "*global*" similarity ranking on $\mathcal{O}^2$.

To make the matter worse, this local treatment would keep the wrong annotations and remove the true labels (see Figure 1.) Suppose that $\zeta_{12} > \zeta_{34} > \zeta_{56}$, *majority voting* will remove the edges from node v_{12} to node v_{56} as they are the minority versus the other direction edges. In particular, the *majority voting* will introduce a cyclic comparison $v_{12} > v_{34} > v_{56} > v_{12}$ which is the well-known **Condorcet's paradox** (Gehrlein 2006; Jiang et al. 2011), and consequently $\mathcal{G}$ will contains cycles. In order to eliminate the cycles in $\mathcal{G}$, the maximum acyclic subgraph are adopted (McFee and Lanckriet 2011; van der Maaten and Weinberger 2012) to replace $\mathcal{G}$.

Summarizing the arguments, we have the following comments on the traditional methods. For one thing, these methods could not utilize the ordinal information $\mathcal{C}_\mathcal{U}$ properly by adopting *majority voting* to outlier detection. For the other thing, an **NP-complete** problem, see maximum acyclic subgraph (Garey and Johnson 1979), makes the whole process more complicated. These are the main motivations for us to propose a new framework which can detect the outlier and obtain the embedding simultaneously.

Framework Formulation

Given the comparison graph $\mathcal{G}$ with inconsistent multiple edge $\mathcal{E} = \{\boldsymbol{e}_c^\mathcal{U}, \boldsymbol{e}_{\bar{c}}^\mathcal{U} \mid c,\ \bar{c} \in \mathcal{C}_\mathcal{U}\}$, there are two goals:

- Detecting the outliers in the edge set $\mathcal{E}$. To this end, we introduce a set of unknown variables $\boldsymbol{\gamma} = \{\gamma_c\} \in \mathbb{R}^{|\mathcal{E}|}$ where each variable γ_c indicates whether the edge $\boldsymbol{e}_c^\mathcal{U}$ is an outlier or not. The outlier detection task in $\mathcal{C}_\mathcal{U}$ thus becomes the problem of estimating $\boldsymbol{\gamma}$ with $\mathcal{G}$.
- Obtaining an embedding $\boldsymbol{X} \in \mathbb{R}^{p\times n}$. Without prior knowledge, the squared Euclidean distance is selected as the dissimilarity measure, that is, given $\boldsymbol{e}_c^\mathcal{U} \in \mathcal{E}$ which represents the correct relative similarity measurement, we hope the embedding $\boldsymbol{X}$ satisfies

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 < \|\boldsymbol{x}_l - \boldsymbol{x}_k\|_2^2,\ c_u = (i,j,l,k)_u \in \mathcal{C}_\mathcal{U}.$$

In contrast to the multi-staged methods, we propose to jointly learn the embedding $\boldsymbol{X}$ by $\mathcal{G}$ and remove outliers globally via $\boldsymbol{\gamma}$ to avoid finding the maximum acyclic subgraph of $\mathcal{G}$. For this purpose, the outlier indicator $\boldsymbol{\gamma}$ and the embedding $\boldsymbol{X}$ are estimated in a unified framework. Given an edge $\boldsymbol{e}_c^\mathcal{U} \in \mathcal{E}$, its corresponding direction indicator y_c is modeled as

$$y_c = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 - \|\boldsymbol{x}_l - \boldsymbol{x}_k\|_2^2 + \gamma_c + \varepsilon_c, \tag{5}$$

where $\varepsilon_c \sim \mathcal{N}(0,\sigma^2)$ is the Gaussian noise with zero mean and a variance σ. The outlier indicator γ_c is assumed to have a larger magnitude than σ. For multiple edge $\boldsymbol{e}_c^\mathcal{U}$, if they are not outliers, we expect $d_{ij} - d_{lk}$ to be approximately equal to y_c, therefore we have $\gamma_c = 0$. On the contrary, when the prediction of $d_{ij} - d_{lk}$ differs greatly from y_c, we can explain $\boldsymbol{e}_c^\mathcal{U}$ as the outliers and compensate for the discrepancy between the prediction and the annotation with a nonzero value of γ_c. The only prior knowledge we have on $\boldsymbol{\gamma}$ is that it is a sparse variable, *i.e.,* in most cases $\gamma_c = 0$.

