Supplementary Material for 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

In this supplementary material, we provide the proof of Proposition 1 and the details about the derivation of the majorizing function in Proposition 2. Moreover, a number of additional simulations are described.

1. Proof of Proposition 1

Proposition 1 (Scale parameter)**.** *Let* $\delta_1,\ \delta_2 > 0$*,* $\delta_1 \neq \delta_2$*. 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* $\operatorname{argmin} \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_2)$ *with parameter* δ_2*.*

Proof. For any constant $c > 0$, we obviously have $c^2 \,\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta) = \mathrm{Err}_{\mathrm{soft}}(cX \mid p, c\delta)$ for all $X \in \mathbb{R}^{n\times p}$. Thus, for all $X \in \mathbb{R}^{n\times p}$, we have $(\delta_2/\delta_1)^2 \,\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_1) = \mathrm{Err}_{\mathrm{soft}}((\delta_2/\delta_1)X \mid p, \delta_2)$. To obtain a contradiction, we assume $(\delta_2/\delta_1)X_{\delta_1}$ does not achieve the minimal value of $\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_2)$, i.e.,

$$\min_{X\in\mathbb{R}^{n\times p}} \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_2) < \mathrm{Err}_{\mathrm{soft}}((\delta_2/\delta_1)X_{\delta_1} \mid p, \delta_2).$$

For all X_1 and X_2 satisfying $\mathrm{Err}_{\mathrm{soft}}(X_1 \mid p, \delta_1) < \mathrm{Err}_{\mathrm{soft}}(X_2 \mid p, \delta_1)$ and all $c > 0$, we have

$$c^2 \,\mathrm{Err}_{\mathrm{soft}}(X_1 \mid p, \delta_1) < c^2 \,\mathrm{Err}_{\mathrm{soft}}(X_2 \mid p, \delta_1).$$

Then, we have

$$\begin{aligned}
\mathrm{Err}_{\mathrm{soft}}((\delta_2/\delta_1)X_{\delta_1} \mid p, \delta_2) &= \left(\frac{\delta_2}{\delta_1}\right)^2 \mathrm{Err}_{\mathrm{soft}}(X_{\delta_1} \mid p, \delta_1) \\
&> \min_{X\in\mathbb{R}^{n\times p}} \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_2) \\
&= \left(\frac{\delta_2}{\delta_1}\right)^2 \min_{X\in\mathbb{R}^{n\times p}} \mathrm{Err}_{\mathrm{soft}}((\delta_1/\delta_2)X \mid p, \delta_1),
\end{aligned}$$

which contradicts $X_{\delta_1} := \arg\min \mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta_1)$. □

2. Majorizing function and Majorization algorithm for SOE

First, we provide the detailed derivation of the majorizing function of the components of $\mathrm{Err}_{\mathrm{soft}}$ (i.e., the proof of Proposition 2) and then we describe a majorization algorithm for SOE.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

2.1. Majorizing the components of the objective function $\mathrm{Err}_{\mathrm{soft}}$

To majorize $\max\left[0,\ d_{ij}(X)+\delta-d_{kl}(X)\right]^2$, we use the following majorization inequality (Groenen et al., 2006),

$$\max[0,\ a_1-a_2]^2 \leq \begin{cases} 2(a_1^2+a_2^2)-2(a_1+a_2)(b_1+b_2)+(b_1+b_2)^2 & \text{if } b_1 \geq b_2, \\ 2(a_1^2+a_2^2)-4a_1b_1-4a_2b_2+2(b_1^2+b_2^2) & \text{if } b_1 < b_2. \end{cases} \tag{1}$$

By this inequality, we have

$$\begin{aligned} &\max\left[0,\ d_{ij}(X)+\delta-d_{kl}(X)\right]^2 \leq \\ &\begin{cases} 2\left[(d_{ij}(X)+\delta)^2+d_{kl}^2(X)\right]-2(d_{ij}(Y)+d_{kl}(Y)+\delta)(d_{ij}(X)+d_{kl}(X)+\delta) & \\ +(d_{ij}(Y)+d_{kl}(Y)+\delta)^2 & \text{if } d_{ij}(Y)+\delta \geq d_{kl}(Y), \\ 2\left[(d_{ij}(X)+\delta)^2+d_{kl}^2(X)\right]-4(d_{ij}(Y)+\delta)(d_{ij}(X)+\delta)-4d_{kl}(Y)d_{kl}(X) & \\ +2\left[(d_{ij}(Y)+\delta)^2+d_{kl}^2(Y)\right] & \text{if } d_{ij}(X)+\delta < d_{kl}(X). \end{cases} \end{aligned} \tag{2}$$

To obtain the final majorizing function, we majorize each term in the right side of the inequality.

