Neurocomputing 141 (2014) 97–109

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Correlation-based embedding of pairwise score data

Marc Strickert [a,*], Kerstin Bunte [b], Frank-Michael Schleif [c,d], Eyke Hüllermeier [a]

[a] Department of Mathematics and Computer Science, SYNMIKRO, Philipps Universität Marburg, Germany
[b] Department of Information and Computer Science, Aalto University School of Science and Technology, Finland
[c] CITEC Center of Excellence, Bielefeld University, Germany
[d] School of Computer Science, University of Birmingham, United Kingdom

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form
24 January 2014
Accepted 27 January 2014
Available online 5 April 2014

Keywords:
Multidimensional scaling
Neighbor embedding
Score data
Visualization

ABSTRACT

Neighbor-preserving embedding of relational data in low-dimensional Euclidean spaces is studied. Contrary to variants of stochastic neighbor embedding that minimize divergence measures between estimated neighborhood probability distributions, the proposed approach fits configurations in the output space by maximizing correlation with potentially asymmetric or missing relationships in the input space. In addition to the linear Pearson correlation measure, the use of soft formulations of Spearman and Kendall rank correlation is investigated for optimizing embeddings like 2D point cloud configurations. We illustrate how this scale-invariant correlation-based framework of multidimensional scaling (cbMDS) helps going beyond distance-preserving scaling approaches and how the embedding results are characteristically different from recent neighborhood embedding techniques.

1. Introduction

Data visualization is an important tool during stages of initial data screening, validation, and analysis. Scatter plot displays are frequently used to study the context of data items and their relationship with other data being represented as point clouds [1]. Traditionally, dimension reduction is being sought for displaying vectorial data items in a scatter plot. Beyond this, complex relational data sets containing pairwise scores, affinities, or preferences are available for being visually inspected nowadays. While dimension reduction problems gave rise to a plethora of methods, the treatment of partially missing relational data beyond mere pairwise distances is a rather young topic [2] being mainly addressed in the current work.

Principal component projection refers to a widely used technique for mapping data tables in a linear fashion to points along axes of maximum attribute variance. The key concept is an eigen-decomposition of the covariance matrix or, alternatively, a singular value decomposition of the data matrix [3]. For the reconstruction of relational pairwise distances in a lower-dimensional Euclidean space, classical multidimensional scaling (cMDS) can be applied to such square input matrices [4]. Kernel principal component analysis (kPCA) can be seen as a generalization of cMDS in the sense that kernel functions are used to construct Gram matrices, i.e. pairwise data similarities, to be visualized [5]; thereby, standard PCA point configurations are obtained by assigning negative squared Euclidean distances as elements of the Gram matrix. The potential computational burden of kPCA can be reduced by low-rank approximations of the kernel matrix [6].

Kernel PCA is also used for local and specialized manifold learning. Locally linear embedding (LLE) models local data neighborhoods as a connection weight matrix, including zeros for data outside the local neighborhood range [7]; ISOMAP compares data points by their geodesic distances along the locally constructed neighborhood graph [8]. Both methods give rise to connectivity matrices specifically derived from the data that can be used as Gram matrices for reconstruction by kPCA. Generally, flexible functional mappings of new data vectors are desirable, but they are not available for pairwise relational representations for which the inclusion of new data points changes the structure of corresponding data similarity matrices.

Kernel PCA offers some flexibility for processing different types of input relationships, however, the above-mentioned methods are often found in the domain of dimension reduction of Euclidean input data, such as the synthetic three-dimensional swiss-roll data set [7] or high-dimensional images like hand gestures or handwritten digits [8]. Some applications of these methods to non-vectorial data exist [9], but a principal restriction remains: the inevitable symmetry of the underlying kernel matrix derived from data comparisons. Such symmetry is not always natural. For example, if – in a set of points – A is nearest neighbor of point B, this does often not hold true the other way round in case of skewed densities.

* Corresponding author.
E-mail address: marc.strickert@uni-marburg.de (M. Strickert).

http://dx.doi.org/10.1016/j.neucom.2014.01.049
0925-2312/

In contrast to algebraic solutions of the point embedding problem by eigen-decomposition of the kernel matrix, non-linear iterative solutions are constructed by non-classical multidimensional scaling (MDS) and neighbor embedding approaches. Iterative scaling methods arrange a set of low-dimensional points of which the pairwise distances approximate given input dissimilarities. These dissimilarities are not necessarily Euclidean distances, but often they use Euclidean distances for further transformation by user-defined stress functions, e.g. for emphasizing on certain scales in the data [10]. Again, symmetric relationships are assumed for the input data, because MDS stress functions match Euclidean distances in the embedding space.

Stochastic neighbor embedding (SNE) is a more flexible concept for approximately reconstructing pairwise data neighborhood relationships in the Euclidean space [11,12]. Adjacent data items are identified as neighborhood probabilities. Based on this concept metric relationships, dissimilarities and pairwise scoring data could be treated in the same manner. The minimization of Kullback–Leibler divergence between probability distributions of neighborhoods in the input and output spaces is the optimization goal. Recently, a generalization of Jensen–Shannon divergences has been proposed to better resolve undesired density concentrations in the input distributions, helping to better control the quality of neighborhood reconstruction [13,14]. While the original articles limit their embedding schemes to dimension reduction of Euclidean distances, more general pairwise relationships can be considered, such as Minkowski distances or string compression distance. Even more generic scoring data can be addressed like greedy string tiling [15] or kernel function values [16]. As opposed to dissimilarity relations, higher score values indicate higher similarity. With suitable embedding algorithms such pairwise relationships can be used as input for creating intuitive visual displays of document topics and gene expression data [17], for example.

In this work we present a non-parametric embedding technique that is conceptually located between iterative MDS and stochastic neighbor embedding approaches. While MDS uses some stress criterion for reconstructing distances identical to the transformed input measure, SNE operates on neighborhood probability distributions. Correlation measures offer a compromise between strict distance matching stress, aiming at vectors to become equal, and divergence measures that operate on vectors of neighborhood density estimates and which are invariant under scalar translation and scaling transformations. Linear Pearson correlation was earlier utilized in the MDS context as a fast and scale-independent way to globally compare distances in input and output spaces [18]. The application of weighted rank correlation measures between object-specific relationships in both spaces was recently proposed as alternative to isotonic regression [19,20] for the reconstruction of local neighborhood orders [21,22]. Other authors have addressed the problem of neighborhood reconstruction earlier. Venna and Kaski propose local MDS to optimize the embedding for matching neighbors in the input space also in the output space and vice versa [23]. While their approach still relies on smooth transformations of distances within pre-defined radii, the RankVisu method compares between ranks of input and output distance matrices and it makes use of force-directed placement of points in the embedding space to minimize the discrepancy [24]. By its design it favors preservation of local neighborhoods rather than global. Its great flexibility comes at the cost of large computational efforts for dealing with discrete ranks. Onclinx et al. proposed a rank-based optimization scheme where rank-induced discrete plateaus in the embedding stress function are avoided by distance-based interpolation between ranks [25].

In the following, a framework for correlation-based multidimensional scaling will be described in detail. Particularly, soft-rank optimization approaches are being discussed. Illustrative examples are provided along with relevant applications to protein data.

2. Embedding framework

Data embedding methods follow a general principle [26] which can be summarized as follows. For a given finite set of n data items some characteristics $\mathrm{char}_{\mathbf{X}}$ are derived and the aim is to match them as well as possible with corresponding characteristics $\mathrm{char}_{\mathbf{Y}}$ in the low-dimensional space:

$$\mathrm{tension}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} \mathrm{m}(\mathrm{char}_{\mathbf{X}}(\mathbf{X},\mathbf{x}_i), \mathrm{char}_{\mathbf{Y}}(\mathbf{Y},\mathbf{y}_i)). \tag{1}$$

Here $\mathrm{m}(\cdot)$ denotes a measure of mismatch between the characteristics, and the index i refers to the ith data object $\mathbf{x}_i$ and its low-dimensional counterpart $\mathbf{y}_i$. The source matrix contains pairwise similarity information about the data items. Optimization of usually low-dimensional point coordinates $\{\mathbf{y}_i\}_{i=1}^{n}$ or of parameters θ of a functional point placement model $\mathbf{Y} = F_\theta(\mathbf{X})$ allows for minimization of the overall tension. Preferably, continuous tension functions are employed for gradient-based coordinate optimization.

In dimension reduction scenarios, pairwise input information refers to distances of high-dimensional input vectors. For visualization, one-, two-, or three-dimensional Euclidean spaces are common embedding targets. Using the above formalism with $\mathrm{m} = \mathrm{m}_{\mathrm{MDS}}$ being the sum of squares and $\mathrm{char}(\cdot,\cdot)$ picking pairwise distances $\mathbf{D}_{ij}$, classical MDS can be expressed as

$$\mathrm{tension}_{\mathrm{MDS}}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} (\mathbf{D}_{ij}^{\mathbf{X}} - \mathbf{D}_{ij}^{\mathbf{Y}})^2. \tag{2}$$

In practice, algebraic eigen-decomposition is used for solving this classical scaling problem efficiently [4]. However, a large variety of modifications exists for modeling embedding stress in customized, e.g. scale sensitive, ways by iterative optimization of suitably designed tension functions m [10].