Thanks to the low computational complexity and convenience in optimization, we will consider a linear model. We replace the embedding matrix $\boldsymbol{X}$ with the Gram matrix $\boldsymbol{G} = \boldsymbol{X}^\top\boldsymbol{X} = \{g_{ij}\}$ in (5) as

$$y_c = g_{ii} - 2g_{ij} + g_{jj} - g_{ll} + 2g_{lk} - g_{kk} + \gamma_c + \varepsilon_c. \tag{6}$$

Next we define the gradient operator (finite difference operator) on graph $\mathcal{G}$ which maps a dissimilarity function on the vertices $d : \mathcal{V} \to \mathbb{R}$ to an edge flow $\operatorname{grad} \nabla d : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ such that $\nabla d\,(v_{ij}, v_{lk}) = d_{ij} - d_{lk}$. For the whole graph $\mathcal{G}$, (6) can be re-written in the matrix form as

$$\boldsymbol{y} = \boldsymbol{Z} \odot \boldsymbol{G} + \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{7}$$

and we define

$$\boldsymbol{Z} \odot \boldsymbol{G} = \boldsymbol{Z}\boldsymbol{g} = \boldsymbol{Z} \cdot vec(\boldsymbol{G}),$$

where $\boldsymbol{Z} = \nabla \in \mathbb{R}^{|\mathcal{E}|\times n^2}$ is the design matrix, $\boldsymbol{y}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are vectors in $\mathbb{R}^{|\mathcal{E}|}$. It is easy to see that (7) is a linear model.

In order to estimate the $|\mathcal{E}| + n^2$ unknown parameters ($|\mathcal{E}|$ for $\boldsymbol{\gamma}$ and n^2 for $\boldsymbol{G}$), we aim to minimize the discrepancy between the annotation $\boldsymbol{y}$ and the prediction $\boldsymbol{Z}\odot\boldsymbol{G}+\boldsymbol{\gamma}$, as well as holding the outlier indicator $\boldsymbol{\gamma}$ sparse. In addition, the estimated $\boldsymbol{G}$ should be a positive semi-definite matrix and its rank would be no more than $p \ll n$. Note that $\boldsymbol{y}$ only contains information about direction, but not how many votes received by each multiple edge $\boldsymbol{e}_c^\mathcal{U}$. The discrepancy thus needs to be weighted by the number of votes, represented by the edge weight vector $\boldsymbol{w}$. To that end, we put a weighted ℓ_2-loss on the discrepancy, a sparsity enhancing penalty on $\boldsymbol{\gamma}$ as well as semi-definite positive and rank constraint on $\boldsymbol{G}$. It gives us the following optimization problem:

$$\begin{aligned} &\underset{\boldsymbol{G},\,\boldsymbol{\gamma}}{\textit{minimize}}\ \ \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{G},\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} \\ &\textit{subject to}\ \ \boldsymbol{rank}(\boldsymbol{G}) = p,\ \boldsymbol{G} \succeq 0, \end{aligned} \tag{8}$$

where

$$\begin{aligned}\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{G},\boldsymbol{\gamma}) &= \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{Z}\odot\boldsymbol{G}-\boldsymbol{\gamma}\|_{2,\boldsymbol{w}}^2 \\ &= \frac{1}{2}\sum_{\boldsymbol{e}_c^{\mathcal{U}}\in\mathcal{E}} w_c^2(y_c-\gamma_c-d_{ij}+d_{lk})^2 \\ &= \frac{1}{2}\|\boldsymbol{W}\boldsymbol{y}-(\boldsymbol{W}\boldsymbol{Z})\odot\boldsymbol{G}-\boldsymbol{W}\boldsymbol{\gamma}\|_2^2,\end{aligned} \tag{9}$$

and

$$\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} = \sum_{\boldsymbol{e}_c^{\mathcal{U}}\in\mathcal{E}} w_c|\gamma_c| = \|\boldsymbol{W}\boldsymbol{\gamma}\|_1, \tag{10}$$

where $\boldsymbol{W} = \mathit{Diag}(\boldsymbol{w})$ is the diagonal matrix of $\boldsymbol{w}$. Solving (8), our framework identifies outliers globally by integrating all relative similarity measurements together, and obtains the embedding matrix $\boldsymbol{X}$ via eigen-decomposition of $\boldsymbol{G}$. The noise term $\boldsymbol{\varepsilon}$ has been ignored in (8) because the discrepancy is mainly caused by outliers due to their larger magnitude.

Optimization

Let $\boldsymbol{y}_{\boldsymbol{w}} = \boldsymbol{W}\boldsymbol{y}$ and $\boldsymbol{g} = \mathit{vec}(\boldsymbol{G})$, we do variable substitution in the weighted ℓ_2-loss (9) as

$$\begin{aligned}&\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{G},\boldsymbol{\gamma}) \\ =\ & \frac{1}{2}\|\boldsymbol{y}_{\boldsymbol{w}}-\boldsymbol{W}\boldsymbol{Z}\cdot\boldsymbol{g}-\boldsymbol{W}\boldsymbol{\gamma}\|_2^2 \\ =\ & \frac{1}{2}\left\|\boldsymbol{y}_{\boldsymbol{w}}-\begin{bmatrix}\boldsymbol{W}\boldsymbol{Z} & \boldsymbol{W}\end{bmatrix}\begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{\gamma}\end{pmatrix}\right\|_2^2 \\ =\ & \frac{1}{2}\|\boldsymbol{y}_{\boldsymbol{w}}-\boldsymbol{A}\cdot\boldsymbol{\beta}\|_2^2 := f(\boldsymbol{\beta}),\end{aligned} \tag{11}$$

and the sparsity enhancing penalty (10) can be re-written as

$$\lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} = \lambda\left\|\begin{bmatrix}\boldsymbol{0} & \boldsymbol{W}\end{bmatrix}\begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{\gamma}\end{pmatrix}\right\|_1 = \lambda\|\boldsymbol{B}\cdot\boldsymbol{\beta}\|_1 := g(\boldsymbol{\beta}). \tag{12}$$