2.1.1. Case I: $d_{ij}(Y)+\delta \geq d_{kl}(Y)$

In the case that $d_{ij}(Y)+\delta \geq d_{kl}(Y)$, we have

$$\begin{aligned} \max\left[0,\ d_{ij}(X)+\delta-d_{kl}(X)\right]^2 \leq & 2d_{ij}^2(X)+2d_{kl}^2(X)-2(d_{ij}(Y)+d_{kl}(Y)-\delta)d_{ij}(X) \\ & -2(d_{ij}(Y)+d_{kl}(Y)+\delta)d_{kl}(X)+(d_{ij}(Y)+d_{kl}(Y)+\delta)(d_{ij}(Y)+d_{kl}(Y)-\delta) \\ & +2\delta^2. \end{aligned}$$

Due to the presence of the terms $d_{ij}(X)$ and $d_{kl}(X)$, the last term of the inequality is not quadratic in X. Note that $d_{ij}(Y)+d_{kl}(Y)+\delta>0$. By the Cauchy-Schwarz inequality

$$-d_{kl}(X) \leq \begin{cases} -\frac{1}{d_{kl}(Y)}(\boldsymbol{x}_k-\boldsymbol{x}_l)^T(\boldsymbol{y}_k-\boldsymbol{y}_l) & \text{if } d_{kl}(Y)>0, \\ 0 & \text{if } d_{kl}(Y)=0. \end{cases} \tag{3}$$

We have a linear majorizing function of $-2(d_{ij}(Y)+d_{kl}(Y)+\delta)d_{kl}(X)$. Unfortunately, we cannot apply the inequality (3) for $-2(d_{ij}(Y)+d_{kl}(Y)-\delta)d_{ij}(X)$ directly since it is not always true that $d_{ij}(Y)+d_{kl}(Y)-\delta>0$. Thus, we consider the case $d_{ij}(Y)+d_{kl}(Y)-\delta\geq 0$ and the case $d_{ij}(Y)+d_{kl}(Y)-\delta<0$ separately.

If $d_{ij}(Y)+d_{kl}(Y)-\delta\geq 0$, we can apply the inequality (3) and obtain a linear majorizing function. Then,

$$\begin{aligned} &\max\left[0,\ d_{ij}(X)+\delta-d_{kl}(X)\right]^2 \\ &\leq \begin{cases} 2d_{ij}^2(X)+2d_{kl}^2(X)-2\frac{d_{ij}(Y)+d_{kl}(Y)-\delta}{d_{ij}(Y)}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{y}_i-\boldsymbol{y}_j) & \\ -2\frac{d_{ij}(Y)+d_{kl}(Y)+\delta}{d_{kl}(Y)}(\boldsymbol{x}_k-\boldsymbol{x}_l)^T(\boldsymbol{y}_k-\boldsymbol{y}_l)+\left(d_{ij}(Y)+d_{kl}(Y)\right)^2+\delta^2 & \text{if } d_{ij}(Y)>0 \text{ and } d_{kl}(Y)>0, \\ 2d_{ij}^2(X)+2d_{kl}^2(X)-2\frac{d_{ij}(Y)+d_{kl}(Y)+\delta}{d_{kl}(Y)}(\boldsymbol{x}_k-\boldsymbol{x}_l)^T(\boldsymbol{y}_k-\boldsymbol{y}_l) & \\ +\left(d_{ij}(Y)+d_{kl}(Y)\right)^2+\delta^2 & \text{if } d_{ij}(Y)=0 \text{ and } d_{kl}(Y)>0, \\ 2d_{ij}^2(X)+2d_{kl}^2(X)-2\frac{d_{ij}(Y)+d_{kl}(Y)-\delta}{d_{ij}(Y)}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{y}_i-\boldsymbol{y}_j) & \\ +\left(d_{ij}(Y)+d_{kl}(Y)\right)^2+\delta^2 & \text{if } d_{ij}(Y)>0 \text{ and } d_{kl}(Y)=0, \\ 2d_{ij}^2(X)+2d_{kl}^2(X)+\left(d_{ij}(Y)+d_{kl}(Y)\right)^2+2\delta^2 & \text{if } d_{ij}=0 \text{ and } d_{kl}(Y)=0. \end{cases} \end{aligned} \tag{4}$$

On the other hand, if $d_{ij}(Y)+d_{kl}(Y)-\delta<0$, we have $-2(d_{ij}(Y)+d_{kl}(Y)-\delta)>0$. Thus, we can apply the following majorization inequality (Groenen et al., 2006)

$$d_{ij}(X) \leq \frac{1}{2}\frac{d_{ij}^2(X)}{d_{ij}(Y)}+\frac{1}{2}d_{ij}(Y). \tag{5}$$

Note that this inequality holds when $d_{ij}(Y) > 0$. If $d_{ij}(Y) = 0$, we should replace it by a small positive value, say τ. As Groenen et al. (2006) pointed out, by choosing τ small enough, this replacement should not have an effect on the convergence properties of the majorization algorithm. Using the inequality (5), we can obtain a quadratic majorizing function of $-2(d_{ij}(Y) + d_{kl}(Y) - \delta)d_{ij}(X)$. Thus, if $d_{ij}(Y) + d_{kl}(Y) - \delta < 0$, we have

$$\begin{aligned}
&\max\left[0,\ d_{ij}(X) + \delta - d_{kl}(X)\right]^2 \\
&\leq \begin{cases}
\frac{d_{ij}(Y)+\delta-d_{kl}(Y)}{d_{ij}(Y)} d_{ij}^2(X) + 2d_{kl}^2(X) - 2\frac{d_{ij}(Y)+d_{kl}(Y)+\delta}{d_{kl}(Y)}(\boldsymbol{x}_k - \boldsymbol{x}_l)^T(\boldsymbol{y}_k - \boldsymbol{y}_l) & \\
+(d_{ij}(Y) + d_{kl}(Y))d_{kl}(Y) + (d_{ij}(Y) + \delta)\delta & \text{if } d_{kl}(Y) > 0, \\
\frac{d_{ij}(Y)+\delta-d_{kl}(Y)}{d_{ij}(Y)} d_{ij}^2(X) + 2d_{kl}^2(X) + (d_{ij}(Y) + d_{kl}(Y))d_{kl}(Y) + (d_{ij}(Y) + \delta)\delta & \text{if } d_{kl}(Y) = 0.
\end{cases}
\end{aligned} \tag{6}$$