In a comparison of distance distributions of high-dimensional Euclidean data points and low-dimensional points it turns out that the former one is shifted to higher average distances with relatively low standard deviation. This phenomenon is referred to as concentration of the norm [27]. In order to embed such distances with their specific properties properly in a low-dimensional space, versions of SNE [12] and the neighbor retrieval visualizer NeRV [28] apply different input and output distributions. Gaussian distributions $\mathbf{P}(\mathbf{X})$ are used in the high-dimensional input space and Student t-distributions $\mathbf{Q}(\mathbf{Y})$ in the low-dimensional output space aiming at minimizing the Kullback–Leibler divergence (KL) between them by adapting low-dimensional points $\mathbf{Y}$. Mismatch between per-object neighborhood probabilities is thus modeled by $\mathrm{m}_{\mathrm{t-SNE}} = \mathrm{KL}(\mathbf{P} \,\|\, \mathbf{Q}(\mathbf{Y}))$:

$$\mathrm{tension}_{\mathrm{t-SNE}}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} \mathrm{KL}(\mathbf{P}_i(\mathbf{X}) \,\|\, \mathbf{Q}_i(\mathbf{Y})). \tag{3}$$

Neighborhoods are expressed in terms of σ_i-localized Gaussian transformations of squared Euclidean distances:

$$\mathbf{P}_{ij} = \frac{\exp(-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma_i)}{\sum_{k\neq i}\exp(-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2/\sigma_i)}. \tag{4}$$

The neighborhood probability is modeled indirectly by setting the bell shape width σ_i for each point to capture to which degree nearby points are considered as neighbors for a fixed radius of 'effective' neighbors. This number is referred to as perplexity parameter and is usually set to $5 \le p \le 50$. Naturally, variations in data densities lead to different σ_i and, consequently, to asymmetric matrices $\mathbf{P}$. Gaussian distributions could be used in the

embedding space too, but in order to embed large input distances with relatively low variability in a low-dimensional space, the heavy-tailed Student t-distribution

$$\mathbf{Q}_{ij}(\mathbf{Y}) = \frac{(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}^{n}(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2)^{-1}} \tag{5}$$

turned out to have more suitable characteristics [12]. The use of distributions has an often desirable smoothing effect on the degree to which points are considered neighbors. From a practical point of view, distances get stretched compared to the Gaussian and can be embedded with less clutter in low-dimensional Euclidean spaces. From a theoretical perspective it is not obvious, though, which underlying process justifies a change from one family of distributions to another for mediating between, say, 30-dimensional and three-dimensional spaces. Concentration of the norm was originally stated for Euclidean spaces to argue for different distribution models in the input and output spaces. For general input data relationships, like pairwise scorings, other demands regarding the distributions and neighborhood probability estimation models for $\mathbf{P}$ can be expected.

Kullback–Leibler divergence is a scale-invariant measure between distributions $\mathbf{P}$ and $\mathbf{Q}(\mathbf{Y})$ [13]. In contrast to potential misconception, this does not imply scale-invariance of embedding points $\mathbf{Y}$, because rescaling of $\mathbf{Y}$ leads to non-proportional changes of $\mathbf{Q}(\mathbf{Y})$, i.e. $\mathbf{Q}(\gamma\cdot\mathbf{Y}) = g(\gamma)\cdot\mathbf{Q}(\mathbf{Y}) \Rightarrow \gamma = 1 \wedge g(1) = 1$, irrespective of $\mathbf{Q}(\mathbf{Y})$ being a Gaussian or Student-t distribution. While an optimum scale might be desirable for restricting the solution space, degenerate solutions may exist, though prevented by numerical limitations in practice. We will seek at scale-invariant embeddings by design.

2.1. Symmetry considerations

While the original SNE formulation takes into consideration asymmetric local distance density estimates $\mathbf{P}$, symmetry $\mathbf{P}_{\text{sym}} \propto \mathbf{P}+\mathbf{P}^\top$ is forced in t-SNE to improve speed and visualization properties, as stated by the authors [12]. Such improvements sacrifice parts of the original neighborhood topology though, but interestingly forcing symmetry seems to better separate between clusters of labeled data clouds in the t-SNE experiments.

For some pairwise scoring data asymmetric information should be maintained as much as possible. For example, in social networks it might be good to know if Jim loves Mary but not vice versa. We follow the convention to consider high scores as high degree of similarity. Asymmetric score data also naturally contributes to bioinformatics tasks like protein sequence alignments. For example, homology in transmembrane proteins was found to be faithfully modeled by the SLIM family of asymmetric amino acid block substitution matrices [29]. Table 1 contains scores for a subset of five amino acids from the original 20×20 SLIM 161 substitution matrix. The number subscripts are ranks, smaller means higher affinity. These ranks point out an asymmetric per-object neighborhood structure that should not be forced into a symmetric rank matrix. This asymmetric structure can be perfectly reproduced in two dimensions as shown by some point configuration in Fig. 1 with corresponding distances and neighbor ranks reported in Table 2. Thus, to a certain degree, asymmetric relationships can be expressed as neighbor ranks in the Euclidean space despite its symmetric distance structure.

Table 1
SLIM 161 subset score matrix. Subscripts indicate row-wise neighbor rank.

	R	C	E	P	S
R	10_1	-2_2	-11_5	-7_4	-4_3
C	-8_3	11_1	-12_5	-11_4	2_2
E	-7_5	-2_3	7_1	-4_4	-1_2
P	-9_4	-7_3	-10_5	11_1	-3_2
S	-8_4	4_2	-9_5	-5_3	6_1

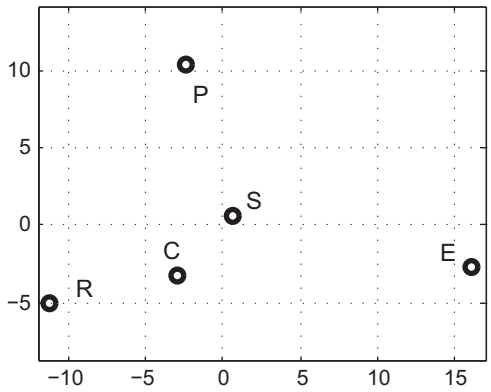

Fig. 1. Neighbor-preserving scatter plot representation of a SLIM 161 subset.

Table 2
SLIM 161 subset distance matrix from points in Fig. 1 rounded in a rank-preserving way to integer values. Ranks are given as subscripts and highlight the asymmetry of neighbor relationships for the five amino acids. These ranks are in perfect correspondence to the original scores from Table 1.

	R	C	E	P	S
R	0_1	8_2	27_5	18_4	13_3
C	8_3	0_1	19_5	14_4	5_2
E	27_5	19_3	0_1	23_4	16_2
P	18_4	14_3	23_5	0_1	10_2
S	13_4	5_2	16_5	10_3	0_1

3. Correlation-based embedding framework

Let $\mathbf{S}$ be an $n\times n$ matrix of pairwise similarity scores denoting the characteristics of the original data. $\mathbf{D}^{\mathbf{Y}}$ is the matrix containing the Euclidean distances of the adjustable m-dimensional objects $\mathbf{Y}$ implying the characteristics of the ith object in the embedding space:

$$\mathbf{D}^{\mathbf{Y}}_{ij} = \left(\sum_{k=1}^{m}(\mathbf{y}^i_k-\mathbf{y}^j_k)^2\right)^{1/2}. \tag{6}$$

Correlation is a common measure of correspondence that can be applied to the vectors of per-object scores $\mathbf{S}_i$ defined as ith row of matrix $\mathbf{S}$ and reconstructed distances $\mathbf{D}^{\mathbf{Y}}_i$. Such measure is a value in the range between -1 (negative correlation, inverse similarity) and $+1$ (positive correlation); values around zero indicate uncorrelatedness. Since high correlation values denote common patterns of similarity, the correlation between *negative* scores and distances is to be maximized. The three most commonly used correlation measures are linear Pearson correlation, Spearman rank correlation, and Kendall rank correlation for quantifying the degree of similarity between two vectors of identical dimension [30].

Pearson correlation is a measure of linear association between real-value vectors; Spearman rank correlation is used for comparing integer-valued rankings, i.e. simultaneously monotonic relationships, of the entries of two vectors; and Kendall correlation is used for quantifying co-occurrences of positive and negative signs in pairwise comparisons of potentially ordinal components in two data vectors. The maximization of Pearson correlation can be seen

as scale-free alternative to least-squares optimization in typical MDS tasks [18]. Spearman rank correlation constitutes a natural measure for neighbor embedding tasks, that is, for reconstructing ordered neighborhoods based on negative input scores and output distances. Kendall correlation is best for use with ordinal scoring information that are even more general than being described by a given score matrix $\mathbf{S}$, like object B being qualitatively more similar to A than C is to A, without knowing quantitative information such as exact rank differences. All three correlation measures are described in separate sections below.

