Local Ordinal Embedding

Yoshikazu Terada[1,2] TERADA@SIGMATH.ES.OSAKA-U.AC.JP
[1]Graduate School of Engineering Science, Osaka University, Japan
[2]CiNet, National Institute of Information and Communications Technology, Japan.

Ulrike von Luxburg[3] LUXBURG@INFORMATIK.UNI-HAMBURG.DE
[3]Department of Computer Science, University of Hamburg, Germany

Abstract

We study the problem of ordinal embedding: given a set of ordinal constraints of the form $distance(i,j) < distance(k,l)$ for some quadruples (i,j,k,l) of indices, the goal is to construct a point configuration $\hat{\boldsymbol{x}}_1, ..., \hat{\boldsymbol{x}}_n$ in $\mathbb{R}^p$ that preserves these constraints as well as possible. Our first contribution is to suggest a simple new algorithm for this problem, Soft Ordinal Embedding. The key feature of the algorithm is that it recovers not only the ordinal constraints, but even the density structure of the underlying data set. As our second contribution we prove that in the large sample limit it is enough to know "local ordinal information" in order to perfectly reconstruct a given point configuration. This leads to our Local Ordinal Embedding algorithm, which can also be used for graph drawing.

1. Introduction

In this paper we consider the problem of ordinal embedding, also called ordinal scaling, non-metric multidimensional scaling, monotonic embedding, or isotonic embedding. Consider a set of objects $x_1, ..., x_n$ in some abstract space $\mathcal{X}$. We assume that there exists a dissimilarity function $\xi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that assigns dissimilarity values ξ_{ij} to pairs of objects (i,j). However, this dissimilarity function is unknown to us. All we get are ordinal constraints

$$\xi_{ij} < \xi_{kl} \text{ for certain quadruples of indices } (i,j,k,l). \quad (\star)$$

Our goal is to construct an embedding $\hat{\boldsymbol{x}}_1, ..., \hat{\boldsymbol{x}}_n$ in some Euclidean space $\mathbb{R}^p$ of given dimension p such that all ordinal constraints are preserved:

$$\xi_{ij} < \xi_{kl} \iff \|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\| < \|\hat{\boldsymbol{x}}_k - \hat{\boldsymbol{x}}_l\|.$$

This problem has first been studied by Shepard (1962a;b) and Kruskal (1964a;b), and lately has drawn quite some attention in the machine learning community (Quist & Yona, 2004; Rosales & Fung, 2006; Agarwal et al., 2007; Shaw & Jebara, 2009; McFee & Lanckriet, 2009; Jamieson & Nowak, 2011a; McFee & Lanckriet, 2011; Tamuz et al., 2011; Ailon, 2012), also in its special case of ranking (Ouyang & Gray, 2008; McFee & Lanckriet, 2010; Jamieson & Nowak, 2011b; Lan et al., 2012; Wauthier et al., 2013).

Soft ordinal embedding. The first main contribution of our paper is to develop a new simple and efficient method for ordinal embedding. We propose a new "soft" objective function that not only counts the number of violated constraints, but takes into account the amount of violation. The resulting optimization problem is surprisingly simple: it does not have any parameters that need to be tuned, and it can be solved by standard *unconstrained* optimization techniques. We develop an efficient majorization algorithm for this purpose. The resulting ordinal embedding has the nice feature that it not only preserves the *ordinal structure* of the data, but it even preserves the *density structure* of the data. This is a key feature for machine learning because the results of learning algorithms crucially depend on the data's density. The code of our algorithm has been published as an official R-package (Terada & von Luxburg, 2014).

Local ordinal embedding. There exists a fundamental theoretical question about ordinal embedding that has received surprisingly little attention in the literature. Namely, in how far does ordinal information as in ($\star$) determine the geometry and the density of an underlying data set? It is widely believed (p. 294 of Shepard, 1966; Section 2.2 of Borg & Groenen, 2005; Section 5.13.2 of Dattorro, 2005) and has recently been proved (Kleindessner & von Luxburg, 2014) that upon knowledge of the ordinal relationships for *all* quadruples (i,j,k,l), the point set $x_1, ..., x_n$ can be approximately reconstructed up to similarity transforms if n is large enough An even more interesting question is whether we really need knowledge about *all* quadruples (i,j,k,l), or whether some subset of

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

quadruples is already sufficient to guarantee the uniqueness of the embedding. Particularly interesting are local ordinal constraints. In the metric world, the whole field of differential geometry is based on the insight that knowledge about local distances is enough to uniquely determine the geometry of a set. Does such a result also hold if our local knowledge only concerns ordinal relationships, not metric distances? As the second main contribution of this paper, we provide a positive answer to this question: if the sample size n is large enough, then it is possible to approximately reconstruct the point set $x_1, ..., x_n$ if we just know who are each point's k-nearest neighbors (for a parameter k to be specified later). That is, the local ordering induced by the distance function already determines the geometry and the density of the underlying point set.

Application to graph drawing. The local point of view suggests ordinal embedding as an interesting alternative to graph drawing algorithms. If vertex i is connected by an edge to vertex j, but not to vertex k, we interpret this constellation as a constraint of the form $\xi_{ij} < \xi_{ik}$. With this interpretation, graph embedding (drawing) for unweighted graphs becomes a special case of ordinal embedding.

2. Related work

Ordinal embedding. Ordinal embedding was invented as a tool for the analysis of psychometric data by Shepard (1962a;b; 1966) and Kruskal (1964a;b). An approach based on Gram matrices, called generalized non-metric MDS (GNMDS), was proposed in Agarwal et al. (2007). This approach solves a relaxed version of the embedding problem as a semi-definite program. In a similar spirit, Shaw & Jebara (2009) introduced structure preserving embedding (SPE) for embedding unweighted nearest neighbor graphs to Euclidean spaces. In practice, both SPE and GNMDS have a number of disadvantages. The computational costs of the semi-definite programs are high, and both algorithms have tuning parameters that have to be chosen by some heuristic. Moreover, as a consequence of relaxation it may happen that even if a perfect embedding exists, it is not a solution of the optimization problem. More on the theoretical side, there is related work on monotone maps and sphericity (Bilu & Linial, 2005), with focus on the question of the minimal achievable dimension p in the non-realizable case, and Alon et al. (2008), with the focus on the worst case distortion guarantees for embedding arbitrary metrics in the Euclidean space. In the machine learning community, the work of Jamieson & Nowak (2011a) investigates a lower bound for the minimum number of queries of the form "Is $\xi_{ij} \leq \xi_{kl}$?" for realizing an ordinal embedding, and similar work exists for the special case of ranking (Jamieson & Nowak, 2011b; Ailon, 2012; Wauthier et al., 2013). There is also a large literature on the special case of graph embedding and graph drawing, see for example the recent monograph by Tamassia (2013). In our experiments, we include some of the most well-known graph drawing algorithms such as the one by Fruchterman & Reingold (1991) and by Kamada & Kawai (1989).