2.1.2. Case II: $d_{ij}(Y) + \delta < d_{kl}(Y)$

In the case that $d_{ij}(Y) + \delta < d_{kl}(Y)$, we have

$$\max\left[0,\ d_{ij}(X) + \delta - d_{kl}(X)\right]^2 \leq 2d_{ij}^2(X) + 2d_{kl}^2(X) - 4d_{ij}(Y)d_{ij}(X) - 4d_{kl}(Y)d_{kl}(X) + 2\left(d_{ij}^2(Y) + d_{kl}^2(Y)\right).$$

Since $d_{ij}(Y),\ d_{kl}(Y) > 0$, by the inequality (3) we have

$$\begin{aligned}
&\max\left[0,\ d_{ij}(X) + \delta - d_{kl}(X)\right]^2 \\
&\leq \begin{cases}
2d_{ij}^2(X) + 2d_{kl}^2(X) - 4(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{y}_i - \boldsymbol{y}_j) - 4(\boldsymbol{x}_k - \boldsymbol{x}_l)^T(\boldsymbol{y}_k - \boldsymbol{y}_l) + 2\left(d_{ij}^2(Y) + d_{kl}^2(Y)\right) & \text{if } d_{ij}(Y) > 0, \\
2d_{ij}^2(X) + 2d_{kl}^2(X) - 4(\boldsymbol{x}_k - \boldsymbol{x}_l)^T(\boldsymbol{y}_k - \boldsymbol{y}_l) + 2\left(d_{ij}^2(Y) + d_{kl}^2(Y)\right) & \text{if } d_{ij}(Y) = 0.
\end{cases}
\end{aligned} \tag{7}$$

2.1.3. Combining the majorization results

Combining the above results, we obtain the majorizing function in Proposition 2:

$$\begin{aligned}
o_{ijkl}\max\left[0,\ d_{ij}(X) + \delta - d_{kl}(X)\right]^2 \leq &\alpha_{ijkl}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + \alpha_{ijkl}^*\|\boldsymbol{x}_k - \boldsymbol{x}_l\| - 2\beta_{ijkl}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{y}_i - \boldsymbol{y}_j) \\
&- 2\beta_{ijkl}^*(\boldsymbol{x}_i - \boldsymbol{x}_k)^T(\boldsymbol{y}_i - \boldsymbol{y}_k) + \gamma_{ijkl},
\end{aligned} \tag{8}$$

where $\alpha_{ijkl}^* = 2o_{ijkl}$,

$$\alpha_{ijkl} = \begin{cases}
2o_{ijkl} & \text{if } d_{ij}(Y) \geq |d_{kl}(Y) - \delta| \text{ or } d_{kl}(Y) > d_{ij}(Y) + \delta, \\
o_{ijkl}\frac{d_{ij}(Y)+\delta-d_{kl}(Y)}{d_{ij}(Y)} & \text{if } d_{ij}(Y) + \delta \geq d_{kl}(Y) \text{ and } d_{ij}(Y) + d_{kl}(Y) < \delta,
\end{cases}$$

$$\beta_{ijkl} = \begin{cases}
o_{ijkl}\frac{d_{ij}(y)+d_{kl}(Y)-\delta}{d_{ij}(Y)} & \text{if } d_{ij}(Y) \geq |d_{kl}(Y) - \delta| \text{ and } d_{ij}(Y) > 0, \\
0 & \text{if } (d_{ij}(Y) + \delta \geq d_{kl}(Y) \text{ and } d_{ij}(Y) + d_{kl}(Y) < \delta) \text{ or } d_{ij}(Y) = 0, \\
2o_{ijkl} & \text{if } d_{ij}(Y) + \delta < d_{kl}(Y),
\end{cases}$$

$$\beta_{ijkl}^* = \begin{cases}
o_{ijkl}\frac{d_{ij}(y)+d_{kl}(Y)+\delta}{d_{kl}(Y)} & \text{if } d_{ij}(Y) \geq |d_{kl}(Y) - \delta| \text{ and } d_{kl}(Y) > 0, \\
0 & \text{if } d_{ij}(Y) \geq |d_{kl}(Y) - \delta| \text{ or } d_{kl}(Y) = 0, \\
2o_{ijkl} & \text{if } d_{ij}(Y) + \delta < d_{kl}(Y),
\end{cases}$$

and

$$\gamma_{ijkl} = \begin{cases}
o_{ijkl}[(d_{ij}(Y) + d_{kl}(Y))^2 + \delta^2] & \text{if } d_{ij}(Y) \geq |d_{kl}(Y) - \delta|, \\
o_{ijkl}[(d_{ij}(Y) + d_{kl}(Y))d_{kl}(Y) + (d_{ij}(Y) + \delta)\delta] & \text{if } d_{ij}(Y) + \delta \geq d_{kl}(Y) \text{ and } d_{ij}(Y) + d_{kl}(Y) < \delta, \\
2o_{ijkl}\left(d_{ij}^2(Y) + d_{kl}^2(Y)\right) & \text{if } d_{ij}(Y) + \delta < d_{kl}(Y).
\end{cases}$$

Note that this function is quadratic in X.