Each object i is described by its relationship to all other $(n-1)$ objects, that is, each row of the Euclidean distance matrix $\mathbf{D}_i^{\mathbf{Y}}$ in the low-dimensional embedding space should match the corresponding row of input scores $\mathbf{S}_i$. Thus, neighborhood embedding might be performed by maximizing the averaged correlations r along all rows of both matrices:

$$\arg\max_{\mathbf{Y}\in\mathbb{R}^m} \bar{r} := \frac{1}{n}\sum_{i=1}^{n} r(-\mathbf{S}_i, \mathbf{D}_i^{\mathbf{Y}}). \tag{7}$$

The signs of the original scores contained in $\mathbf{S}$ are flipped, such that smaller values appear analogous to smaller distances as higher degree of similarity. Integrating this way over per-object correspondences is an object-conditional reconstruction problem, which is similarly modeled in variants of SNE.

Object-conditional reconstruction is in contrast to unconditional (also called matrix-conditional or global) reconstruction where object placement depends on the whole matrix with contributions also across rows. Classical MDS with squared distance stress criterion belongs to both types, because the average sum of row-wise or column-wise squares is proportional to the average of all squares. Nonmetric versions of MDS on the other hand are traditionally defined for vectorized all-pair distances of the whole input and output matrices and not separately for rows or columns [31]. For example, individual row maxima that might be used for object-specific rescaling purposes differ from the global maximum distance. The gradient of the cost function (7) for the proposed row-wise correlation-based multidimensional scaling (cbMDS) approach can be found in Appendix A. In the following sections we will introduce and discuss different correlation measures r for the cost function (7) that share the property of being invariant under scaling and translation of the arguments, i.e. $r(-\mathbf{S}_i, \mathbf{D}_i^{\mathbf{Y}}) = r(-\gamma_1\cdot\mathbf{S}_i+\delta_1, \gamma_2\cdot\mathbf{D}_i^{\mathbf{Y}}+\delta_2)$ for $\gamma_1,\gamma_2\in\mathbb{R}^+, \delta_1,\delta_2\in\mathbb{R}$.

3.1. Pearson correlation

In multidimensional scaling, scatter plots of reconstructed distances against original distances, so-called Shepard diagrams, are commonly used to visually assess the quality of fit [32]. Perfect reconstruction is obtained if this scatter of reconstructed and original distances coincides with the diagonal line. Yet, any straight line with positive slope and arbitrary intercept represents a good reconstruction target, because the relationships between distances are still maintained in a scale-invariant manner. Formally, this goal can be expressed by maximizing the linear Pearson correlation between reconstructed and original distances.

The Pearson correlation coefficient $r_P(\boldsymbol{w},\boldsymbol{u})\in[-1,1]$ between two vectors $\boldsymbol{w}$ and $\boldsymbol{u}$ is given by

$$r_P(\boldsymbol{w},\boldsymbol{u}) = \frac{\sum_{i=1}^{n}(\boldsymbol{w}_i-\mu_{\boldsymbol{w}})\cdot(\boldsymbol{u}_i-\mu_{\boldsymbol{u}})}{\sqrt{(\sum_{i=1}^{n}(\boldsymbol{w}_i-\mu_{\boldsymbol{w}})^2)\cdot(\sum_{i=1}^{n}(\boldsymbol{u}_i-\mu_{\boldsymbol{u}})^2)}} \tag{8}$$

Setting $\boldsymbol{w}=-\mathbf{S}_i$ and $\boldsymbol{u}=\mathbf{D}_i^{\mathbf{Y}}$ and replacing r by r_P in (7) allows to maximize the Pearson correlation between input space similarity scores and Euclidean distances in the low-dimensional space. The gradient of r_P can be found in Appendix B.

We point out that unweighted Pearson correlation is structurally related to the Cauchy–Schwarz divergence via logarithmic transformation:

$$\mathrm{d}_{\mathrm{CS}}(\boldsymbol{w},\boldsymbol{u}) = \tfrac{1}{2}\cdot\log(\langle\boldsymbol{w},\boldsymbol{w}\rangle\cdot\langle\boldsymbol{u},\boldsymbol{u}\rangle) - \log(\langle\boldsymbol{u},\boldsymbol{w}\rangle). \tag{9}$$

In contrast to dissimilarity vectors in the Pearson correlation, here $\boldsymbol{w}$ and $\boldsymbol{u}$ represent distributions, that is, they remain uncentered items. Some benefits of Cauchy–Schwarz divergence over Kullback–Leibler divergence have been discussed for pattern recognition scenarios recently [33,13].

Beyond the discussed least-squares modeling of reconstructed distances against true distances or negative scores, real neighbor embedding requires more advanced models: recovery of local distributions is one option [34], ranking is another one [35]. Since ranking aims, in a non-parametric way, at putting objects into a given order, this approach is ideal for proper neighborhood reconstruction if per-object orderings are being optimized. Thus, for directly modeling neighborhood, order-based rankings instead of linear correlation quantities should be considered for bridging from the reconstruction of distance to neighbor relationships, as addressed in the following.

3.2. Soft Spearman rank correlation

The Spearman rank correlation coefficient r_ρ is easily obtained by first converting data vectors into the order ranks of their elements. These rank vectors are used as arguments of Pearson correlation in (8)

$$r_\rho(\boldsymbol{w},\boldsymbol{u}) = r_P(\mathrm{rnk}(\boldsymbol{w}), \mathrm{rnk}(\boldsymbol{u})) \tag{10}$$

For deriving a continuous ranking approach which is not based on discrete sorting operations, the ranking $\mathrm{rnk}(\cdot)$ of vector elements is alternatively achieved by summing up rows of the indicator matrix $\boldsymbol{Z}$:

$$\mathrm{rnk}(\boldsymbol{u}) = \begin{pmatrix} \sum_{i=1}^{n}\mathsf{Z}(\boldsymbol{u}_1,\boldsymbol{u}_i) \\ \cdots \\ \sum_{i=1}^{n}\mathsf{Z}(\boldsymbol{u}_n,\boldsymbol{u}_i) \end{pmatrix} \quad \text{for}$$

$$\boldsymbol{Z}(\boldsymbol{u}) = \begin{pmatrix} \mathsf{Z}(\boldsymbol{u}_1,\boldsymbol{u}_1) & \cdots & \mathsf{Z}(\boldsymbol{u}_1,\boldsymbol{u}_n) \\ & \cdots & \\ \mathsf{Z}(\boldsymbol{u}_n,\boldsymbol{u}_1) & \cdots & \mathsf{Z}(\boldsymbol{u}_n,\boldsymbol{u}_n) \end{pmatrix}. \tag{11}$$

For the Heaviside step function $\mathsf{Z}(\boldsymbol{u}_k,\boldsymbol{u}_l) = H(\boldsymbol{u}_k-\boldsymbol{u}_l)$, providing zero for negative arguments and else one, correct ranks are obtained for vector elements $\boldsymbol{u}_k$ in the absence of ties. We focus on $\boldsymbol{u}$ for ranking points distances in the iteratively adapted embedding, while $\boldsymbol{w}$ in 10 can be assumed as precomputed ranks of fixed input object scores. As a continuous approximation of distance ranking in the embedding, the step function $H(\boldsymbol{u}_k-\boldsymbol{u}_l)$ can be replaced by a differentiable sigmoid

$$\begin{aligned}\mathsf{Z}(\boldsymbol{u}_k,\boldsymbol{u}_l) &= \mathrm{sgd}_\kappa^{kl} + \frac{1}{2n} = \mathrm{sgd}_\kappa\left(\frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_u}\right) + \frac{1}{2n} \\ &= \frac{1}{1+e^{-\kappa\cdot(\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}}} + \frac{1}{2n}\end{aligned} \tag{12}$$

with mid-tied ranks of vector elements $\boldsymbol{u}_i$ being approximated for $\kappa\to\infty$. Self-comparisons $(\boldsymbol{u}_i-\boldsymbol{u}_i)$ along the diagonal lead to sigmoid values of 1/2, thus, n times the offset of $1/(2n)$ lets the approximated ranks start at 1 without affecting derivatives. Generally, large κ are preferred for strict rank approximations, but the induced flat tails of the sigmoid derivative complicate the optimization. In practice $\kappa<100$ is considered as numerically feasible in most cases. Adaptation signals from derivatives of the sigmoid may else vanish numerically to zero whenever $\exp(-746)$ or smaller, that is, double precision underflow, is encountered.