Metric embeddings. There exists a huge body of work on algorithms that embed data points based on *metric information*. An overview over the traditional approach of metric multidimensional scaling is available in Borg & Groenen (2005). Many of the recent embedding algorithms follow the paradigm that it is enough to preserve local distances: Isomap (Tenenbaum et al., 2000), locally linear embeddings (Roweis & Saul, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2003), stochastic neighbor embedding (SNE; Hinton & Roweis, 2002), t-SNE (van der Maaten & Hinton, 2008), and so on. Related theoretical work includes the one of metric k-local embeddings, where the target is a Johnson-Lindenstrauss-type theorem under the assumption that only local distances have to be preserved (Gottlieb & Krauthgamer, 2011; Bartal et al., 2011). We are not aware of any theoretical work on *local ordinal embeddings.*

3. Soft ordinal embedding

3.1. A soft objective function

Consider a set of n objects with pairwise dissimilarity scores ξ_{ij}. To encode ordinal constraints, we introduce a subset $\mathcal{A} \subset \{1, \ldots, n\}^4$ of quadruples such that

$$(i, j, k, l) \in \mathcal{A} \iff \xi_{ij} < \xi_{kl}.$$

Note that at this point the set $\mathcal{A}$ is allowed to be any subset of $\{1, \ldots, n\}^4$. For given ordinal information $\mathcal{A}$ and given dimension p, the aim of ordinal embedding is to find a p-dimensional embedding $X = (x_{is})_{n \times p}$ that preserves the given ordinal information as well as possible. Denote by $d_{ij}(X) := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ the Euclidean distances between the embedded points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The most natural objective function for ordinal embedding is

$$\mathrm{Err}_{\mathrm{hard}}(X \mid \mathcal{A}) := \textstyle\sum_{(i,j,k,l) \in \mathcal{A}} \mathbb{1}[d_{ij}(X) \geq d_{kl}(X)].$$

However, this objective function is discrete and difficult to optimize. Moreover, it does not take into account the "amount" by which a constraint is violated. This has already been observed by Johnson (1973), who suggested the alternative penalty function

$$\frac{\sum_{(i,j,k,l) \in \mathcal{A}} \max\left[0, d_{ij}^2(X) - d_{kl}^2(X)\right]^2}{\sum_{(i,j,k,l) \in \mathcal{A}} (d_{ij}^2(X) - d_{kl}^2(X))}.$$

The numerator is a continuous version of $\mathrm{Err}_{\mathrm{hard}}$ and the denominator's purpose is to prevent the degenerate solution $X \equiv 0$. However, since the denominator depends on

X, this objective function is cumbersome to optimize. In particular, no majorization algorithm exists for this type of stress function (nor for similar stress functions such as (1) in Kruskal, 1964a;b or (2) in Kruskal, 1968).

We now suggest an alternative approach. To overcome the problem of degeneracy, we introduce a scale parameter $\delta > 0$ and propose the objective function

$$\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta) := \\ \textstyle\sum_{i<j} \sum_{k<l} o_{ijkl} \max\left[0, d_{ij}(X) + \delta - d_{kl}(X)\right]^2, \quad (1)$$

where $o_{ijkl} = 1$ if $(i,j,k,l) \in \mathcal{A}$ and $o_{ijkl} = 0$ otherwise. We call the problem of minimizing $\mathrm{Err}_{\mathrm{soft}}$ the *soft ordinal embedding problem* (SOE). Note that in the realizable case, where the original point set comes from $\mathbb{R}^p$, the true point configuration is a global minimum of the objective function $\mathrm{Err}_{\mathrm{soft}}$. The following proposition shows that the parameter $\delta > 0$ just controls the scale of the embedding and has no further effect on the solution (the proof is straightforward and can be found in the supplement).

Proposition 1 (Scale parameter)**.** *Let* $\delta_1, \ \delta_2 > 0$. *If* $X_{\delta_1} := \arg\min \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_1)$ *is an optimal solution for parameter* δ_1*, then* $(\delta_2/\delta_1)X_{\delta_1}$ *is an optimal solution of* $\operatorname{argmin} \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_2)$ *for parameter* δ_2*.*

3.2. Majorization Algorithm for SOE

In order to minimize the objective function $\mathrm{Err}_{\mathrm{soft}}$, we propose a majorization algorithm. Let us briefly recap this general method. Let $\mathcal{X}$ be a non-empty set and $f : \mathcal{X} \to \mathbb{R}$ a real-valued function. A function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a majorizing function of f if it satisfies

(i)$f(x_0) = g(x_0, x_0)$ for all $x_0 \in \mathcal{X}$,
(ii) $f(x) \leq g(x, x_0)$ for all $x_0, \ x \in \mathcal{X}$.

For given $x_0 \in \mathcal{X}$, let $\tilde{x} \in \mathcal{X}$ be such that $g(\tilde{x}, x_0) \leq g(x_0, x_0)$. This implies $f(\tilde{x}) \leq g(\tilde{x}, x_0) \leq g(x_0, x_0) = f(x_0)$. Consequently, we can optimize the original function by minimizing a majorizing function instead of the original one. This can be of considerable advantage if the majorizing function g is easier to handle than the original function f. The update step of a majorization algorithm for minimizing f is $x_{t+1} = \arg\min_{x \in \mathcal{X}} g(x, x_{t-1})$.