2.2. Majorization algorithm for SOE

Here, we describe the majorization algorithm for the general ordinal embedding problem. By inequality (8), we have

$$\begin{aligned}&\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta)\\ &\leq \sum_{i<j}^{n}\sum_{k<l}^{n}\Big[\alpha_{ijkl}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2+\alpha^*_{ijkl}\|\boldsymbol{x}_k-\boldsymbol{x}_l\|-2\beta_{ijkl}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{y}_i-\boldsymbol{y}_j)-2\beta^*_{ijkl}(\boldsymbol{x}_k-\boldsymbol{x}_l)^T(\boldsymbol{y}_k-\boldsymbol{y}_l)+\gamma_{ijkl}\Big].\end{aligned} \tag{9}$$

By reformulating the right side of (9), we obtain

$$\mathrm{Err}_{\mathrm{soft}}(X \mid p, \delta) \leq \sum_{s=1}^{p}\left[\boldsymbol{x}_s^T M^* \boldsymbol{x}_s - 2\boldsymbol{x}_s^T H^* \boldsymbol{y}_s\right] + \gamma, \tag{10}$$

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

$$m^*_{ij} = \begin{cases}\sum_{j'\neq i} m^*_{ij'} & \text{if } i = j,\\ -(\alpha_{ij\cdot\cdot} + \alpha^*_{\cdot\cdot ij}) & \text{if } i \neq j,\end{cases} \qquad h^*_{ij} = \begin{cases}\sum_{j'\neq i} \eta_{ij'} & \text{if } i = j,\\ -(\beta_{ij\cdot\cdot} + \beta^*_{\cdot\cdot ij}) & \text{if } i \neq j.\end{cases}$$

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 formula

$$x_{is} := \frac{\sum_{j=1}^{n} h_{ij}y_{js} - \sum_{j\neq i} m_{ij}x_{js}}{m_{ii}}.$$

Therefore, a majorization algorithm for minimizing Q is given by Algorithm 1. In this algorithm, the computational complexity of each iteration is $O(c)$ where c is the number of ordinal constraints (i.e., $c := \#(\mathcal{A})$).

Algorithm 1 SOE-MM: Majorization algorithm for SOE

Set $\delta > 0$ to a scale parameter and set X_0 to some initial coordinate matrix.
Set iteration counter $t := 0$.
Set $X_{-1} := X_0$.
Set $\varepsilon > 0$ to a small vale as the convergence criterion (e.g., $\varepsilon = 10^{-5}$).
while $t = 0$ or $\mathrm{Err}_{\mathrm{soft}}(X_{t-1} \mid p, \delta) - \mathrm{Err}_{\mathrm{soft}}(X_t \mid p, \delta) \geq \varepsilon$ **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**
      Compute

$$x_{is} := \frac{\sum_{j=1}^{n} h_{ij}y_{js} - \sum_{j\neq i} m_{ij}x_{js}}{m_{ii}}.$$

    end for
  end for
  Set $X_t := X$.
end while

3. Additional numerical experiments

3.1. A low sample size example in the realizable case

Section 6.2. of the main paper describes a realizable embedding problem. Because some of the algorithms could not cope with the large sample size, we additionally conducted experiments with a smaller sample size. To this end, we

constructed the unweighted kNN graph from the two-dimensional original data in Figure 1 (upper left) and embedded this graph in $\mathbb{R}^2$ by various embedding methods. We compare our approach to Laplacian eigenmaps (LE), the Kamada and Kawai algorithm (KK), the Fruchterman Reingold algorithm (FR), generalized non-metric scaling (GNMDS) with $\lambda = 0.05$, structure preserving embedding (SPE) and t-distributed stochastic neighbor embedding (t-SNE). In Figure 1 we can observe the same effect that also happened for larger sample size in the main paper: while most of the methods get the rough layout correct, LOE is the only method that is capable to capture the original density and geometric structure of the data.

3.2. Choice of k

Next, we describe the relationship between the number of constraints and the number of nearest neighbors. We sampled 500 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 LOE using the Laplacian eigenmaps as the initial embedding, see Figure 2. Figure 1 in the main paper shows the difference between the original embedding and the procrustes transformed LOE embedding with each k. In Figure 2, we show the actual LOE embeddings for a variety of values of k. The number of constraints can be considered as the amount of information that is given about the original density structure. The LOE embeddings around the most informative value $k = 250$ recover the original embedding nearly perfectly.