With (11), (12) and ignoring the constraints on $\boldsymbol{G}$, (8) is equivalent to a *LASSO* formulation

$$\arg\min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}) := f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}). \tag{13}$$

Let $\mathcal{F}$ be the solution set of (13) and suppose $\boldsymbol{G}^*$ is a optimal solution of (8), it holds that $\boldsymbol{G}^* \in \mathcal{F}$. As a consequence, we come to a semi-definite programming with rank equality constraint

$$\begin{aligned}&\textit{find}\ \ \boldsymbol{G},\ \boldsymbol{\gamma} \\ &\textit{subject to}\ \ \boldsymbol{G},\ \boldsymbol{\gamma}\in\mathcal{F},\ \boldsymbol{G}\succeq 0,\ \boldsymbol{rank}(\boldsymbol{G}) = p.\end{aligned} \tag{14}$$

Solving a SDP with rank equality constraints like (14) is notoriously difficult. It is proposed in (Dattorro 2010) to solve this problem via iteratively solving the following two convex problems:

$$\begin{aligned}&\underset{\boldsymbol{G},\,\boldsymbol{\gamma}}{\textit{minimize}}\ \ \langle\boldsymbol{G},\boldsymbol{K}^*\rangle && \text{(15a)}\\ &\textit{subject to}\ \ \boldsymbol{G},\ \boldsymbol{\gamma}\in\mathcal{F},\ \boldsymbol{G}\succeq 0, \\ &\underset{\boldsymbol{G}}{\textit{minimize}}\ \ \langle\boldsymbol{G}^*,\boldsymbol{K}\rangle && \text{(15b)}\\ &\textit{subject to}\ \ \boldsymbol{trace}(\boldsymbol{K}) = n-p,\ \boldsymbol{0}\preceq\boldsymbol{K}\preceq\boldsymbol{I},\end{aligned}$$

where $\langle\boldsymbol{G},\boldsymbol{K}\rangle = \boldsymbol{trace}(\boldsymbol{K}^\top\boldsymbol{G})$, $\boldsymbol{G}^*$ is an optimal solution of (15a) and $\boldsymbol{K}^*$ is an optimal solution of (15b). However, this iteration of the convex problem sequence generally produces a solution of $\boldsymbol{G}$ which satisfies $\mathit{rank}(\boldsymbol{G}) \le p$[1]. Meanwhile, the nuclear-norm heuristic which replaces the rank equality constraint in (14) with nuclear norm regularization often recovers a minimum-rank solution of an SDP feasibility problem (Recht, Fazel, and Parrilo 2010). Tuning the free parameter in these methods to generate a rank-p solution is computational intensive. Rounding methods, *e.g.* low-rank projection which finds the rank-p matrix that is closest to the positive semi-definite solution in some norm, can also be adopted to solve (14). But this method just provides the low-rank approximated solutions instead the exact solution. To obtain the rank-p solution of $\boldsymbol{G}$ explicitly, we leverage the ran-reduction for semi-definite programming (Lemon, So, and Ye 2016).

First, we solve the following optimization problem

$$\begin{aligned}&\textit{find}\ \ \boldsymbol{G},\ \boldsymbol{\gamma} \\ &\textit{subject to}\ \ \boldsymbol{G},\ \boldsymbol{\gamma}\in\mathcal{F},\ \boldsymbol{G}\succeq 0.\end{aligned} \tag{16}$$

For any $L > 0$, consider the following quadratic approximation of $F(\boldsymbol{\beta}) := f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$ at a given point $\boldsymbol{\beta}_0$:

$$\begin{aligned}Q_L(\boldsymbol{\beta},\boldsymbol{\beta}_0) &= f(\boldsymbol{\beta}_0) + \langle\boldsymbol{\beta}-\boldsymbol{\beta}_0,\nabla f(\boldsymbol{\beta}_0)\rangle \\ &\quad + \frac{L}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + g(\boldsymbol{\beta}),\end{aligned} \tag{17}$$

which admits a unique minimizer

$$\begin{aligned}&P_L(\boldsymbol{\beta}) \\ =\ & \arg\min_{\boldsymbol{\beta}}\left\{g(\boldsymbol{\beta}) + \frac{L}{2}\left\|\boldsymbol{\beta}-\left(\boldsymbol{\beta}_0-\frac{1}{L}\nabla f(\boldsymbol{\beta}_0)\right)\right\|^2\right\}.\end{aligned}$$