To construct a majorizing function for our objective function $\mathrm{Err}_{\mathrm{soft}}$, we take inspiration from Groenen et al. (2006). Given any current candidate point configuration Y, we consider the following quadratic majorizing function:

Proposition 2 (Majorizing function)**.** *A majorizing function for each component of* $\mathrm{Err}_{\mathrm{soft}}$ *is given by*

$$\begin{aligned} & o_{ijkl} \max\left[0, d_{ij}(X) + \delta - d_{kl}(X)\right]^2 \\ & \leq \alpha_{ijkl} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + \alpha^*_{ijkl} \|\boldsymbol{x}_k - \boldsymbol{x}_l\|^2 \\ & \quad - 2\beta_{ijkl}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{y}_i - \boldsymbol{y}_j) \\ & \quad - 2\beta^*_{ijkl}(\boldsymbol{x}_k - \boldsymbol{x}_l)^T(\boldsymbol{y}_k - \boldsymbol{y}_l) + \gamma_{ijkl}. \end{aligned} \quad (2)$$

The parameters $\alpha_{ijkl}, \ \alpha^*_{ijkl}, \ \beta_{ijkl}, \ \beta^*_{ijkl}$*, and* γ_{ijkl} *only depend on* Y*. Their closed form expressions are provided in the supplement of the paper.*

The proof of this proposition is provided in the supplementary material, as well as the pseudocode for solving the soft ordinal embedding problem based on this majorizing function. For the special case of local ordinal embedding we explicitly state the majorization algorithm below.

4. Local ordinal embedding

The problem. The potential number of ordinal constraints of the form $(\star)$ is of the order $O(n^4)$, much too large to be practical. It is an interesting question whether it is possible to significantly reduce this number of constraints, without giving up on embedding quality. In particular, we are interested in the case of "local" ordinal constraints. By kNN$(i) \subset \{1, ..., n\}$ we denote the set of indices of the nearest neighbors of point x_i. Note that such a set encodes a particular subset of ordinal constraints:

$$j \in \text{kNN}(i) \text{ and } l \notin \text{kNN}(i) \implies \xi_{ij} < \xi_{il} \qquad (\star\star)$$

It is well known from the area of manifold algorithms that we can reconstruct a set of points if we know the *distances* of each point to its k-nearest neighbors. We now want to show the surprising fact that we do not even need to know the distances — it is enough to know the indices in the sets kNN(i) to reconstruct the point set.

It is convenient to formalize the neighborhood information in the form of the k-nearest neighbor graph, in which each point is connected to its k nearest neighbors by a directed, unweighted edge. We define the problem of *local ordinal embedding (LOE)* as follows: *Given a directed, unweighted kNN graph G, construct an embedding* $\hat{\boldsymbol{x}}_1, ..., \hat{\boldsymbol{x}}_n \in \mathbb{R}^p$ *such that the kNN graph of the new points coincides with the given graph G.*

Our algorithm. Consider a directed, unweighted kNN-graph with adjacency matrix $A = (a_{ij})_{i,j=1,...,n}$. With the notation $a^*_{ijk} := a_{ij}(1 - a_{ik})$, our objective for local ordinal embedding is a special realization of $\mathrm{Err}_{\mathrm{soft}}$, namely

$$\mathrm{Err}_{\mathrm{local}}(X \mid p, \delta) := \\ \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k \neq i}^{n} a^*_{ijk} \max\left[0, d_{ij}(X) + \delta - d_{ik}(X)\right]^2. \quad (3)$$

Based on Proposition 2, we obtain the following majorizing function for $\mathrm{Err}_{\mathrm{local}}$:

$$\mathrm{Err}_{\mathrm{local}}(X \mid p, \delta) \leq \sum_{s=1}^{p} \left[\boldsymbol{x}_s^T M \boldsymbol{x}_s - 2\boldsymbol{x}_s^T H \boldsymbol{y}_s\right] + \gamma,$$

where $\boldsymbol{x}_s = (x_{1s}, \dots, x_{ns})^T$, $\boldsymbol{y}_s = (y_{1s}, \dots, y_{ns})^T$, $M = (m_{ij})_{n\times n}$, $H = (h_{ij})_{n\times n}$, $\gamma = \sum_{i=1}^n \sum_{j\neq i}^n \sum_{k\neq i}^n \gamma_{ijk}$,

$$m_{ij} = \begin{cases} \frac{1}{2}\sum_{j\neq i}(m_{ij} + m_{ji}) & \text{if } i = j, \\ -2(\alpha_{iji\cdot} + \alpha^*_{i\cdot ij}) & \text{if } i \neq j, \end{cases}$$

$$h_{ij} = \begin{cases} \sum_{j\neq i}(h_{ij} + h_{ji}) & \text{if } i = j, \\ -(\beta_{iji\cdot} + \beta^*_{i\cdot ij} + \beta_{jij\cdot} + \beta^*_{j\cdot ji}) & \text{if } i \neq j. \end{cases}$$

Note that the diagonal elements of M are positive. For given Y and x_{js} $(j \neq i)$, we can optimize the majorization function with x_{is} by the following update rule:

$$x_{is} := \frac{2\sum_{j=1}^n h_{ij}y_{js} - \sum_{j\neq i}(m_{ij} + m_{ji})x_{js}}{2m_{ii}}. \qquad (4)$$

The pseudocode of the resulting majorization algorithm is presented in Algorithm 1, called LOE-MM in the following. The computational complexity of each of its iterations is $O(kn^2)$. Compare this to the complexity of structure preserving embedding (SPE), an algorithm that has been designed explicitly for the purpose of embedding k-nearest neighbor graphs: here, the complexity of each iteration is $O(n^3 + c^3)$ with the number of ordinal constraints $c = (n-k)kn$, so $O(k^3n^6)$ altogether. A similar complexity bound applies to the GNMDS algorithm.

As a final remark, note that local ordinal embedding applies in a straightforward manner to **general graph embedding** problems. Given a graph $G = (V, E)$, we formulate the constraints that $\xi_{ij} < \xi_{ik}$ if $(i, j) \in E$ and $(i, k) \notin E$. Then we continue as above. In the supplementary material we demonstrate that LOE works gracefully for visualizing moderately sized graphs and can outperform standard graph drawing algorithms.