To cope with potentially large variations in differences $\boldsymbol{u}_k - \boldsymbol{u}_l$, i.e. for a scale-invariant setting of κ, a division by the standard deviation $\sigma_{\boldsymbol{u}}$ is carried out. The usefulness of this rescaling is shown by expressing standard deviation $\sigma_{\boldsymbol{u}}$ as the square root of variance, naturally written as overall average of squared paired differences:

$$\sigma_{\boldsymbol{u}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \mu_{\boldsymbol{u}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{u}_i - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{u}_j\right)^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2. \tag{13}$$

For gradient-based optimization the gradient vector of the soft Spearman rank correlation is provided:

$$\frac{\partial r_\rho(\boldsymbol{w},\boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{J}_{r_P|\mathrm{rnk}(\boldsymbol{w})}(\mathrm{rnk}(\boldsymbol{u}))\boldsymbol{J}_{\mathrm{rnk}(\boldsymbol{u})}(\boldsymbol{u}). \tag{14}$$

Therein, $\boldsymbol{J}_{r_P}$ denotes the gradient of the Pearson correlation r_P, and $\boldsymbol{J}_{\mathrm{rnk}(\boldsymbol{u})}$ is the Jacobian matrix of the approximated ranks. The necessary derivatives are given in Appendix C. Substituting $\boldsymbol{w} = -\mathbf{S}_i$ and $\boldsymbol{u} = \mathbf{D}_i^{\mathbf{Y}}$ this gradient is plugged into (7) for optimizing a point set $\mathbf{Y}$ with ranks of Euclidean relationships best matching the ranks of original data relationships. This way neighborhood ranks can be reconstructed by continuous optimization.

3.3. Soft Kendall correlation

The Spearman correlation is a basic measure for the comparison of two given rank vectors and thus a good candidate for the reconstruction of neighborhoods. In this section we consider an alternative measure that is not based on whole rankings but on pairwise comparisons.

For optimization let one score be less than another score for a fixed object in the input space; then, the corresponding Euclidean distance relationship in the embedding space should be reversed. More generally, for every three objects in the input space the inverse similarity relationships should be reestablished by point distances in the Euclidean embedding space. For reaching this goal we make use of the Kendall correlation coefficient r_τ for comparing vectors $\boldsymbol{w}$ of negative scores from the input space and $\boldsymbol{u}$ of corresponding Euclidean distances from the embedding space by assessing the local ordering of all their elements and counting the number of concordant (C_{ij}) and discordant (D_{ij}) pairs:

$$C_{ij} = (\boldsymbol{w}_i > \boldsymbol{w}_j \wedge \boldsymbol{u}_i > \boldsymbol{u}_j) \vee (\boldsymbol{w}_i < \boldsymbol{w}_j \wedge \boldsymbol{u}_i < \boldsymbol{u}_j), \tag{15}$$

$$D_{ij} = (\boldsymbol{w}_i > \boldsymbol{w}_j \wedge \boldsymbol{u}_i < \boldsymbol{u}_j) \vee (\boldsymbol{w}_i < \boldsymbol{w}_j \wedge \boldsymbol{u}_i > \boldsymbol{u}_j). \tag{16}$$

For all $i,j \in \{1,\ldots,n\}$ let $\#C$ be the total number of occasions whenever C_{ij} evaluates to TRUE, and likewise $\#D$ summarizes D_{ij}. Ignoring the pairs $i=j$, the maximum number of mutually exclusive concordant and discordant pairs is $\#C+D = (n\cdot n-n)/2$. Thus, the normalized difference of the pair counts is used to quantify trends of positive or negative correlation:

$$r_\tau(\boldsymbol{w},\boldsymbol{u}) = 2\cdot\frac{\#C-\#D}{n(n-1)} \in [-1,1]. \tag{17}$$

This common definition does not consider ties, and we also assume their absence in the following.

In order to get a differentiable measure we use a soft version of Kendall correlation:

$$\widehat{r}_{\tau,\kappa}(\boldsymbol{w},\boldsymbol{u}) = 1 - 2\cdot\frac{(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Z}(\tilde{p}_{ij})) - q}{n(n-1)}, \tag{18}$$

which is based on the products of differences:

$$\tilde{p}_{ij} = (\boldsymbol{w}_j - \boldsymbol{w}_i)\cdot(\boldsymbol{u}_i - \boldsymbol{u}_j) \quad \begin{cases} <0 & \text{for concordant pairs} \\ >0 & \text{for discordant pairs} \end{cases} \tag{19}$$

and an indicator matrix $\boldsymbol{Z}$ constructed from all pairs:

$$\boldsymbol{Z}(\tilde{p}) = \begin{pmatrix} \mathrm{Z}(\tilde{p}_{11}) & \ldots & \mathrm{Z}(\tilde{p}_{1n}) \\ & \ldots & \\ \mathrm{Z}(\tilde{p}_{n1}) & \ldots & \mathrm{Z}(\tilde{p}_{nn}) \end{pmatrix}. \tag{20}$$

For the step function $\mathrm{Z}(z) = \frac{1}{2}(\mathrm{sign}(z)+1)$, providing zero for negative arguments, 0.5 for zero and 1 otherwise, the Kendall correlation $r_\tau = \widehat{r}_{\tau,\kappa}$ is recovered by (18) if $q = n/2$. This value of q compensates for the fact that zero differences occurring for pairs $i=j$ get counted as partial discordances $\mathrm{Z}(0) = 0.5$. In (18) twice the relative amount of discordant pairs is subtracted from the highest possible correlation value. This is valid, because untied data induces a complementary amount of concordant pairs.

In order to get a differentiable measure we use again the sigmoidal approximation of the step function as soft indicator function:

$$\mathrm{Z}(\tilde{p}_{ij}) = \mathrm{sgd}_\kappa\left(\frac{\tilde{p}_{ij}}{\sigma_{\boldsymbol{w}}\sigma_{\boldsymbol{u}}}\right). \tag{21}$$

Again, the larger the value of κ the steeper the sigmoidal transition from zero to one, such that $\lim_{\kappa\to\infty}\widehat{r}_{\tau,\kappa}(\boldsymbol{w},\boldsymbol{u}) = r_\tau(\boldsymbol{w},\boldsymbol{u})$. Rescaling of $\tilde{p}_{ij}$ by the standard deviations of $\boldsymbol{w}$ and $\boldsymbol{u}$ follows the same argument as outlined in (13) for Spearman correlation. Again, the variables $\boldsymbol{w}$ and $\boldsymbol{v}$ are substituted back to scores and distances $\boldsymbol{w} = -\mathbf{S}_i$ and $\boldsymbol{u} = \mathbf{D}_i^{\mathbf{Y}}$.

All building blocks of soft Kendall correlation are differentiable, and the overall gradient of $\widehat{r}_{\tau,\kappa}$ is derived in Appendix D. Generally, per-object gradients – determining effective directions of point movements – can be used to assess the goodness of the embedding for individual objects. Points fall to stationary positions after convergence, but there are net forces of $n-1$ relations acting on any point. These forces are described by the sum of absolute values of the components of the per-object input–output correlation gradients (B.1), (C.1) and (D.1). Large values, for example displayed by larger glyphs, highlight points under strong tension that either cannot be embedded easily or which serve as important hubs of the connectivity structure.

4. Practical issues

In this section some statements are made regarding optimization, runtime, and output standardization.

4.1. Optimization

Gradients of Pearson, soft Spearman and soft Kendall correlation allow for utilization of unconstrained gradient-based optimization methods. The maximization of correlation by moving embedding points is a non-convex problem. Similar to other neighbor embedding methods, partially optimal configurations may fail to converge optimally on the large scale. Some approaches try to circumvent local optima by stochastic gradient descent [12]. For our problem, we found that the memory-limited quasi-Newton l-BFGS gradient batch optimization scheme with built-in default parameters provides good convergence, with termination triggered at changes of objective function values below 10^{-7}.

For exploiting some structure of the data, initial embedding coordinates can be obtained by treating rank vectors of per-object input scores as features to be mapped by random linear projection into the embedding space. According to the Johnson–Lindenstrauss lemma this helps to map nearby matrix rows into similar regions of the low-dimensional embedding space [36].

4.2. Runtime

The runtime of embeddings based on Pearson correlation is $\mathcal{O}(n^2)$, that is $\mathcal{O}(n)$ for the correlation over n objects, with a factor depending on the number of iterations to convergence. This is a common complexity when pairwise scoring data are operated on without advanced methods like Nystrom approximation [6]. In contrast to this and to the original sorting-based $\mathcal{O}(n \cdot \log(n))$ formulations [37], soft Spearman and Kendall correlation already involve an $\mathcal{O}(n^2)$ complexity for the comparison of each of n objects in the input and embedding space via the construction of soft indicator matrices. The total runtime for those approaches is thus $\mathcal{O}(n^3)$. It will be illustrated that rather good embeddings can be obtained even though many pairs of input relationships are not considered during the optimization. Thus, by using sparse input score matrices, the runtime is substantially reduced. For example, removing 50% of the input relationships leads to a quarter of runtime for evaluating the indicator matrices for thinned-out rows $\mathbf{S}_i$ and $\mathbf{D}_i^{\mathbf{Y}}$.

Soft correlation problems can be efficiently solved on GPUs, because the data transfer-to-processing ratio is $1/n$ to evaluate the indicator matrices. Problems with up to about 5000 items can be processed on consumer graphics boards within two days, and the break even point in favor of GPU over CPU is reached at about 500 items for a current quad-core machine. For large problems a speedup of a factor of eight of GPU against CPU is obtained for the provided MATLAB implementation.