Note that the regularization $g(\boldsymbol{\beta})$ is a constant in term of $\boldsymbol{G}$, we introduce projected gradient descent to find a semi-definite positive solution and the iterative scheme is

$$\boldsymbol{G}_{t+1} = \Pi_{\mathcal{S}_+^n}\left(\boldsymbol{G}_t - \frac{1}{L}\,\nabla_{\boldsymbol{G}}\,f(\boldsymbol{\beta}_t)\right), \tag{18}$$

where $\Pi_{\mathcal{S}_+^n}$ is the semi-definite positive projection of a matrix in the Frobenius norm. Since the ℓ_1 norm is separable, the computation of $\boldsymbol{\gamma}$ reduces to solving a one-dimensional minimization problem for each of its components,

$$\boldsymbol{\gamma}_{t+1} = \mathcal{T}_{\boldsymbol{\mu}}\left(\boldsymbol{\gamma}_t - \frac{1}{L}\,\nabla_{\boldsymbol{\gamma}}\,f(\boldsymbol{\beta}_t)\right), \tag{19}$$

where $\mathcal{T}_{\boldsymbol{\mu}} : \mathbb{R}^{|\mathcal{E}|} \to \mathbb{R}^{|\mathcal{E}|}$ is the shrinkage operator

$$[\mathcal{T}_{\boldsymbol{\mu}}(\boldsymbol{\gamma})]_i = \max(|\gamma_i|-\mu_i,0)\cdot\boldsymbol{sign}(\gamma_i),$$

and $\boldsymbol{\mu} = \frac{\lambda}{L}\boldsymbol{w}$.

[1] Because this iterative scheme for constraining rank in semidefinite program is not a projection method, it can find a rank-p solution $\boldsymbol{G}^*$ only if at least one exists in the feasible set of (14). When a rank-p feasible solution to (14) exists, it remains an open problem to state conditions under which $\langle\boldsymbol{G}^*,\boldsymbol{K}^*\rangle = \sum_{i=p+1}^n \sigma_i(\boldsymbol{G}^*) = 0$ is achieved by iterative solution of (15a) and (15b).

Suppose $\boldsymbol{\beta}^+$ is a solution of (16)

$$\boldsymbol{\beta}^+ = [\boldsymbol{G}^+,\ \boldsymbol{\gamma}^+],\ \boldsymbol{G}^+ \in \mathcal{F} \cap \mathbb{S}_+^n,$$

where $\mathbb{S}_+^n$ is the *PSD* cone of n-dimensional matrix. The main step of rank reduction is finding a new solution $\boldsymbol{G}^*$ which satisfies $\boldsymbol{rank}(\boldsymbol{G}^*) < \boldsymbol{rank}(\boldsymbol{G}^+)$. Here we assume that $\boldsymbol{null}(\boldsymbol{G}^*) \subset \boldsymbol{null}(\boldsymbol{G}^+)$. Since $\boldsymbol{G}^+ \in \mathbb{S}_+^n$, there exists a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ such that $\boldsymbol{G}^+ = \boldsymbol{U}\boldsymbol{U}^\top$. We hope $\boldsymbol{G}^*$ has the following update rule:

$$\begin{aligned} \boldsymbol{G}^* &= \boldsymbol{G}^+ + \alpha\, \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^\top \\ &= \boldsymbol{U}(\boldsymbol{I} + \alpha\boldsymbol{\Delta})\boldsymbol{U}^\top, \end{aligned} \tag{20}$$

where $\alpha < 0$ is a step size and $\boldsymbol{\Delta} \in \mathbb{S}_+^p$ is the update direction. Here we reformulate (16) as a standard SDP

$$\begin{aligned} &\textit{minimize} && \|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} \\ &\textit{subject to} && \langle \boldsymbol{G}, \boldsymbol{A}_c\rangle + \gamma_c = y_c,\ \boldsymbol{e}_c^{\mathcal{U}} \in \mathcal{E}, \\ & && \boldsymbol{G} \succeq 0, \end{aligned} \tag{21}$$

where $\boldsymbol{A}_c \in \mathbb{R}^{n\times n}$ is a symmetric matrix which has zero entry everywhere except on the entries corresponding to $c = (i, j, l, k)$ which has the form

$$\boldsymbol{A}_c = \begin{array}{c} \\ i \\ j \\ l \\ k \end{array} \begin{array}{c} \begin{array}{cccc} i & j & l & k \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \end{array}. \tag{22}$$

Due to $\boldsymbol{G}^+,\ \boldsymbol{\gamma}^+$ is a solution of (21), we need $\boldsymbol{G}^*$ satisfies the equality constraints

$$\langle \boldsymbol{G}^*, \boldsymbol{A}_c\rangle + \gamma_c^+ = y_c,\ \boldsymbol{e}_c^{\mathcal{U}} \in \mathcal{E}. \tag{23}$$