Algorithm 1 LOE-MM: Majorization minimization algorithm for local ordinal embedding

```
Set $\delta > 0$ to a scale parameter and set $X_0$ to some initial $n \times p$ coordinate matrix.
Set iteration counter $t := 0$ and $X_{-1} := X_0$.
Set $\varepsilon > 0$ to a small value as the convergence criterion (e.g., $\varepsilon = 10^{-5}$).
while $t = 0$ or $\mathrm{Err}_{\mathrm{local}}(X_{t-1} \mid p, \delta) - \mathrm{Err}_{\mathrm{local}}(X_t \mid p, \delta) \geq \varepsilon$ do
    $t := t + 1$.
    Set $Y := X_{t-1}$.
    Compute $M$ and $H$.
    for $i = 1$ to $n$ do
        for $s = 1$ to $p$ do
            Update $x_{is}$ by the formula (4).
        end for
    end for
    Set $X_t := X$.
end while
```

5. Local ordinal embedding: consistency

In this section, we prove that local ordinal embedding is statistically consistent: in the large sample limit, it recovers the original point position up to a small error. This establishes that local ordinal information is indeed sufficient to reconstruct the geometry and density of a point set. To be able to state the theorem, we first need to introduce a distance function between sets of points X and Y:

$$\Delta(X, Y) := \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{y}_i)^T(\boldsymbol{x}_i - \boldsymbol{y}_i)$$
$$\Delta_{sim}(X, Y) := \inf_{\substack{a>0,\boldsymbol{b}\in\mathbb{R}^p, \\ O:\,\text{orthonormal}}} \Delta(X, a \cdot OY + \boldsymbol{1}\boldsymbol{b}^T),$$

where $\boldsymbol{1} = (1, \dots, 1)^T$ is the p-dimensional one-vector.

Theorem 3 (Consistency of LOE)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^p$ is compact, connected, convex, has a smooth boundary, and has full dimensionality in the sense that there exists some $\varepsilon > 0$ such that the set $\mathcal{X}_\varepsilon := \{x \in \mathcal{X} \mid d(x, \partial\mathcal{X}) > \varepsilon\}$ is nonempty and connected. Let f be a probability density function with support $\mathcal{X}$. We assume that f is continuously differentiable and is bounded away from 0. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an i.i.d. sample from f, and $\hat{\boldsymbol{X}}_1, ..., \hat{\boldsymbol{X}}_n \in \mathbb{R}^p$ be a global optimum of the LOE objective function* $\mathrm{Err}_{\mathrm{local}}$*. Then, as $n \to \infty$, $k \to \infty$ such that $k/n \to 0$ and $k^{p+2}/(n^2 \log^p n) \to \infty$, we have that $\Delta_{sim}(X, \hat{X}_{\mathrm{LOE}}) \to 0$ in probability.*

Proof sketch. In Proposition 4 below we prove that upon knowledge of the unweighted, directed kNN graph it is possible to consistently estimate all pairwise Euclidean distances $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|$, up to a global multiplicative constant and a small additive error ε. Consequently, the distance matrix of any point configuration $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ that has the same kNN graph as the original set $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$, has to agree with the original distance matrix (up to error ε and a global constant). But it is well known in the context of classic multidimensional scaling that if two distance matrices agree up to entry-wise deviations of ε, then the Δ_{sim}-distance between the two corresponding point configurations is bounded by ε^2 times a constant (Sibson, 1979). Taken together, any point configuration $\hat{\boldsymbol{X}}_1, ..., \hat{\boldsymbol{X}}_n$ that is a solution of LOE has to agree with the original point set, up to similarity transforms and an error converging to 0. □

The key ingredient in this proof is the following statement.

Proposition 4 (Estimating distances from kNN)**.** *Under the assumptions of Theorem 3, consider the unweighted, directed kNN graph on the sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. Then we can construct estimates $\hat{d}_{ij}$ such that with probability $1 - p_n$ the following holds: There exists an unknown constant C_n,* *a scaling constant $s_n > 0$ and an $\varepsilon_n > 0$ such that for all $i, j \in \{1, ..., n\}$, we have $|C_n \cdot s_n \cdot \hat{d}_{ij} - \|\boldsymbol{X}_i - \boldsymbol{X}_j\|\,| \le \varepsilon_n$. In particular, if $n \to \infty$, $k \to \infty$, $k/n \to 0$, and $k^{p+2}/(n^2 \log^p n) \to \infty$, then $\varepsilon_n \to 0$ and $p_n \to 0$, hence the distance estimates converge to the true Euclidean distances uniformly in probability.*

Proof sketch. It has been proved in von Luxburg & Alamgir (2013) that if we are given an unweighted, directed k-nearest neighbor graph on a sample of points, then under the conditions stated in the proposition it is possible to consistently estimate the underlying density $f(\boldsymbol{X}_i)$ at each data point, that is there exist estimates $\hat{f}_i$ such that $\hat{f}_i \to f(\boldsymbol{X}_i)$ a.s., uniformly over all sample points. We now use these estimates to assign edge weights to the previously unweighted kNN graph: if edge (i, j) exists, it gets the weight $r_{n,k}(i) := (1/\hat{f}_i)^{1/p}$. The key is now to prove that the shortest path distances in the re-weighted kNN graph converge to the underlying Euclidean distances. Assume first that we knew the underlying density values, that is $\hat{f}_i = f(\boldsymbol{X}_i)$. Then under the conditions on n and k stated in the proposition, the distance between a point $\boldsymbol{X}_i$ and its k-nearest neighbor is highly concentrated around its expectation, and this expectation is proportional to $r_{n,k}(i)$. To see that $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|$ is lower bounded by the rescaled shortest path distance between i and j, we take the straight line between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ and chop it into small pieces $[\boldsymbol{a}_l, \boldsymbol{a}_{l+1}]$ of length proportional to $r_{n,k}$ (this length varies with the density as we go along the line). Now we replace each of the intermediate points $\boldsymbol{a}_l$ by its closest sample point $\boldsymbol{b}_l$. With some care we can ensure that $\boldsymbol{b}_l$ is connected to $\boldsymbol{b}_{l+1}$ in the graph, and in this case the length of the path $\boldsymbol{X}_i, \boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{X}_j$ is an upper bound for the rescaled shortest path distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in the re-weighted graph. The other way round, consider a shortest path in the re-weighted graph. It is straightforward to see that its length is approximately proportional to the sum of Euclidean distances between subsequent vertices, which is an upper bound on the Euclidean distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. The same analysis holds if $\hat{f}_i$ does not coincide with $f(\boldsymbol{X}_i)$, but consistently converges to $f(\boldsymbol{X}_i)$ up to a constant. In this case, the matrix of pairwise shortest path distances approximates a constant times the original Euclidean distance matrix. □

For an illustration of Proposition 4, we provide a number of simulations in the supplementary material. They show the convergence behavior of the shortest path distance in the re-weighted graph.

Choice of k. Theorem 3 states that statistical consistency of local ordinal embedding occurs if $k/n \to 0$ but $k/n^{2/(p+2)} \to \infty$ (ignoring log factors). This requirement is inherited from von Luxburg & Alamgir (2013), but we believe that this condition on k can be significantly lowered to k some power of $\log n$ (this is ongoing work).