3.3. Some standard graph-drawing examples

Here, we describe applications for 6 classical unweighted graphs: the Frucht graph, the Icosahedral graph, the Chvatal graph, the Folkman graph, the Thomassen graph and the Meredith graph. The adjacency matrices of these graphs are available in the igraph package on R (Csárdi & Nepusz, 2006). To these graphs, we applied various embedding methods: LOE, 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). Here, to choose the tuning parameter λ for GNMDS, we also tried the candidate values $\{0.5, 1, 2, \ldots, 100\}$ and chose the one that leads to the best GARI value. Figure 3 shows the adjacency matrices (0 : blue and 1 : yellow) of the 6 graphs and embeddings of each method with the values of GARI and $\mathrm{Err}_{\mathrm{local}}$. We can see that LOE performs very nicely for most of these graph drawing tasks. Moreover, for the Desargues graph whose 2-dimensional embeddings are described in the main paper, Figure 3 shows 3-dimensional embeddings of each method with the values of GARI and $\mathrm{Err}_{\mathrm{local}}$.

3.4. A non-local embedding task solved by the general SOE algorithm

In this experiment, we consider a general, non-local embedding problem and apply our SOE algorithm. As initial data we use a matrix of "**driving**" distances between 21 cities in Europe. This data is available as 'eurodist' in the stats library of R. Note that there are no perfect embeddings in the 2-dimensional space. Among all ordinal distance comparisons (about 20000 comparisons in total), we randomly selected 1000 ordinal comparisons. We now compare our SOE algorithm to GNMDS. The parameter for GNMDS was set to $\lambda = 0.15$, which was selected as the value that provides the minimal number of violations of ordinal constraints. We compare SOE and GNMDS to two hard competitors: metric MDS, which is allowed to uses the actual distance scores between all cities, and standard non-metric MDS, which is allowed to use all ordinal constraints. Figure 4 shows the procrustes rotated embeddings of each of the methods. SOE provides a very similar embedding to the one of non-metric MDS with full ordinal embedding. The solution of GNMDS is somewhat different. The figure shows that SOE provides a much better embedding than GNMDS. This is also in accordance with the number of violations of constraints, which is $94/1000$ for GNMDS and $35/1000$ for SOE.

3.5. Illustration of the proof of Theorem 3

The main step in the consistency proof is Proposition 4: the density estimate of von Luxburg & Alamgir (2013) can be used to re-weight the graph, and the shortest path distances in this re-weighted graph converge to the true underlying Euclidean distances.

We demonstrate this convergence for a simple toy data set: a mixture of two Gaussians in $\mathbb{R}^2$, where the weight of one Gaussian is slightly larger than the weight of the other Gaussian (see Figure 5, left side). We draw $n = 1000$ points according to this distribution, build the unweighted kNN graph with $k = 50$, and estimate the density as described in

von Luxburg & Alamgir (2013). Then we re-weight the edges in the kNN graph with the weights $r_{n,k}$ as described in the proof of Proposition 4 of the main paper. In this re-weighted kNN graph we then compute the shortest path distances. In Figure 5 we depict these shortest path distances (right plot) and the true underlying Euclidean distances (middle plot). It is obvious that the distance structure is exactly the same.

Note that once we have an approximation of the full underlying Euclidean distance matrix, as provided in Proposition 4, we could simply use classic multidimensional scaling to embed the original graph to the Euclidean space. This embedding is going to be consistent, by a similar argument as the one of Theorem 3. On the same toy data as above, we show the result of this embedding in Figure 6. It is obvious that it gets the structure correct, which illustrates the relevance of Theorem 3. However, we do not recommend this approach for practice, because estimating the density to construct an embedding violates the principle that one never should solve too difficult problems as an intermediate step.

References

Csárdi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, Complex Systems: 1695, 2006. URL http://igraph.sf.net.