Substituting (20) into them and simplifying give the conditions

$$\left\langle \boldsymbol{\Delta},\ \boldsymbol{U}^\top \boldsymbol{A}_c \boldsymbol{U} \right\rangle = 0,\ \boldsymbol{e}_c^{\mathcal{U}} \in \mathcal{E}. \tag{24}$$

For convenience we define the mapping $\mathcal{A}_{\boldsymbol{U}} : \mathbb{S}_+^p \to \mathbb{R}^{|\mathcal{E}|}$ such that

$$\mathcal{A}_{\boldsymbol{U}}(\boldsymbol{\Delta}) = \begin{bmatrix} \left\langle \boldsymbol{\Delta},\ \boldsymbol{U}^\top \boldsymbol{A}_1 \boldsymbol{U} \right\rangle \\ \vdots \\ \left\langle \boldsymbol{\Delta},\ \boldsymbol{U}^\top \boldsymbol{A}_{|\mathcal{E}|} \boldsymbol{U} \right\rangle \end{bmatrix} = \begin{bmatrix} \left\langle \boldsymbol{A}_1,\ \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^\top \right\rangle \\ \vdots \\ \left\langle \boldsymbol{A}_{|\mathcal{E}|},\ \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^\top \right\rangle \end{bmatrix}.$$

Then we can express the condition (24) as

$$\mathcal{A}_{\boldsymbol{U}}(\boldsymbol{\Delta}) = 0,\ \text{ or } \ \boldsymbol{\Delta} \in \boldsymbol{null}(\mathcal{A}_{\boldsymbol{U}}).$$

Moreover, we choose $\alpha < 0$ to make $\boldsymbol{rank}(\boldsymbol{G}^*) < \boldsymbol{rank}(\boldsymbol{G}^+)$ and keep $\boldsymbol{G}^*$ be a solution of (16). It means that

$$\boldsymbol{I} + \alpha\boldsymbol{\Delta} \in \mathbb{S}_+^n,$$

and $\boldsymbol{I} + \alpha\boldsymbol{\Delta}$ is singular. The process of rank reduction is given as Algorithm 1. The following theorem guarantees the existence of $\boldsymbol{G}^*$, the part of a solution of (16) with

$$\boldsymbol{rank}(\boldsymbol{G}^*) = p,\ \frac{p(p+1)}{2} \le |\mathcal{C}|,$$

where $|\mathcal{C}|$ is the number of linear equality constraints in (16). As the low-dimensional embedding requires that $p \ll n$ and the number of comparisons $|\mathcal{C}|$ is suggested to be $\boldsymbol{O}(pn\log n)$ (Jain, Jamieson, and Nowak 2016), such a $\boldsymbol{G}^*$ could be solved efficiently via Algorithm 1.

Theorem 1. *(Lemon, So, and Ye 2016) If* (16) *is solvable, then it has a solution contains $\boldsymbol{G}^*$ with $\boldsymbol{rank}(\boldsymbol{G}^*) = p$ such that $p(p+1)/2 \le |\mathcal{C}|$. Moreover, Algorithm 1 efficiently finds such $\boldsymbol{G}^*$.*

Algorithm 1: ***rank-reduction***$(\boldsymbol{\beta}^+, p)$

Input: $\boldsymbol{\beta}^+ = [\boldsymbol{G}^+, \boldsymbol{\gamma}^+]$ is a solution of (21), p is the embedding dimension.
Output: $\boldsymbol{G}^*$, which satisfies $\boldsymbol{rank}(\boldsymbol{G}^*) = p$.

Initialize $\boldsymbol{G}^* = \boldsymbol{G}^+$;
repeat
  $\boldsymbol{G}^* = \boldsymbol{U}\boldsymbol{U}^\top,\ \boldsymbol{U} \in \mathbb{R}^{n\times p}$;
  find a nonzero $\boldsymbol{\Delta} \in \boldsymbol{null}(\mathcal{A}_{\boldsymbol{U}})$ (if possible);
  find a maximum-magnitude eigenvalue σ_1 of $\boldsymbol{\Delta}$;
  update

$$\boldsymbol{G}^* = \boldsymbol{U}\left(\boldsymbol{I} - \frac{1}{\sigma_1}\boldsymbol{\Delta}\right)\boldsymbol{U}^\top.$$

until $\boldsymbol{null}(\mathcal{A}_{\boldsymbol{U}}) = \{0\}$;

At the end of this section, we summarize the whole optimization algorithm as Algorithm 2. The reproducible code can be found here[2].

Algorithm 2: FISTA with rank reduction for (14)

Input: Comparison graph $\mathcal{G}$, the edge direction flag $\boldsymbol{y}$, multiple edge weight $\boldsymbol{w}$, the regularization parameter λ and the embedding dimension p.
Output: $\boldsymbol{G}^*, \boldsymbol{\gamma}^*$.