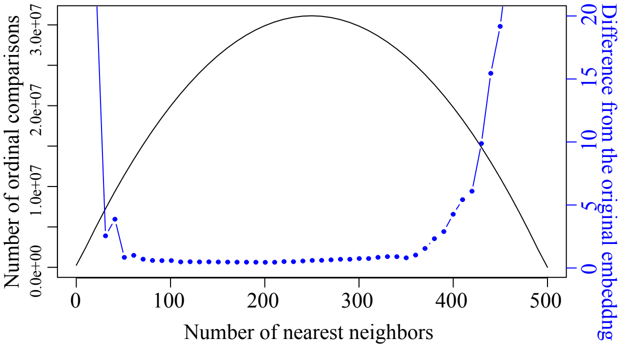


Figure 1. Relationship between the number of constraints and the embedding error, on the data described in Section 6.2 (with $n = 500$). The "difference from the original embedding" refers to the squared Frobenius distance between the original coordinate matrix and the procrustes transformed LOE embedding.

A natural question is whether the quality of the embedding increases dramatically if we increase k. Note that the number of ordinal constraints of the type ($\star\star$) that are encoded in a kNN graph is $nk(n - k)$, which takes its maximum for $k = n/2$. However, Figure 1 shows empirically that once we passed some reasonably small value of k, the error of the embedding stays about the same for a wide range of k, and only increases when k gets extremely large again. Further figures illustrating the behavior of LOE with respect to the choice of k can be found in the supplementary material.

We should contrast our number of constraints with the results in Jamieson & Nowak (2011a). Here the authors showed that at least $\Omega(n \log n)$ *actively chosen* queries are necessary to uniquely determine a point embedding. They conjecture that there is also a matching upper bound. In our case, however, we are not interested in the case of arbitrary comparisons of type ($\star$), but in the particular case of *local* comparisons of type ($\star\star$). So even if their conjecture turns out to be true, this is not in conflict with our results.

6. Experiments with local ordinal embedding

In our experiments we focus on the case of local ordinal embedding (experiments for more general soft ordinal embedding are provided in the supplementary material).

6.1. Evaluation criterion: Graph adjusted rand index

To measure a recovery rate of ordinal information in an unweighted graph, we need an appropriate criterion. Let $A_n := (a_{ij})_{n \times n}$ be a given adjacency matrix and $\hat{A}_n := (\hat{a}_{ij})_{n \times n}$ be a recovered adjacency matrix. The naive approach would be to consider the error function

$$\mathrm{Err}(\hat{A}_n, A_n) := \frac{1}{n(n-1)} \|\hat{A}_n - A_n\|_F^2.$$

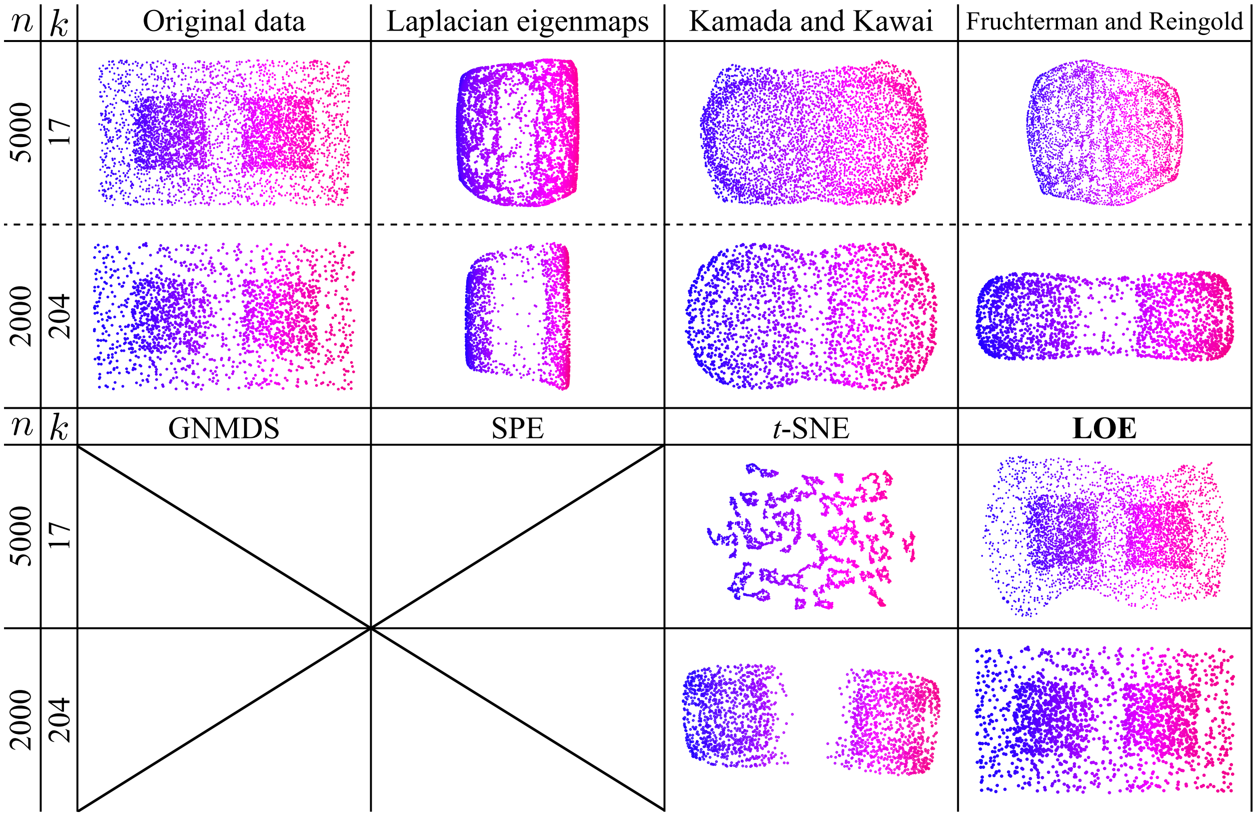


Figure 2. Two-dimensional embeddings of different methods in the realizable case. Our LOE algorithm is the only one that captures the density information of the original data.

		$\hat{a}_{ij}$		Total row
		1	0	
a_{ij}	1	m_i	$k_i - m_i$	k_i
	0	$k_i - m_i$	$n-1-2k_i+m_i$	$n-1-k_i$
Total col		k_i	$n-1-k_i$	$n-1$

Table 1. Contingency table of the i-th rows of A_n and $\hat{A}_n$. Used for deriving the graph adjusted rand index.