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

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

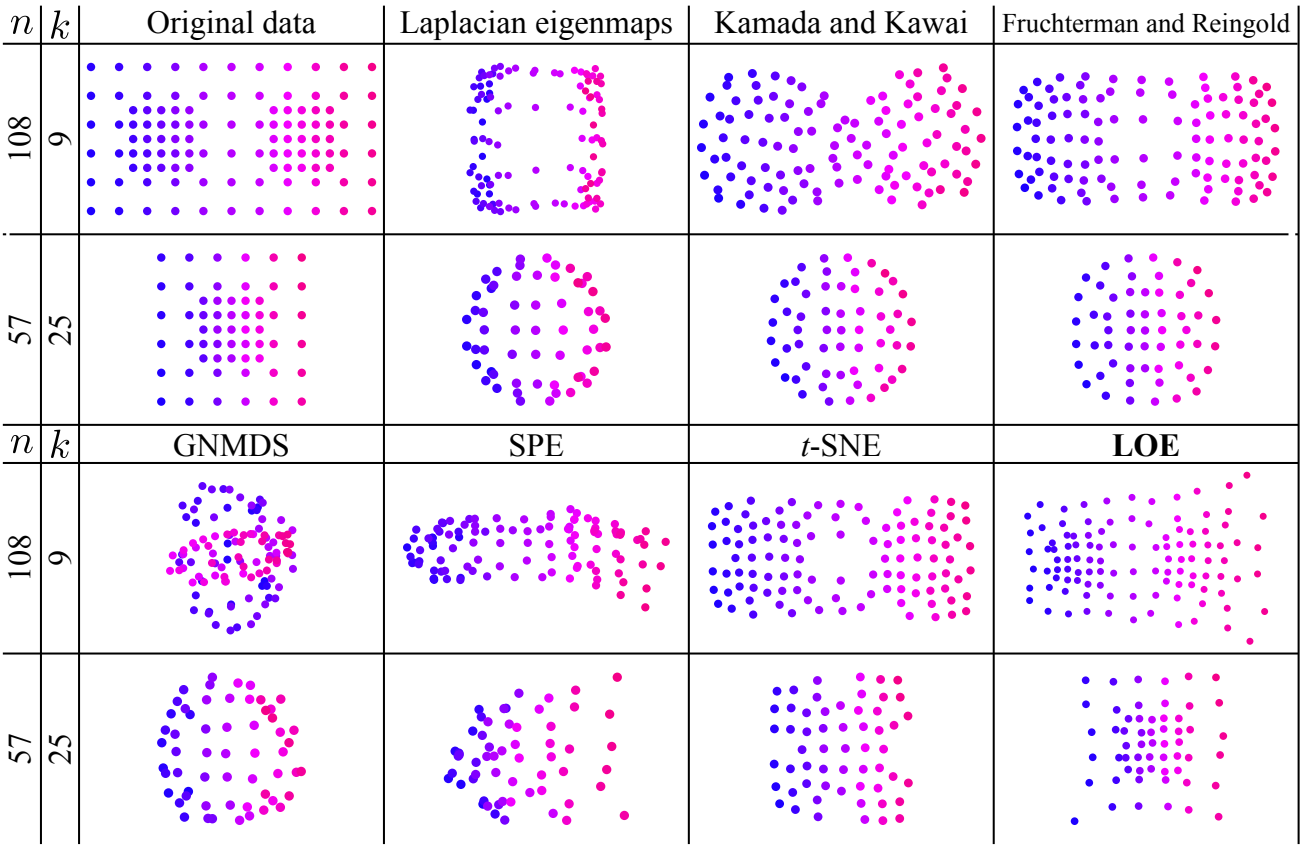


Figure 1: Two dimensional embeddings of 7 methods in the realizable and small sample case.

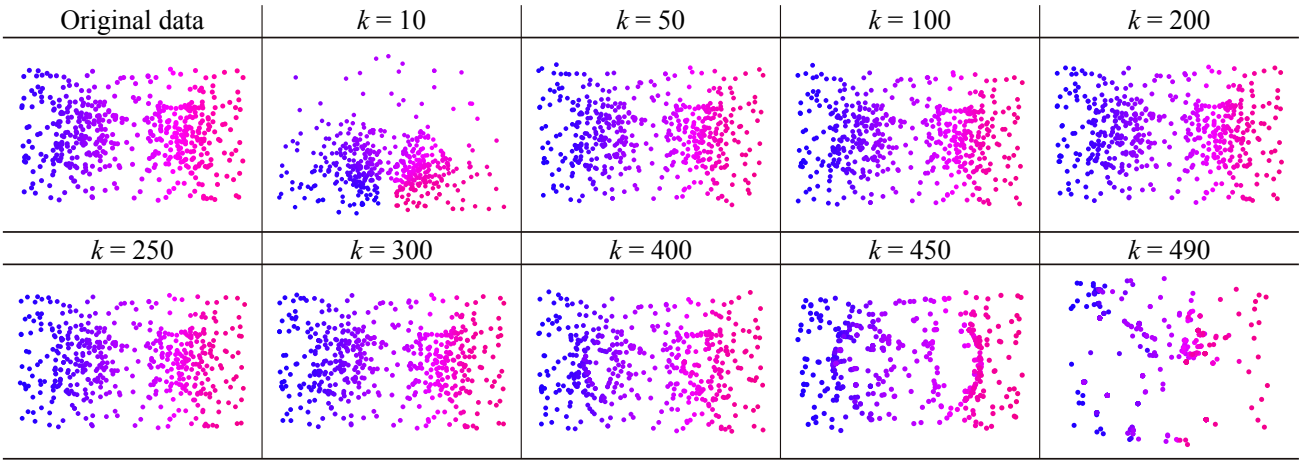


Figure 2: LOE embeddings for different values of k.