Initialize $\boldsymbol{G}_0 \in \mathbb{R}^{n\times n}, \boldsymbol{\gamma}_0 \in \mathbb{R}^{|\mathcal{E}|}, \boldsymbol{\beta}_0 = [\boldsymbol{G}_0, \boldsymbol{\gamma}_0]$; Set $L_0 > 0, \eta > 1, t_1 = 1, k = 1, \boldsymbol{\beta}_1^* = \boldsymbol{\beta}_0$;
while *not satisfies the stopping rules* **do**
  Find the smallest nonnegative integer i_k such that with $\bar{L} = \eta^{i_k} L_{k-1}$

$$F(P_{\bar{L}}(\boldsymbol{\beta}_k^*)) \le Q_{\bar{L}}(P_{\bar{L}}(\boldsymbol{\beta}_k^*),\ \boldsymbol{\beta}_k^*)$$

  Set $L_k = \eta^{i_k} L_{k-1}$ and update

$$\begin{aligned} \boldsymbol{\beta}_k &= P_{L_k}(\boldsymbol{\beta}_k^*) \text{ via (18) and (19)} \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ \tilde{\boldsymbol{\beta}}_{k+1} &= \boldsymbol{\beta}_k + \frac{t_k - 1}{t_{k+1}}\left(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\right) \\ \boldsymbol{\beta}_{k+1}^* &= \boldsymbol{rank\text{-}reduction}(\tilde{\boldsymbol{\beta}}_{k+1}) \end{aligned}$$

  $k \leftarrow k + 1$;
end

[2] https://github.com/alphaprime/ROE

Experiments

Simulation

Dataset. The simulated dataset consists of 100 points $\{\boldsymbol{x}_i\}_{i=1}^{100} \subset \mathbb{R}^{10}$, where $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{20}\boldsymbol{I})$, $\boldsymbol{I} \in \mathbb{R}^{10\times 10}$ is the identity matrix. The possible similarity triple-wise comparisons are generated based on the Euclidean distances between $\{\boldsymbol{x}_i\}$. We randomly sample 10000 correct triplets as the basic training data and a validation set is build with the same number of triplets. The remains are served as the test set.

Settings. The existing ordinal embedding methods adopt triple-wise comparisons as the constraints. The triple-wise comparison set $\mathcal{T} = \{(i, j, k)\}$ is the special case of quadruplet which means $l = i$ in $c = (i, j, l, k) \in \mathcal{C}$. The differences between the above triplets setting and the generalized formulation are two-fold, (i) the edge of triple-wise comparison graph $\mathcal{G}_\mathcal{T} = \{\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T}\}$ only links a pair of nodes which share the same item, such as v_{ij} and v_{ik}, (ii) the equality constraints in (21) could be modified as $\langle \boldsymbol{G}, \boldsymbol{A}_t\rangle + \gamma_t = y_t$, $\boldsymbol{e}_t^\mathcal{U} \in \mathcal{E}_\mathcal{T}$, where $\boldsymbol{A}_t$ is a symmetric $n \times n$ matrix indicated by $t = (i, j, k)$ as

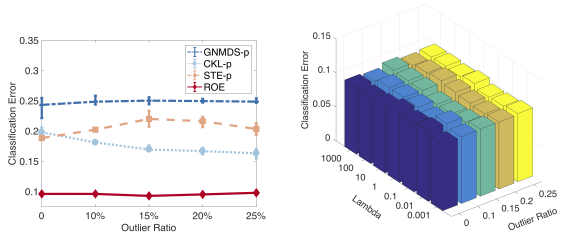
$$\boldsymbol{A}_t = \begin{array}{c} \\ i \\ j \\ k \end{array} \begin{array}{c} \begin{array}{ccc} i & j & k \end{array} \\ \begin{pmatrix} 0 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \end{array}. \tag{25}$$

With these modifications, the proposed method employs triple-wise comparisons $\mathcal{T}$ to construct the comparison graph $\mathcal{G}_\mathcal{T}$ for fair competition.

The basic training comparisons are augmented to represent the votes received on a triplet. For each triplet t, there will be s copies, $t_1, \ldots, t_s$, $15 \le s \le 50$. Second, errors are then synthesized according to the different ratios: we assume that at most $q\%$ of all relative comparisons are not consistent with the ground-truth metric information and we change the position of j and k in each randomly chosen triplet, where q ranges from 10% to 25%. Thus the outliers of these comparisons could not be the minority of the edges between a pair of nodes. Specially, we conduct the experiments with the "noiseless" case where $q = 0$ which indicates that the local outlier detection can detect all outliers in the augmented training comparisons. Actually, this setting is just an ideal case which would not be expected in real applications. At last, the comparison graph $\mathcal{G}_\mathcal{T}$ is constructed, where $\mathcal{T}$ is the contaminated training set.

Evaluation Metrics. We employ the classification error to evaluate the performance of various algorithms. The learned Gram matrix $\boldsymbol{G}$ can predict the direction of an unseen but possible edge in the graph. The percentage of wrong prediction, which is not consistent with the ground-truth metric information, is the classification error of the learned embedding. Larger classification error means lower quality of the learned embedding.