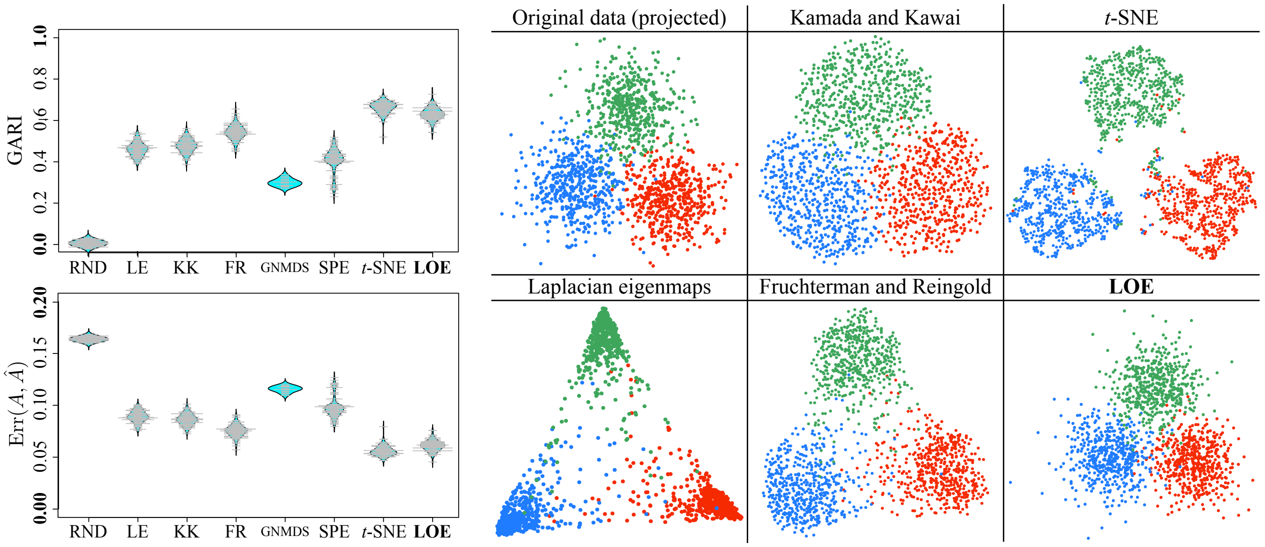
However, this function is unsuitable if n is large and $k = o(n)$, because then $\mathrm{Err}(\hat{A}_n, A_n) \le \frac{2k}{n-1} \to 0$ as $n \to \infty$. Thus, we introduce an adjusted recovery measure for an unweighted graph, called *graph adjusted rand index.* Let k_i be the out-degree of vertex i and $m_i := \#\{j \mid a_{ij} = 1$ and $\hat{a}_{ij} = 1\}$. Consider the contingency table presented as Table 1. As with Hubert & Arabie (1985), for each i we assume m_i is drawn from a hypergeometric distribution, that is $\hat{a}_{ij}$ takes 1 or 0 randomly such that k_i is fixed. Under this assumption, we have $\mathbb{E}[m_i] = k_i^2/(n-1)$. For $M_i := (n-1) - 2(k_i - m_i)$, we have $\mathbb{E}[M_i] = (n-1) + 2k_i(k_i - n + 1)/(n-1)$. We define the *graph adjusted rand index* $\mathrm{GARI}(A_n, \hat{A}_n)$ between A_n and $\hat{A}_n$ as

$$\mathrm{GARI}(A_n, \hat{A}_n) := \frac{\sum_{i=1}^{n}(\hat{M}_i - \mathbb{E}[M_i])}{\sum_{i=1}^{n}(\max M_i - \mathbb{E}[M_i])}$$

where $\hat{M}_i := \sum_{j \neq i} \mathbb{1}[a_{ij} = \hat{a}_{ij}]$. GARI is bounded from above by 1, and $\mathrm{GARI}(A_n, \hat{A}_n) = 1 \iff A_n = \hat{A}_n$. A high GARI score implies that many of the ordinal constraints have been satisfied by the solution. Note, however, that GARI does not take into account the amount by which a constraint is violated. We will see below that there exist embeddings that have a high GARI score, but do not preserve the density information. In this sense, a high GARI score is a necessary, but not a sufficient criterion for a good ordinal embedding.

6.2. kNN graph embedding in the realizable case

We first consider a simple case in which a perfect embedding to $\mathbb{R}^2$ exists. We sampled n points in $\mathbb{R}^2$ from a distribution that has two uniform high-density squares, surrounded by a uniform low density region. See Figure 2 (upper left). We then constructed the **unweighted** kNN graph and embedded this graph in $\mathbb{R}^2$ by various embedding methods, see Figure 2. We compare our approach to Laplacian eigenmaps (LE), the Kamada and Kawai algorithm (KK), the Fruchterman Reingold algorithm (FR), and t-distributed stochastic neighbor embedding (t-SNE). We also wanted to compare it to generalized non-metric scaling



(a) (b)

Figure 3. (a) Beanplots of GARI (high is good) and $\mathrm{Err}(\hat{A}_n, A_n)$ (low is good) for 2-dimensional embeddings of each method with 100 datasets, (b) Two-dimensional embeddings of 5 methods for a unweighted k-NN graph with $n = 1500$.

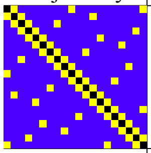
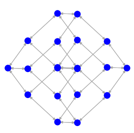
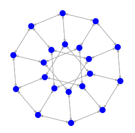
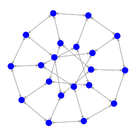
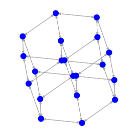
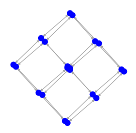
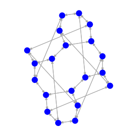
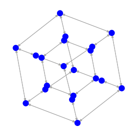
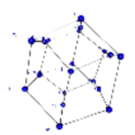
Adjacency	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE-2D**	**LOE-3D**
								
GARI	0.29	0.41	0.39	0.49	0.29	**0.68**	0.33	**1.00**
$\mathrm{Err}_{\mathrm{local}}$	4.10	4.22	4.21	9.50	4.30	5.35	**3.58**	**0.00**

Figure 4. 2-dimensional embeddings of each method and a 3-dimensional perfect embedding (LOE-3D) for the Desargues graph.