4.3. Output standardization

The discussed embedding techniques do not provide a unique output, because correlation is a scale-invariant measure, and distance information is invariant to rotation and reflection. Thus, for standardizing rotational components, posterior PCA is applied to the mean-centered output points, i.e. without changing the dimensionality. For planar plots this defines a rotation to align the point cloud horizontally along the axis of maximum variance. Reflection invariance is addressed by flipping to positive skewness along the axes. Point clouds are finally scaled to attain maximum axis variance of one. These three actions are a useful subset of steps of the more general Procrustes analysis problems [38].

5. Quality measures

The assessment of data embeddings is a non-trivial task, because qualitative (appearance) and quantitative (neighborhood reconstruction performance) goals might be conflicting. Users may feel comfortable by looking at *a posteriori* labeled point clouds of embedded data. At the same time, calculated measures of neighborhood preservation may point out problems for local or global neighborhood sizes. Even worse, people look differently at point clouds, and quality measures for confusion, correlation or exact reconstruction of original and embedded neighborhoods yield values that are hard to compare. As a compromise, relaxed measures are used for comparing the overlap between neighborhoods in the original and the embedding space.

Embedding procedures yield sets of points for which the utility can be either visually assessed or by comparison of the obtained neighborhood configurations with the original neighborhoods. The co-ranking framework was designed exactly for such quantitative studies [39]. The framework allows one to measure not only neighborhoods ranking errors of the embedding, but also to describe the behavior if original neighborhoods are missed in the reconstruction ('extrusion') or if false neighborhoods are created ('intrusion').

Let $\phi_{ij} = |\{k : \mathbf{D}_{ik}^{\mathbf{Y}} \le \mathbf{D}_{ij}^{\mathbf{Y}}\}|$ be neighborhood ranks of object j given object i for the reconstructed distances of the embeddings, and let $\xi_{ij} = |\{k : \mathbf{S}_{ik} \ge \mathbf{S}_{ij}\}|$ be ranks of the input scores. Then the co-ranking matrix counting pairwise rank combinations is defined as

$$\mathbf{R} = [\mathbf{R}_{kl}]_{1 \le k,l \le n-1} \quad \text{with}$$
$$\mathbf{R}_{kl} = |\{(i,j) : \xi_{ij} = k \text{ and } \phi_{ij} = l\}| \in \{1 \ldots n\}. \tag{22}$$

Then, K-ary neighborhoods of size K are described by

$$U_{\mathrm{N}(K)} = \frac{1}{nK} \sum_{k=1}^{K} \sum_{l=k+1}^{K} \mathbf{R}_{kl},$$
$$U_{\mathrm{X}}(K) = \frac{1}{nK} \sum_{l=1}^{K} \sum_{k=l+1}^{K} \mathbf{R}_{kl},$$
$$U_{\mathrm{P}}(K) = \frac{1}{nK} \sum_{k=1}^{K} \mathbf{R}_{kk}. \tag{23}$$

For a perfectly embedded K-ary neighborhood all entries in the first K elements of the diagonal of the count matrix $\mathbf{R}$ are n. Thus, $1/nK$ is needed for scaling $U_{\mathrm{P}}(K)$, the paired matches, to a maximum of one. More generally, at most $n \cdot K$ counts can occur – in a mutually dependent way – in the $K \times K$ quadratic sub-matrix of $\mathbf{R}$. Entries might also occur in undesired far-away regions of the adjacent rectangular $K \times (n-1-K)$ and $(n-1-K) \times K$ sub-matrices counting missed or false neighborhood reconstruction. The expressions $U_{\mathrm{N}}(K)$ and $U_{\mathrm{X}}(K)$ describe mild intrusions and extrusions, respectively, which refer to the correct set of neighbors in the K-ary neighborhood up to permutation errors. Thus, the combination of all three descriptors

$$Q_{\mathrm{NX}}(K) = U_{\mathrm{N}}(K) + U_{\mathrm{X}}(K) + U_{\mathrm{P}}(K) \in [0,1] \tag{24}$$

characterizes the quality of reconstruction. The largest value of $Q_{\mathrm{NX}}(K) = 1$ indicates complete neighborhood retrieval which must not be confused with perfect reconstruction. The behavioral quantity

$$B_{\mathrm{NX}}(K) = U_{\mathrm{N}}(K) - U_{\mathrm{X}}(K) \in [-1,1] \tag{25}$$

becomes positive for intrusive and negative for extrusive embeddings. The measures $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ can be related to precision and recall in information retrieval: intrusion events are false positively created neighbors that decrease the precision, while extrusions decrease the recall. More details on the theoretical foundations of the co-ranking framework are found in separate works [39,40].

Using the co-ranking framework the distance matrix of the points embedded in the 2D visual plane is compared to the negative source similarity matrix in the experiments, that is, the original score ranking is again reversed for matching the attained distance ranks, and the assessments in (24) and (25) are sensitive to transposition of the score matrix.

6. Applications

Different scenarios of correlation-based multidimensional scaling are studied with three data sets. A synthetic two-dimensional data set allows for investigating reconstruction using different neighborhood models and sparse data. A 4096-dimensional face image database is used for dimension reduction purposes, and a protein database is used in the last application for processing asymmetric score data. Depending on the application, different methods are used for comparison, like non-metric MDS, kPCA, or SNE and t-SNE from the MATLAB toolbox for dimensionality reduction [2].

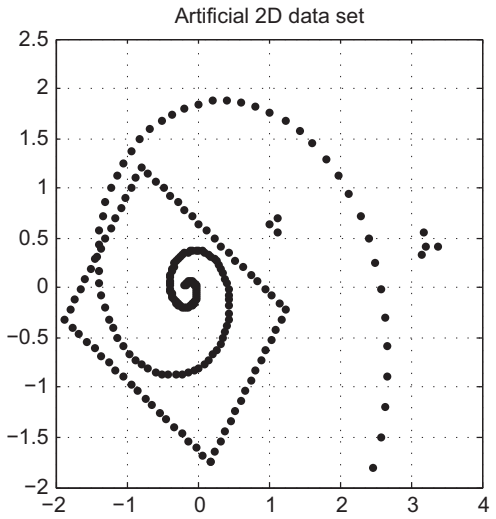

Fig. 2. Synthetic 2D data set featuring changing densities in the logarithmic spiral, an rectangle overlay with equi-distant points, and two mini clusters.

6.1. Synthetic data

An illustrative data set is used to demonstrate a number of properties of the correlation-based embedding techniques. Different from the 2D SwissRoll manifold designed for optimum embedding with geodesic distances by LLE and IsoMap [7,8] and from the hollow sphere designed for being torn and cut at fluctuating point densities on its surface by stochastic neighbor methods [14], the proposed data set exhibits more structure. A logarithmic spiral with increasing point distance is combined with a rectangle composed of equidistant points, and two mini-clusters occur as satellites. Since this 2D Euclidean data set shown in Fig. 2 contains an intelligible number of only 236 points, effects of re-embedding them into the plane can be studied by visual inspection. In a classical MDS context such data is not very valuable, because a perfect reconstruction can be expected. In neighbor embedding scenarios, though, it is not intuitively clear how well object-conditional neighborhood models perform in data reconstruction.

Perfect embedding results are obtained for all three correlation measures in the cbMDS framework for the trivial tasks of reconstructing the given distance matrix. These results are not shown here, but they can easily be verified by running the demo script included in the software package. We focus on potentially more interesting aspects in the following.

6.1.1. Reconstruction from neighborhood probabilities

From an ordering point of view distance-based neighbor ranks and neighborhood probability estimates should represent the same configuration for a given data set. Yet, ranks induce uniform distributions even under scalar shifts, scalings, or monotone transformations of the relational input data, while distance-based probability distributions are usually strongly skewed. Two experiments aim at the following aspects: a valid reconstruction of probability-induced neighbor ranks and the properties of reconstructions of symmetrized probability matrices. The former aspect takes relationships as a SNE view of the asymmetric neighborhood probability matrix and the latter represents the symmetric probability matrix as used by t-SNE. Both input matrices are estimated by Gaussian distributions by using the `x2p` function from the MATLAB toolbox for dimensionality reduction [2].

The top right panel in Fig. 3 contains the results by cbMDS for soft Spearman with $\kappa = 5$ for a typical perplexity value of $p = 15$ for the input neighborhood probability estimation. Soft Spearman is chosen as embedding method, because neighbor probabilities and ranks are closely related, while Pearson would reconstruct rather the raw probability values, and the triangle-based reconstruction in soft Kendall would not fully utilize the available complete ranking information.

Black diamond points in the top right panel in Fig. 3 show the very good results for the *asymmetric* input matrix, if compared to the original data set in Fig. 2. Mild bending artifacts occur only where the rectangle and the spiral are very close. Almost perfect rank-based embedding results, being supported by the excellent corresponding quality and behavior graph in the upper left panel, can be expected. This is because the Gaussian neighbor probability estimation is a monotonic function of the distance (radius). For comparison, the embedding results of SNE are found in the lower left panel. Due to the necessarily fixed perplexity value (here 15), this number of points is always accounted for in the neighborhood model. Such a forced inclusion of points turns the logarithmic spiral into a more linear spiral, and the edges of the rectangle are bent into curves. The general topology is validly reconstructed, though, including the two mini-clusters.