Competitors. We compare the proposed algorithm with three well-known ordinal embedding methods: GNMDS (Agarwal et al. 2007), CKL (Tamuz et al. 2011) and STE (van der Maaten and Weinberger 2012). Note that we adopt the strategies proposed by (Jain, Jamieson, and Nowak 2016), which performs projected gradient descent with line search. The learned matrix are projected onto the subspace spanned by the top p eigenvalues at each iteration, *i.e.* setting the smallest $n-p$ eigenvalues to 0. We call the three new algorithms: GNMDS-p, CKL-p and STE-p, correspondingly. The parameters of these competitors are tuned on the validation set.



(a) Comparative evaluation. (b) Parameter sensitivity.

Figure 2: The results on the synthetic data. (a) The classification performance comparative evaluation. Smaller classification error means better embedding result. The mean and standard deviation of each method over 20 trials are shown in the plots. (b) The classification error with different choices of λ.

Comparative Results. The embedding performance of various models are evaluated when different ratios of outliers are considered. The results are shown in Figure 2a and Table 1. We provide more details in supplementary materials. It shows clearly that ***ROE*** significantly outperforms the three alternatives for a wide range of noise density. This validates the effectiveness of ***ROE***. In particular, it can be observed that: (i) the improvement over the three competitors demonstrates the superior generalization ability of ***ROE*** thanks to the unified framework rather than phased methodology. The traditional methods rely heavily on majority voting and maximum acyclic subgraph approximation as the preprocessing, but their models ignore the intrinsic inconsistencies between the noise comparisons and the ground-truth similarity relationship of $\boldsymbol{X}$. More important, these inconsistencies, especially the global ones, would not be conquered by the local outlier detection methods. Consequently, the three alternatives would suffer from the wrong training comparisons and the classification error would be amplified and accumulated. However, the proposed ***ROE*** method incorporate the global outlier detection scheme with ordinal embedding. This unified framework not only benefits from the correct training samples which would be pruned by majority voting and maximum acyclic subgraph approximation, but also gets rid of the contamination from the outliers. (ii) Even under the "noiseless" situation, the proposed method also shows the superiority. This improvement comes from the exact rank-p solution of Algorithm 1 and the regression-based framework which aggregates all the votes on a comparison.

Parameter Sensitivity Analysis. To show the sensitivity of ***ROE*** toward the free parameter λ in (8) changes, we record the average classification error over 20 runs for the synthetic datasets with different λ. The corresponding result are

Table 1: Classification error results on the synthetic dataset.

(a) Clean Data (without outlier)

Methods	min	median	max	std
GNMDS-p	0.2061	0.2434	0.2607	0.0194
CKL-p	0.1879	0.1979	0.2088	0.0062
STE-p	0.1829	0.1884	0.1930	0.0031
Ours	**0.0953**	**0.0961**	**0.1049**	**0.0024**

(b) Contaminated Data (with 25% outliers)

Methods	min	median	max	std
GNMDS-p	0.2373	0.2487	0.2586	0.0063
CKL-p	0.1443	0.1632	0.1809	0.0103
STE-p	0.1828	0.2035	0.2250	0.0127
Ours	**0.0960**	**0.1025**	**0.1070**	**0.0032**

Table 2: Classification error results on the music artists dataset.

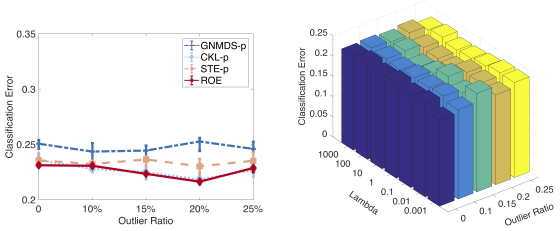
(a) Clean Data (without outlier)

Methods	min	median	max	std
GNMDS-p	0.2373	0.2506	0.2828	0.0101
CKL-p	**0.2239**	0.2358	0.2813	0.0117
STE-p	0.2278	0.2356	0.2763	0.0122
Ours	0.2289	**0.2312**	**0.2356**	**0.0016**

(b) Contaminated Data (with 25% outliers)

Methods	min	median	max	std
GNMDS-p	0.2310	0.2460	0.2887	0.0121
CKL-p	0.2099	**0.2120**	0.2505	0.0120
STE-p	**0.2027**	0.2203	0.2579	0.0116
Ours	0.2226	0.2287	**0.2389**	**0.0047**

shown in Figure 2b. We find that the performance of ***ROE*** keeps relatively stable overall in a wide range, from 10^{-3} to 10^{3}.



(a) Comparative evaluation. (b) Parameter analysis.

Figure 3: The results on the music data. (a) The classification comparative evaluation. Smaller classification error means better embedding result. The mean and standard deviation of each method over 20 trials are shown in the plots. (b) The sensitive analysis with different choices of λ.