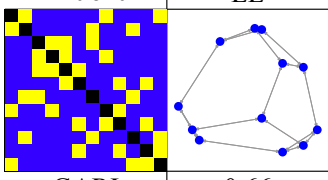
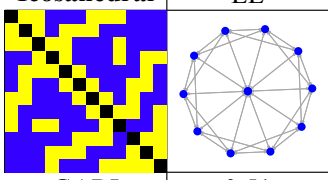
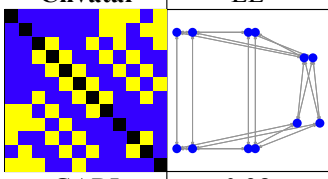
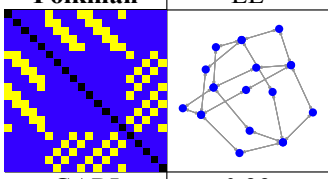
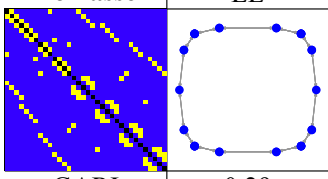
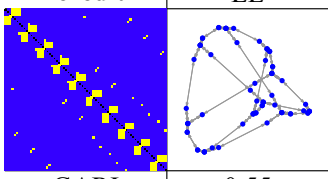
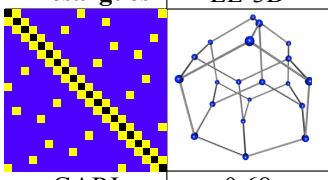
Frucht	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.66	0.73	0.77	0.73	0.69	0.84	**1.00**
$\text{Err}_{\text{local}}$	0.47	0.36	0.50	2.79	0.50	0.62	**0.00**
Icosahedral	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.51	0.57	0.57	**0.69**	0.57	0.57	0.63
$\text{Err}_{\text{local}}$	1.54	1.44	1.42	3.38	1.40	1.82	**1.32**
Chvatal	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.38	0.12	0.12	**0.48**	0.28	0.41	0.35
$\text{Err}_{\text{local}}$	2.25	2.69	2.65	3.30	2.45	2.42	**2.03**
Folkman	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.29	0.37	0.37	0.43	0.37	**0.52**	0.37
$\text{Err}_{\text{local}}$	6.27	6.04	5.96	11.9	6.18	7.30	**5.34**
Thomassen	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.29	0.42	0.38	0.30	0.30	**0.68**	0.53
$\text{Err}_{\text{local}}$	6.18	4.52	4.32	30.2	5.79	7.80	**3.04**
Meredith	LE	KK	FR	GNMDS	SPE	*t*-SNE	**LOE**
							
GARI	0.55	0.45	0.45	**0.59**	0.40	0.48	0.58
$\text{Err}_{\text{local}}$	29.0	21.7	21.5	24.2	27.8	154.6	**12.6**
Desargues	LE-3D	KK-3D	FR-3D	GNMDS-3D	SPE-3D	*t*-SNE-3D	**LOE-3D**
							
GARI	0.68	0.84	0.88	0.68	0.80	0.76	**1.00**
$\text{Err}_{\text{local}}$	4.14	4.26	4.21	9.50	4.30	5.17	**0.00**

Figure 3: Adjacency matrices (0 : blue and 1 : yellow) and embeddings of 7 methods for some classical graphs.