The colored points in the top right panel represent the cbMDS results for the *symmetric* neighborhood probability matrix as seen by t-SNE. Many distortions become apparent. A valid reconstruction of the original point configurations is not possible, because of the loss of information by symmetrization, that is, by averaging pairs of probabilities (p_{ij}, p_{ji}). Consequently, as seen in the lower right panel, a t-SNE embedding does not capture the overall topology correctly, but local features as rectangle edges and mini-clusters get reproduced. A large diversity of results is achieved by t-SNE, from which a representative one is shown. This points out the difficulty of solving this low-dimensional reconstruction problem by matching Gaussian and Student-t distributions of the neighborhood probabilities by KL divergence.

6.1.2. Embedding of sparse relationship data

Currently, unknown pairwise relationships are unsupported in most implementations of neighbor embedding methods. Setting the corresponding entries in the neighborhood probability matrix to zero would induce zero neighborhood probabilities in the embedding, but this would lead to repelling instead of neutral embedding forces for these relationships. Still, SNE and t-SNE are inherently sparse methods, because they look only for a given number of effective neighbors per row: at a perplexity of p and for n data points, there is a fraction of p/n significantly contributing entries in each row. Thus, most entries of the neighborhood probability matrix are close to zero if $p \ll n$.

The proposed correlation-based methods are implemented to truly ignore unknown input relationships by skipping quantitative contributions to the correlation measures for tagged entries of the original score matrix. This is a potentially useful feature, because generally pairwise object relationships contain vastly redundant information about relations in a data set for deriving still valid embeddings.

For example, the above set of 236 2D points is being represented by 27 730 informative distance pairs. We randomly removed 215 entries from each line of the original distance matrix (i.e. $\approx 91\%$ dropped) and applied cbMDS for soft Kendall and Spearman for the reconstruction. The results of cbMDS are shown in the middle two panels. Generally, rather good results are found, but the Kendall embedding looks a bit more scattered locally compared to the Spearman approach. This is also shown in the co-ranking quality graph, but undesired intrusion of Spearman can be observed for larger neighborhood sizes in the behavior graph. Generally, the success of reconstruction of unknown relationships

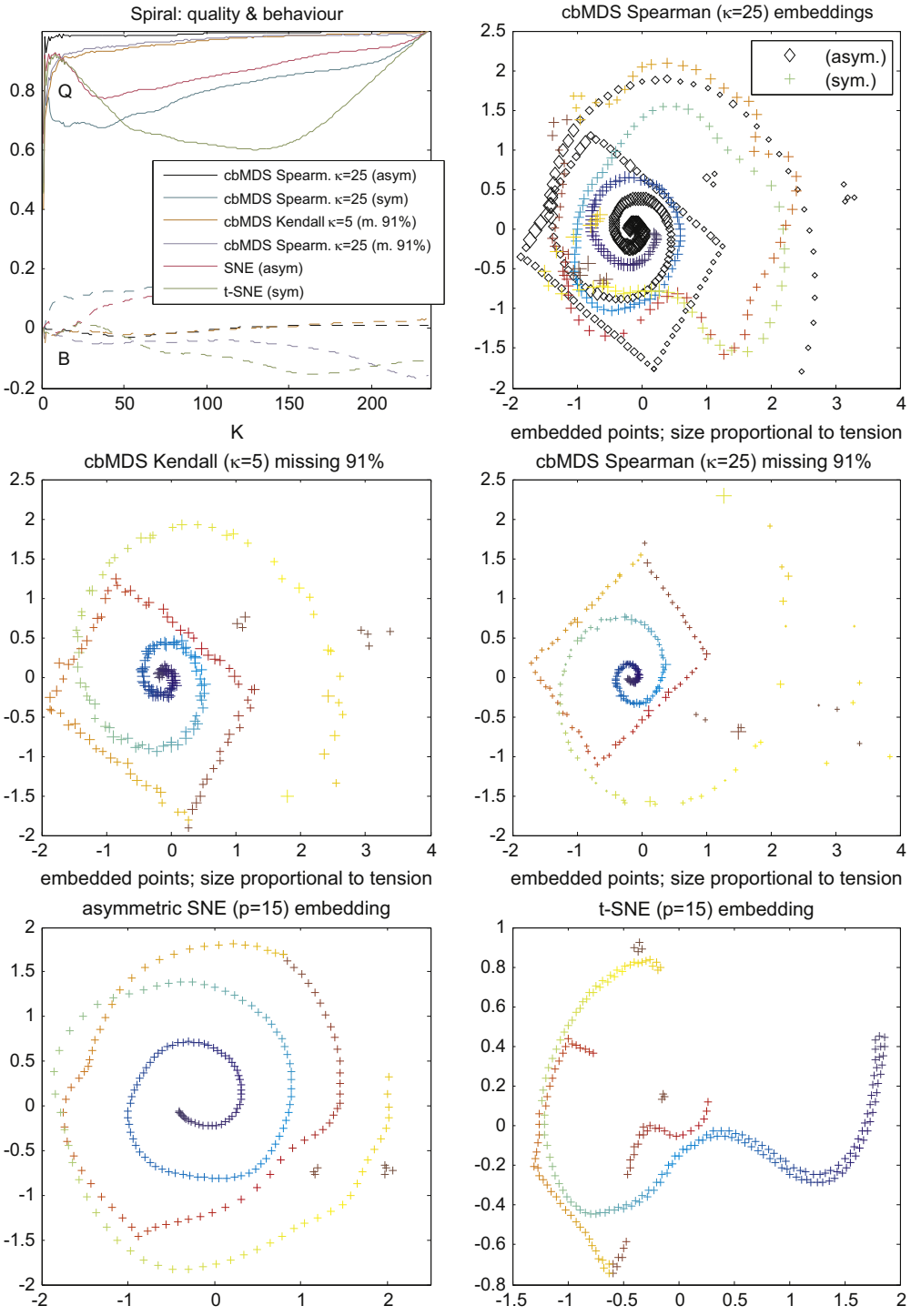

Fig. 3. Comparison of embeddings of the synthetic spiral data. In the top left co-ranking panel the quality and behavior values Q and B on the ordinate correspond to (24) and (25); Q and B are both plotted against neighborhood size K on the abscissa.

depends mostly on the data, and especially objects with naturally few relationships might be difficult to reconstruct under such an uncertainty.

6.1.3. *Olivetti faces data set*

The Olivetti faces database contains 400 gray images of 40 individuals in 10 different poses [41]. The number of dimensions is 4096 per image. In contrast to the previous synthetic reconstruction task, this one is a true dimension reduction problem, because negative pairwise Euclidean distances between these image vectors are used to constitute $\mathbf{S}_{ij}$. In this dimension reduction scenario, it is better to account for the distance information. This is better respected by Pearson correlation rather than by rank correlation measures. Again t-SNE is used for comparison. Results are shown in Fig. 4.

Pearson correlation exhibits a visually cluttered output in the top right panel, but the corresponding quality graph shows best performance beyond a neighborhood size of about 40. Since each person is represented by 10 faces, this means that faces of different persons are visually confused. For t-SNE a perplexity value of $p=15$ was used to allow for some tolerance in dealing with

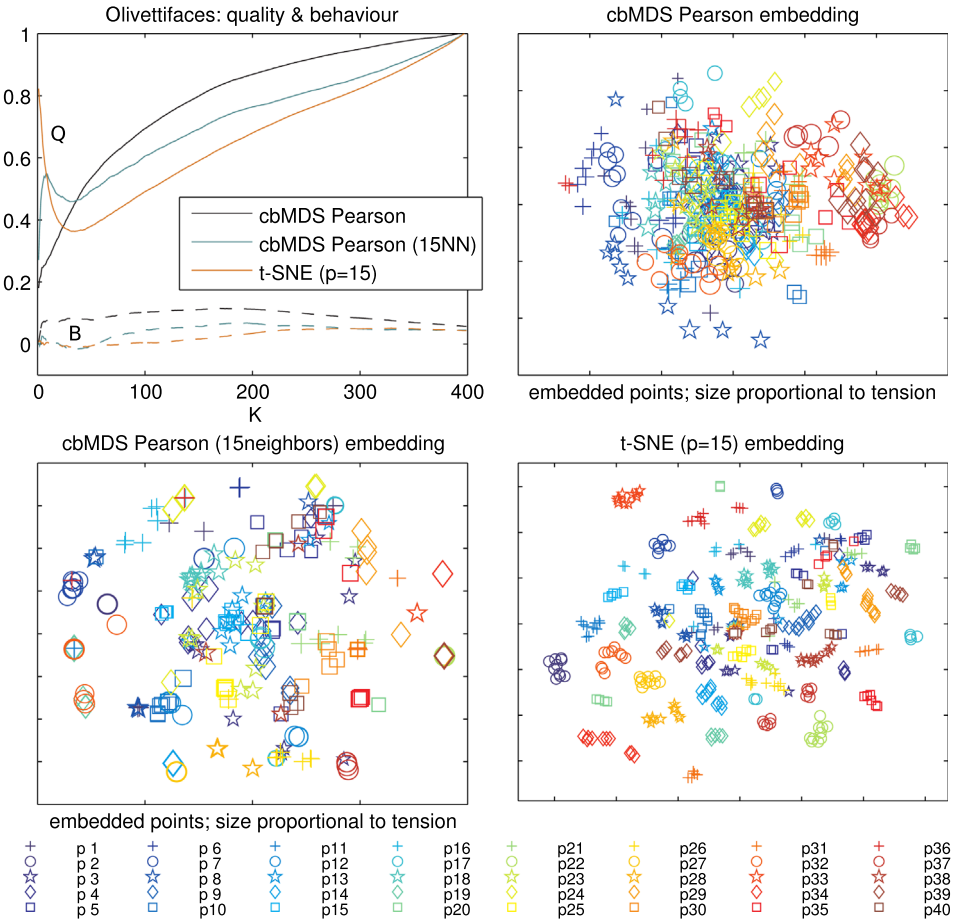

Fig. 4. Comparison of embeddings of the Olivetti faces data. Images of 40 people in different poses are referred by the 40 symbols in the scatter plots. Magnitudes of gradients can be identified as tension and they are shown by proportional point size in cbMDS.