Music Artists Similarity

Dataset. The music artist data is collected by (Ellis et al. 2002) via a web-based survey in which $1,032$ users provided triple-wise comparisons on the similarity of 412 music artists. The traditional methods like (van der Maaten and Weinberger 2012) adopt the majority voting and the maximum acyclic subgraph approximation to prune the inconsistence comparisons. Therefore, a much smaller version[3], which has only $9,107$ triplets for $n = 400$ artists, is employed by the existing methods. For fair comparison, we evaluate ***ROE*** on this subset. The size of training samples is $5,000$ and the validation set contains $2,000$ triplets. 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. It's worth noting that these $9,107$ triplets still include outliers duo the inappropriate preprocessing. Accordingly, the evaluation on this data verifies that the local outlier detection and maximum acyclic subgraph approximation should be eliminated.

Comparative Results. Without the ground truth of music artists similarity values, different models were evaluated indirectly via classification accuracy based on the noise test comparisons in Figure 3a. The following observations can be made: (i) the wrong data would damage the performance of ordinal embedding methods. The outliers in training data cause the three alternatives to generate the sub-optimal solutions as they can't prune the outliers without preprocessing. The global outlier detection is more accurate but the wrong comparisons in validation set and test set would lead the evaluation metric of ***ROE*** to be higher than these competitors. This phenomenon proves the effectiveness of ***ROE*** from the opposite side. (ii) Benefit from the global outlier detection ability, the variance of ***ROE*** is much smaller than these three alternatives because ***ROE*** can prune the outliers and keep the result more stable.

Conclusions

In this paper we introduce a novel unified robust framework to construct the representation of items in the Euclidean space $\mathbb{R}^p$ with contaminated comparisons. The key advantage of our method over the existing approaches is that our model infers the embedding and detects the outliers jointly by minimizing a global ranking inconsistence cost. It can be formulated as a partial penalized LASSO optimization problem. Efficient algorithm is proposed to obtain a positive semi-definite solution which satisfies the rank equality constraint. Experimental studies are conducted with both synthetic and real-world data. Our results suggest that the local outlier detection is not a reliable tool for ordinal embedding with contaminated comparisons.

[3] https://lvdmaaten.github.io/ste/Stochastic_Triplet_Embedding.html

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.

Chen, Y.; Garcia, E. K.; Gupta, M. R.; Rahimi, A.; and Cazzanti, L. 2009. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research* 10:747–776.

Chen, X.; Bennett, P. N.; Collins-Thompson, K.; and Horvitz, E. 2013. Pairwise ranking aggregation in a crowd-sourced setting. In *ACM International Conference on Web Search and Data Mining*, 193–202.

Dattorro, J. 2010. *Convex optimization & Euclidean distance geometry*. Lulu. com.

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*.

Garey, M. R., and Johnson, D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Gehrlein, W. V. 2006. *Condorcets paradox*. Springer.

Huber, P., and Ronchetti, E. 2009. *Robust Statistics*. Wiley.

Jain, L.; Jamieson, K. G.; and Nowak, R. 2016. Finite sample prediction and recovery bounds for ordinal embedding. *Annual Conference on Neural Information Processing Systems* 2711–2719.

Jamieson, K., and Nowak, R. 2011. Active ranking using pairwise comparisons. *Annual Conference on Neural Information Processing Systems* 2240–2248.

Jiang, X.; Lim, L.; Yao, Y.; and Ye, Y. 2011. Statistical ranking and combinatorial hodge theory. *Mathematical Programming* 127(1):203–244.

Kleindessner, M., and Luxburg, U. 2014. Uniqueness of ordinal embedding. *Conference on Learning Theory* 40–67.

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.

Lemon, A.; So, A. M.; and Ye, Y. 2016. Low-rank semidefinite programming: Theory and applications. *Foundations and Trends in Optimization* 2(1-2):1–156.

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.

McFee, B., and Lanckriet, G. 2011. Learning multi-modal similarity. *Journal of Machine Learning Research* 12:491–523.

Recht, B.; Fazel, M.; and Parrilo, P. A. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* 52(3):471–501.

Rousseeuw, P. J., and Leroy, A. M. 2005. *Robust regression and outlier detection*, volume 589. John wiley & sons.

Rousseeuw, P., and Yohai, V. 1984. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*. Springer. 256–272.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American statistical association* 79(388):871–880.

She, Y., and Owen, A. B. 2011. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* 106(494):626–639.

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.

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.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Wei, X.; Zhu, J.; Feng, S.; and Su, H. 2018. Video-to-video translation with global temporal consistency. In *ACM Conference on Multimedia*, 18–25.

Welinder, P.; Branson, S.; Belongie, S. J.; and Perona, P. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2424–2432.

Xu, Q.; Huang, Q.; Jiang, T.; Yan, B.; Lin, W.; and Yao, Y. 2012. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia* 14(3):844–857.