(GNMDS) and structure preserving embedding (SPE), but those two algorithms could not cope with this sample size for computational reasons. Even for moderate 500 sample points, GNMDS and SPE failed due to out of memory on a 3 GHz Intel core i7 with 8 GB of memory (we provide their results on smaller data sets in the supplementary material). In Figure 2 we can see that while most of the methods get the rough point layout correct, LOE is the only method that is capable to capture the original density and geometric structure of the data. For the Kamada and Kawai algorithm we can give a theoretical explanation for the distortion. This algorithm tries to find an embedding such that the Euclidean distances between the embedded vertices correspond to the shortest path distances in the graph. However, as has been proved in Alamgir & von Luxburg (2012), the shortest path distances in unweighted kNN graphs do not converge to the Euclidean distances — to the opposite, any embedding based on shortest path distances generates an embedding that distributes points as uniformly as possible (Alamgir et al., 2014).

6.3. kNN graph embedding in the non-realizable case

Next we consider a higher dimensional Gaussian mixture model with three components. We define the mean vectors of the three components as $\boldsymbol{\mu}_l := A\boldsymbol{c}_l$ $(l = 1, 2, 3)$, where

$$\boldsymbol{c}_1 := \left(\frac{4}{\sqrt{3}}, 0\right), \ \boldsymbol{c}_2 := \left(-\frac{4}{2\sqrt{3}}, \frac{4}{2}\right), \ \boldsymbol{c}_3 := \left(-\frac{4}{2\sqrt{3}}, -\frac{4}{2}\right)$$

and A is a random $p \times 2$ orthonormal matrix. The points $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{ip}]^T$ $(i = 1, \ldots, n)$ are generated as $\boldsymbol{X}_i := \sum_{l=1}^{3} u_{il}\boldsymbol{\mu}_l + \boldsymbol{\varepsilon}_{il}$, where $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})$ and $\boldsymbol{\varepsilon}_{ik}$ are independently generated from the multinomial distribution for three trials with equal probabilities and the p-dimensional standard normal distribution $N_p(\boldsymbol{0}, I_p)$, respectively. Based on these observations, we then construct the unweighted kNN graph. We set the true number of nearest neighbors $k \approx 2 \log n$, the number of original dimensions $p = 5$. In order to run a more thorough statistical evaluation, we chose the small sample size $n = 90$ $(k = 8)$ and constructed 100 such unweighted kNN graphs. To these graphs, we applied various embedding methods: LOE, a random embedding (RND) which just uses sample points from the standard normal distribution, Laplacian eigenmaps (LE), Kamada and Kawai algorithm (KK), Fruchterman Reingold algorithm (FR), generalized non-metric multidimensional scaling (GNMDS), structure preserving embedding (SPE), and t-distributed stochastic neighbor embedding (t-SNE). To choose the tuning parameter λ for GNMDS, we tried the candidate values $\{0.5, 1, 2, \dots, 100\}$ and chose the one that leads to the best GARI value. This means that we had to run GNMDS 101 times for each data set. It took approximately 2 months to get the solutions of GNMDS for 10 data sets, after which we stopped (in this experiment, GNMDS was performed on a 1.9 GHz Intel core i5 with 8 GB of memory). For SPE, we allowed the original algorithm to select the tuning parameter C. Figure 3(a) shows beanplots of the GARI scores for 2-dimensional embeddings. In this figure, LOE and t-SNE perform best. However, the GARI scores do not tell the whole story as they do not evaluate the actual distortion, but just the number of violated ordinal constraints. To investigate the preservation of density information, we settle on the larger sample size of $n = 1500$ ($k = 14 \approx 2 \log n$) and compare various embeddings in Figure 3(b). The top left figure shows the original data, projected on the space spanned by the mean vectors. It is obvious that while most methods do something reasonable, LOE is the only method that is able to recover the Gaussian density structure of the data. This finding also indicates that the GARI score alone is not enough to evaluate quality of embeddings. To compare the computational costs of the methods, we measured the required time for 50 iterations of each algorithm, for each parameter choice. This experiment was performed on a 3 GHz Intel core i7 with 8 GB of memory. The results are depicted in Figure 5. It is obvious that the two semi-definite programming methods, GNMDS and SPE, are way too expensive. On the other hand, the standard graph embedding algorithms such as LE, KK, FR, t-SNE are pretty fast. The methods based on LOE are in the intermediate range. More specifically, we additionally compared three different methods to minimize our objective functions: LOE-SD (the steepest descent algorithm, as implemented in C and R, with the step size 1 and the rate parameter of the backtracking line search 0.5); LOE-BFGS (a Newton-like algorithm, the Broyden-Fletcher-Goldfarb-Shanno algorithm implemented as the optim function in the R stats package), and our majorization algorithm LOE-MM. Figure 5 shows that our majorization algorithm is the fastest among these three.

6.4. Further simulations: graph drawing and SOE

We also applied our algorithm to standard graph-drawing tasks. Figure 4 shows one example, the classic Desargues graph. The 3-dimensional LOE-embedding can perfectly recover the original graph structure. It seems that the 2-dimensional embedding of LOE is a projection of the 3-dimensional one. We provide more graph drawing examples in the supplementary material. The general bottom line is that LOE outperforms competing algorithms in many cases. Finally, we also ran experiments with the general SOE approach, with varying amounts of ordinal constraints. The results can be found in the supplement.

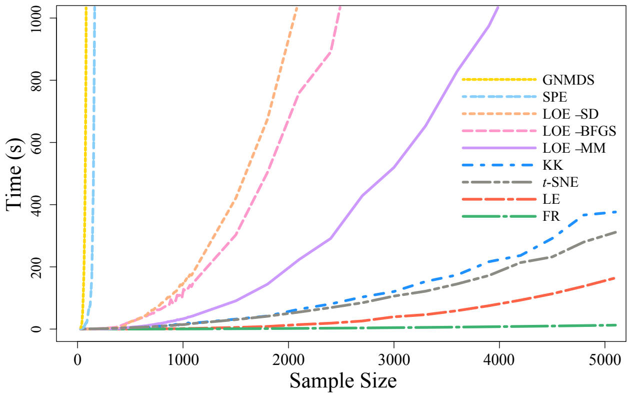


Figure 5. Running times of various methods, cf. Section 6.3.