between-person class boundaries. The output looks well-structured with strong separation of the 40 persons. Accordingly, the quality graph shows very high values, starting at 0.8, for very small neighborhood sizes of 8 points, but being less accurate on the overall scale. Such characteristic is to be expected for this prime example of what t-SNE was originally designed for: local neighborhood preservation in dimension reduction based on Euclidean distances.

To simulate the t-SNE concept of effective neighborhood size, per-face distance ranks were squashed by a mirrored sigmoid transformation, as shown in Fig. 5. The function approximates neighborhood probabilities of 1 for low distance ranks and 0 for high ranks with a turning point of 50% probability at a rank of 15, using a transition width of $\kappa = 2$. The resulting transformed distance matrix was used as score-type input data for cbMDS. Again Pearson correlation is employed, because results of soft Spearman and Kendall correlation do not possess the desired properties under this monotonic distance transformation. The resulting scatter plot is shown in the lower left panel. A structure is obtained that shares some properties of t-SNE and of the original distance embedding by cbMDS. This observation is confirmed by the quality and behavior graph being located between both. The straight-forward sigmoidal rank transformation led to the desired emphasis on local neighborhood reconstruction. At the same time, also global distance features are comparably well recovered. While t-SNE performs best for the important goal of local neighborhood reconstruction, cbMDS can be successfully modulated to focus on localized features and still achieve good average overall results.

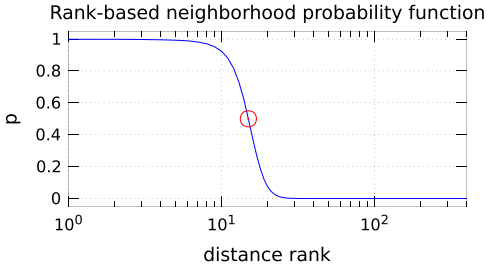

Fig. 5. Distance transformation function. Distances are turned into neighborhood probabilities by mapping their per-object ranks, here using $f(x) = (1 + \exp(-(\text{rank} - 15)/2))^{-1}$.

6.2. Protein data

The data set consists of 2900 samples of protein sequences taken as a random subset from the full Swiss-Prot database [42]. The database contains protein sequences of length from 30 to more than 1000 amino acids related to 32 common classes such as globin, cytochrome a and b, tubulin, kinases, as provided by the Prosite labeling [43]. The basic local alignment search tool (BLAST) was used to calculate pairwise sequence similarities, leading to an asymmetric scoring matrix because of local sequence comparisons [44]. The maximum score varies strongly between 8.5 and 6353 across all sequences, which indicates the need for object-conditional embedding techniques unless data are being normalized in a desirably appropriate preprocessing step.

Different embedding methods are used for comparing cbMDS with kernel PCA, non-metric MDS, and t-SNE. Since nonmetric multidimensional scaling and t-SNE rely on dissimilarity data **D** as inputs, original score data in **S** need to be transformed before their application. The following common score transformation is carried out, according to the work of Pekałska [45]:

$$\mathbf{D}_{ij} = \sqrt{\mathbf{S}_{ii} + \mathbf{S}_{jj} - \mathbf{S}_{ij} - \mathbf{S}_{ji}}. \tag{26}$$

The resulting symmetric dissimilarity matrix comes at the price of some loss of information which we can avoid by using cbMDS directly on the original score matrix. Note that t-SNE even suffers a second loss of information by involving a transformation of **D** into a symmetrized neighborhood probability matrix **P**.

Results are shown in Fig. 6 and summarized here:

- Soft Kendall correlation with a sigmoidal approximation of $\kappa = 5$ leads to a cross-like scatter point configuration with some protein classes becoming visible fairly well in the top right panel.
- Soft Spearman correlation and different choices of the sigmoidal approximation κ are not shown, because they turned out be similar to soft Kendall, but with slightly worse quality values. Pearson correlation yields the structurally different, more circular scatter plot shown in the middle left panel.
- As frequently observed, t-SNE provides a visually appealing plot with rather specific class assignments for a perplexity value of $p = 50$; this is displayed in the middle right panel.
- For the symmetrized BLAST score matrix $\frac{1}{2}(\mathbf{S}+\mathbf{S}^\top)$ taken as Gramian, rays are being formed for the three major protein classes by kernel PCA, as shown in the lower left panel. This is a clear visual result, but it collects most other protein classes in a singular spot near the coordinate origin.
- The scatter plot in the lower right panel for nonmetric scaling based on isotonic regression resembles convection patterns with protein classes appearing as neighbored patches and ribbons. Class members are rather contiguously grouped, but

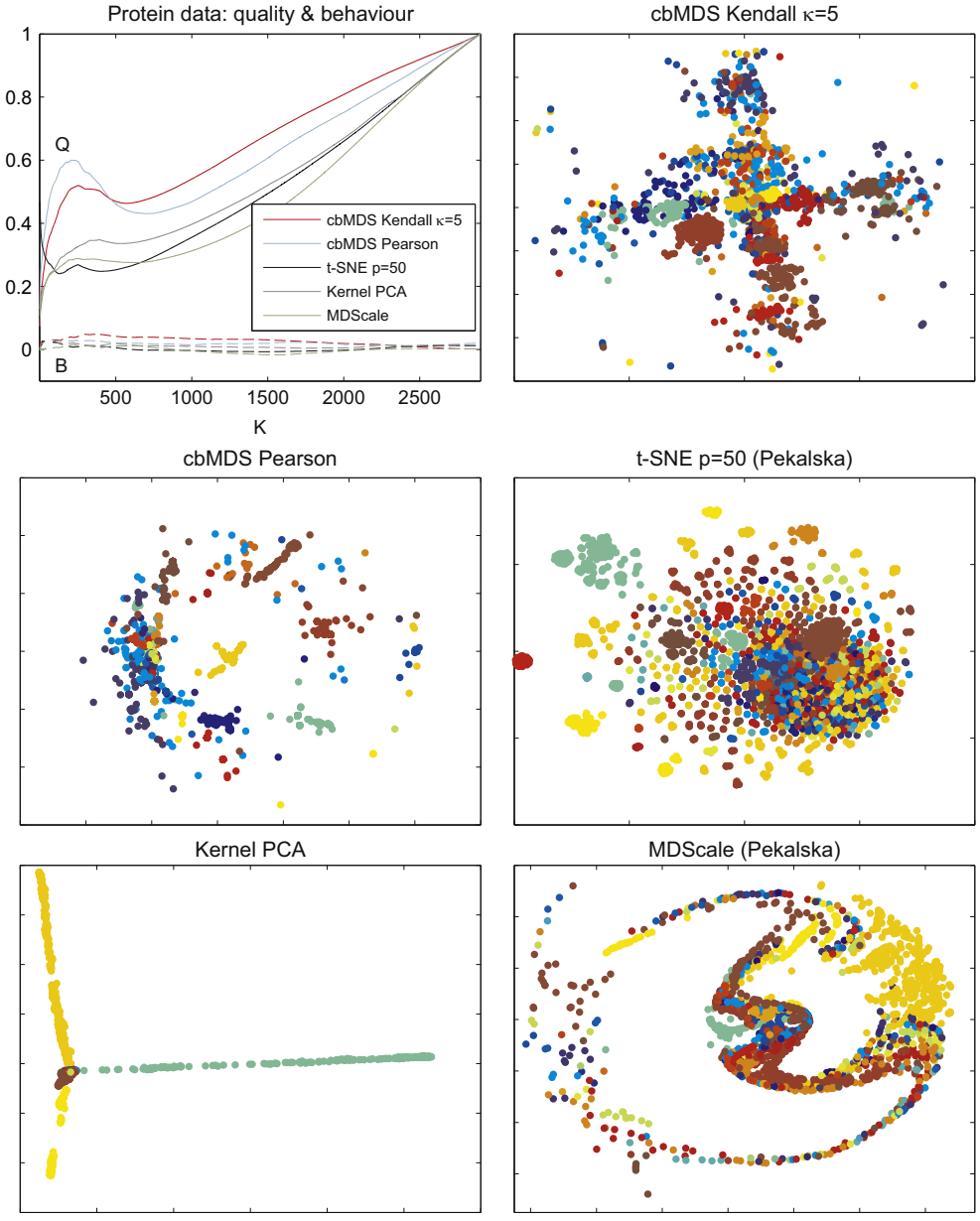

Fig. 6. Comparison of embeddings of the protein data. Colors indicate 32 different protein classes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

the overall arrangements are structurally different from the other embedding results. Note that 153 proteins are not shown, because they are placed widely scattered far beyond the chosen axis intervals.