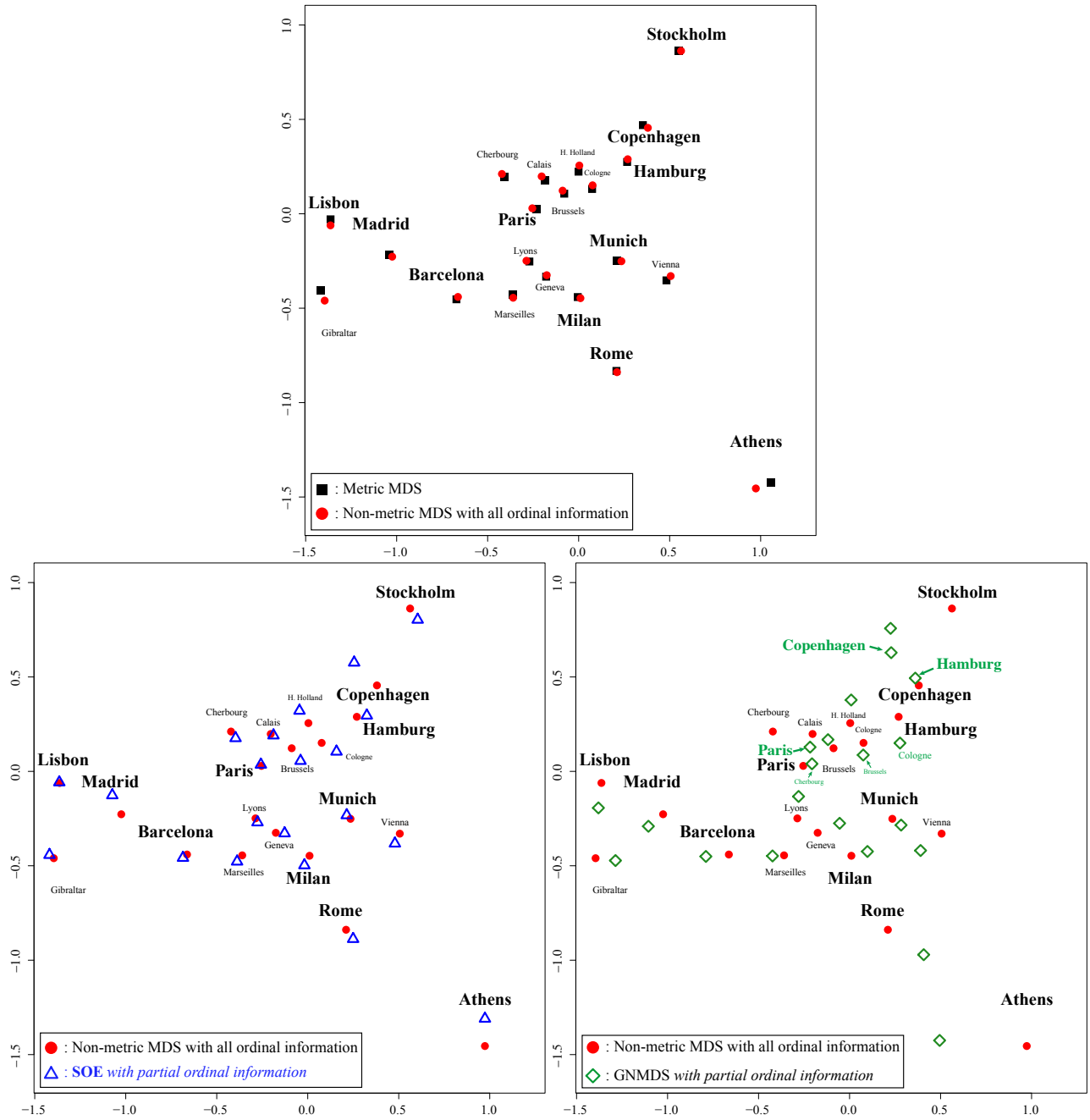


Figure 4: Two-dimensional embeddings of 21 cities in Europe (Upper figure: metric MDS with the original distance matrix and non-metric MDS with all ordinal comparisons, Lower left figure: non-metric MDS with all ordinal comparisons and SOE with 1000 ordinal comparisons, Lower right figure: non-metric MDS with all ordinal comparisons and GNMDS with 1000 ordinal comparisons).

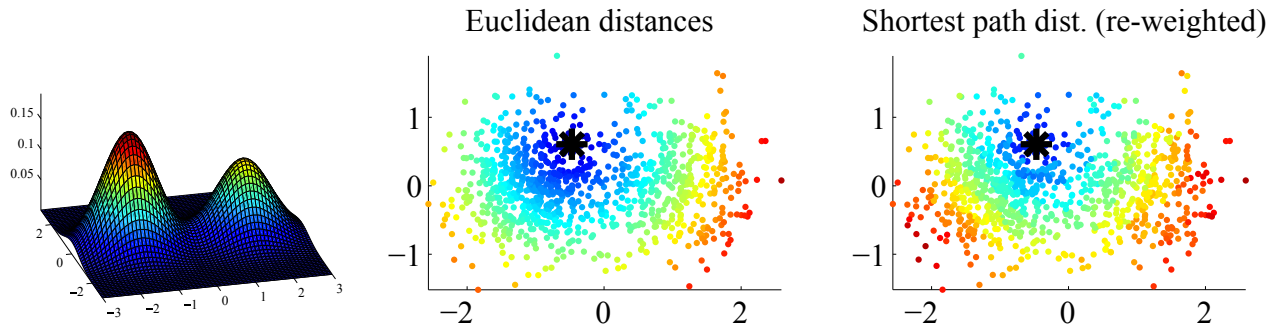


Figure 5: Left side: Density of the mixture of Gaussians used in the simulation. Middle and right: we fix one data point X, marked by a black star. The color of the remaining points encodes the distance of the corresponding point to X. We can see that the true Euclidean distance (middle plot) is very well approximated by the shortest path distance in the re-weighted graph (right plot).

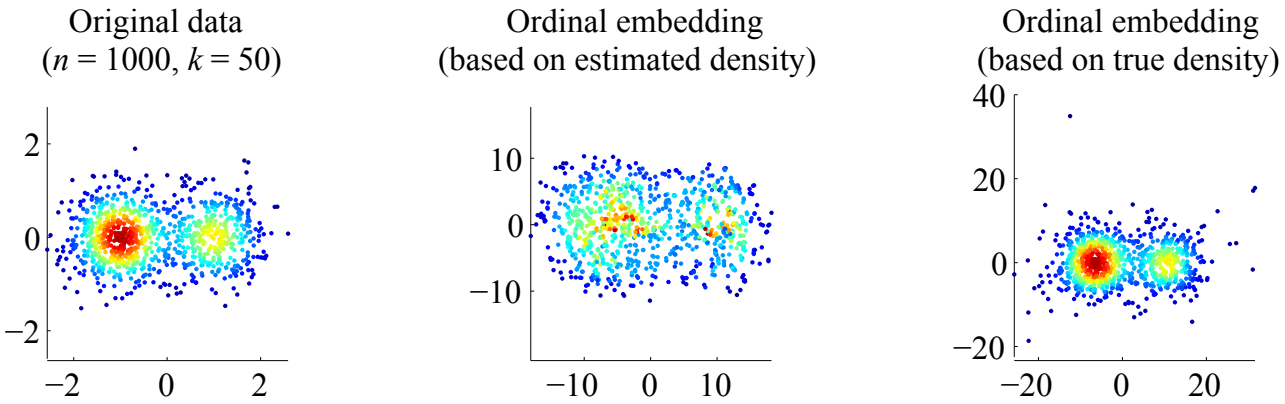


Figure 6: Left side: original sample from the mixture of Gaussians. The color encodes the true density values. Middle: reconstructed point set. We first estimated the density based on the unweighted kNN graph, then re-weighted the graph edges as in the Proof of Proposition 4, computed the shortest path distances in this reweighted graph, and then applied classic (metric) multidimensional scaling. The colors show the values of the estimated density. Right: same procedure as in the middle, but the graph was reweighted based on the true distance values. We can see that this results in a nearly perfect embedding.