7. Conclusion and Discussion

In this paper we suggest a new soft objective function for ordinal embedding. It not only takes into account the number of ordinal constraints that are violated, but the actual amount of distance by which these constraints are violated. Optimizing this objective function leads to an ordinal embedding algorithm that is able to recover the density structure of the underlying data set in a much better way than many other methods. Our approach for optimizing this objective function is based on majorizing functions and has been published as an R-package (Terada & von Luxburg, 2014). As a second contribution, we prove that ordinal embedding is even possible if not all ordinal constraints are given, but we just get to know the indices of the k nearest neighbors of each data point. This theoretical insight is new. As a special case of local ordinal embedding, we consider the problem of graph embedding. We ran extensive simulations to compare our algorithms to its competitors (the main paper and the supplementary material). They show that our method is very good at discovering the density structure of data sets and the geometric structure of graphs.

Acknowledgments

Y.T. was partially supported by Grant-in-Aid for JSPS Fellows Number 24 · 2466. U.v.L. was supported by the German Research Foundation (grant LU1718/1-1 and Research Unit 1735 ”Structural Inference in Statistics: Adaptation and Efficiency”).

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. Generalized non-metric multidimensional scaling. In *AISTATS*, pp. 11–18, 2007.

Ailon, N. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *J. Mach. Learn. Res.*, 13:137–164, 2012.

Alamgir, M. and von Luxburg, U. Shortest path distance in random k-nearest neighbor graphs. In *ICML*, pp. 1031–1038, 2012.

Alamgir, M., Lugosi, G., and von Luxburg, U. Density-preserving quantization with application to graph downsampling. In *Conference of Learning Theory (COLT)*, 2014.

Alon, N., Bădoiu, M., Demaine, E., Farach-Colton, M., Hajiaghayi, M., and Sidiropoulos, A. Ordinal embeddings of minimum relaxation: general properties, trees, and ultrametrics. *ACM Trans. Alg.*, 4(4):46:1–46:21, 2008.

Bartal, Y., Recht, B., and Schulman, L. Dimensionality reduction: beyond the Johnson-Lindenstrauss bound. In *SODA*, pp. 868–887, 2011.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6): 1373–1396, 2003.

Bilu, Y. and Linial, N. Monotone maps, sphericity and bounded second eigenvalue. *J. Combin. Theory Ser. B*, 95(2):283–299, 2005.

Borg, I. and Groenen, P. J. F. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

Dattorro, J. *Convex optimization and Euclidean distance geometry*. Meboo Publishing, 2005.

Fruchterman, T. and Reingold, E. Graph drawing by force-directed placement. *Software: Practice and experience*, 21 (11):1129–1164, 1991.

Gottlieb, L. and Krauthgamer, R. A nonlinear approach to dimension reduction. In *SODA*, pp. 888–899, 2011.

Groenen, P. J. F., Winsberg, S., Rodíguez, O., and Diday, E. I-scal: Multidimensional scaling of interval dissimilarities. *Comput. Stat. Data Anal.*, 51(1):360–378, 2006.

Hinton, G. E. and Roweis, S. T. Stochastic neighbor embedding. In *NIPS*, pp. 833–840, 2002.

Hubert, L. and Arabie, P. Comparing partitions. *J. Classification*, 2(1):193–218, 1985.

Jamieson, K. and Nowak, R. Low-dimensional embedding using adaptively selected ordinal data. In *Conference on Communication, Control, and Computing*, pp. 1077–1084, 2011a.

Jamieson, K. and Nowak, R. Active ranking using pairwise comparisons. In *NIPS*, pp. 2240–2248, 2011b.

Johnson, R. M. Pairwise non metric multidimensional scaling. *Psychometrika*, 38(1):11–18, 1973.

Kamada, T. and Kawai, S. An algorithm for drawing general undirected graphs. *Inform. Process Lett.*, 31(1):7–15, 1989.

Kleindessner, M. and von Luxburg, U. Uniqueness of ordinal embedding. In *Conference of Learning Theory (COLT)*, 2014.

Kruskal, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964a.

Kruskal, J. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964b.

Kruskal, J. *How to Use M-D-SCAL: A Program to do Multidimensional Scaling and Multidimensional Unfolding*. Bell Telephone Laboratories, Murray Hill, New Jersey, 1968.

Lan, Y., Guo, J., Cheng, X., and Liu, T. Statistical consistency of ranking methods in a rank-differentiable probability space. In *NIPS*, pp. 1241–1249, 2012.

McFee, B. and Lanckriet, . Learning multi-modal similarity. *J. Mach. Learn. Res.*, 12:491–523, 2011.

McFee, B. and Lanckriet, G. Partial order embedding with multiple kernels. In *ICML*, pp. 721–728, 2009.

McFee, B. and Lanckriet, G. Metric learning to rank. In *ICML*, pp. 775–782, 2010.

Ouyang, H. and Gray, A. Learning dissimilarities by ranking: from SDP to QP. In *ICML*, pp. 728–735, 2008.

Quist, M. and Yona, G. Distributional scaling: An algorithm for structure-preserving embedding of metric and nonmetric spaces. *J. Mach. Learn. Res.*, 5:399–420, 2004.

Rosales, R. and Fung, G. Learning sparse metrics via linear programming. In *KDD*, pp. 367–373, 2006.

Roweis, S. and Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323 – 2326, 2000.

Shaw, B. and Jebara, T. Structure preserving embedding. In *ICML*, pp. 937–944, 2009.

Shepard, R. The analysis of proximities: Multidimensional scaling with an unknown distance function (I). *Psychometrika*, 27 (2):125–140, 1962a.

Shepard, R. The analysis of proximities: Multidimensional scaling with an unknown distance function (II). *Psychometrika*, 27 (3):219–246, 1962b.

Shepard, R. Metric structures in ordinal data. *J. Math. Psych.*, 3 (2):287–315, 1966.

Sibson, R. Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *J. Roy. Statist. Soc. Ser. B*, 41(2):217–229, 1979.

Tamassia, R. (ed.). *Handbook of graph drawing and visualization*. CRC Press, 2013.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. Adaptively learning the crowd kernel. In *ICML*, pp. 673–680, 2011.

Tenenbaum, J., de Silva, V., and Langford, J. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319 – 2323, 2000.

Terada, Y. and von Luxburg, U. loe: Local ordinal embedding, 2014. URL `http://cran.r-project.org/web/packages/loe/index.html`.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008.

von Luxburg, U. and Alamgir, M. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *NIPS*, pp. 225–233, 2013.

Wauthier, F., Jordan, M., and Jojic, N. Efficient ranking from pairwise comparisons. In *ICML*, pp. 109–117, 2013.