- Finally, quality and behavior graphs for the visual outputs are combined in the upper left panel. According to its design, t-SNE performs best for small neighborhood sizes, but drops quality for larger neighborhoods. Pearson-based cbMDS performs very well for neighborhood sizes of about 180–200. Kendall-based cbMDS provides good average overall results. Kernel PCA focusses too much on three major protein classes and has a rather poor quality performance. MDScale, despite its generally interesting grouping, provides the worst results which may be partially caused by (26) for the score-to-dissimilarity transformation or its unconditional, i.e. non-object-specific, nature.

7. Discussion and outlook

The present work is supposed to strengthen the yet underrepresented method development of methods for directly processing relational, potentially asymmetric score data. Reconstructing optimized object neighborhoods, that is, data topology, via score ranking becomes feasible in the proposed soft Spearman and soft Kendall correlation approaches for embedding complex data relationships in a Euclidean space. Essentially, both correlation measures are based on soft counting operations modeled by a sigmoidal function. Theoretically, hard approximations of the step function would be desirable, but softer transitions turn out to be more robust and easier to optimize in practice.

A synthetic data set was used to illustrate distortion-free embeddings where state-of-the-art neighbor embedding methods would generate local neighborhood magnifications. Also, a perspective was taken on forced symmetries of inherently, though usually weakly, asymmetric neighbor relations. As an expected result, the original data topology could not be validly recovered for these structural modifications. Further experiments showed that a very good reconstruction is possible by the proposed soft-rank cbMDS methods, even after deleting more than 90% of the input relationships.

One of the current limitations of rank-based approaches is their weak statement about conceptually indistinguishable relations. For example, it is important to properly order high-scoring items, while low scores would not need good ordering precision, and rankings would even be misleading in the presence of noise. Robustness under noise is one of the major advantages of localized embedding methods like t-SNE, because items far beyond the effective neighborhood do not contribute much to the result. As shown for the face dimension reduction problem, it is possible to emulate such a behavior by using the proposed cbMDS with Pearson correlation on sigmoidal transformations of score ranks.

Processing of asymmetric pairwise protein scoring data turned out to be a great challenge. Embedding results of five different models looked very different. Two of the methods suffered from the need for symmetric dissimilarity data, thus, inducing lossy data transformation. A good compromise of local and global neighbor reconstruction quality was found for the developed soft Kendall cbMDS method.

Generally, the proposed correlation-based MDS methods are conceptually located between usual 'least-square' MDS and recent neighbor embedding approaches. The framework is rather generic with correlation taking over the role of scale-free mediator between input scores and embedding distances. Depending on the problem at hand, Pearson correlation can be applied, if input score distributions shall be accounted for. Alternatively, one of soft Spearman or Kendall correlation can be used for distribution-invariant reconstruction. Although specialized methods tend to perform better for specific tasks, the applications ranging from sparse reconstruction via dimension reduction to embedding of asymmetric score data illustrate how the cbMDS framework can be used for a versatile set of different tasks.

Implementations of the cbMDS methods discussed here, including data, are available at http://mloss.org as package 'cbMDS' for MATLAB/GNU-Octave.

Acknowledgment

This work was kindly supported by the Marburg Research Center for Synthetic Microbiology (SYNMIKRO), funded by the state Hesse, Germany, within the LOEWE initiative. F.-M. Schleif was kindly supported by a Marie Curie Intra-European Fellowship FP7-PEOPLE-2012-IEF (FP7-327791-ProMoS); additional funding was provided by the Cluster of Excellence277 CITEC funded in the framework of the German Excellence Initiative. Furthermore, the Finnish Centre of Excellence in Computational Inference Research (COIN) and the Helsinki Institute for Information Technology HIIT are thankfully acknowledged.

Appendix A. Derivative of $\bar{r}$

This section contains the gradient of $\bar{r}$ (cf. (7)) with respect to the point positions $\mathbf{y}_i$:

$$\boldsymbol{J}_{\bar{r}|-\mathbf{S}}(\mathbf{y}_i)=\frac{1}{n}\boldsymbol{J}_{r|-\mathbf{S}_i}(\mathbf{D}_i^{\mathbf{Y}})\boldsymbol{J}_{\mathbf{D}_i^{\mathbf{Y}}}(\mathbf{y}_i).$$

The notation $\boldsymbol{J}_{\bar{r}|-\mathbf{S}}(\mathbf{y}_i)$ refers to the Jacobian of the correlation function $\bar{r}$ given the scores $\mathbf{S}$ with respect to $\mathbf{y}_i$. The distance matrix derivatives are given by

$$\boldsymbol{J}_{\mathbf{D}_i^{\mathbf{Y}}}(\mathbf{y}_i)=\begin{pmatrix}\partial\mathbf{D}^{\mathbf{Y}}_{i1}/\partial\mathbf{Y}_i^1 & \dots & \partial\mathbf{D}^{\mathbf{Y}}_{i1}/\partial\mathbf{Y}_i^d\\ & \dots & \\ \partial\mathbf{D}^{\mathbf{Y}}_{in}/\partial\mathbf{Y}_i^1 & \dots & \partial\mathbf{D}^{\mathbf{Y}}_{in}/\partial\mathbf{Y}_i^d\end{pmatrix}$$

$$\text{with}\quad \frac{\partial\mathbf{D}^{\mathbf{Y}}_{ij}}{\partial\mathbf{Y}_i^k}=\frac{\mathbf{Y}_i^k-\mathbf{Y}_j^k}{\mathbf{D}^{\mathbf{Y}}_{ij}}$$

where $\mathbf{Y}_i^d$ denotes the dth attribute of $\mathbf{y}_i$.

Appendix B. Derivative of Pearson correlation r_P

This section contains the gradient of r_P (cf. (8)) with respect to $\mathbf{y}_i$, using the abbreviation $r_P:=\mathcal{B}/\sqrt{\mathcal{C}\cdot\mathcal{D}}$:

$$\frac{\partial r_P(\boldsymbol{w},\boldsymbol{u})}{\partial\boldsymbol{u}}=\boldsymbol{J}_{r_P|\boldsymbol{w}}(\boldsymbol{u})=r_P(\boldsymbol{w},\boldsymbol{u})\cdot\left(\frac{\boldsymbol{u}-\mu_{\boldsymbol{u}}}{\mathcal{B}}-\frac{\boldsymbol{w}-\mu_{\boldsymbol{w}}}{\mathcal{D}}\right)\tag{B.1}$$

$$\text{with}\ \mu_{\boldsymbol{w}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i\quad\text{and}\quad\mu_{\boldsymbol{u}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i.$$

Appendix C. Derivative of r_ρ

The standard deviation $\boldsymbol{\sigma_u}$ and its derivative are

$$\sigma_{\boldsymbol{u}}=\left(\frac{1}{n-1}\cdot\sum_{i=1}^{n}(\boldsymbol{u}_i-\mu_{\boldsymbol{x}})^2\right)^{1/2}\quad\text{and}\quad\frac{\partial\sigma_{\boldsymbol{u}}}{\partial\boldsymbol{u}_t}=\frac{\boldsymbol{u}_t-\mu_{\boldsymbol{x}}}{(n-1)\cdot\sigma_{\boldsymbol{u}}}$$

Derivatives for (11) and (12) are

$$\frac{\partial Z(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_k} = \frac{\partial \mathrm{sgd}_\kappa((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}})}{\partial \boldsymbol{u}_k} = \left(\frac{1}{\sigma_{\boldsymbol{u}}} - \frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial \sigma_u}{\partial \boldsymbol{u}_k}\right)\cdot \mathrm{sgd}'^{kl}_\kappa$$

$$\frac{\partial Z(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_l} = \frac{\partial \mathrm{sgd}_\kappa((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}})}{\partial \boldsymbol{u}_l} = \left(\frac{-1}{\sigma_{\boldsymbol{u}}} - \frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial \sigma_{\boldsymbol{u}}}{\partial \boldsymbol{u}_l}\right)\cdot \mathrm{sgd}'^{kl}_\kappa$$

$$\frac{\partial Z(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_m} = \frac{\partial \mathrm{sgd}_\kappa((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}})}{\partial \boldsymbol{u}_m} = -\frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial \sigma_{\boldsymbol{u}}}{\partial \boldsymbol{u}_m}\cdot \mathrm{sgd}'^{kl}_\kappa$$

$$\text{with}\quad \mathrm{sgd}'^{kl}_\kappa = \kappa\cdot \mathrm{sgd}^{kl}_\kappa\cdot(\mathrm{sgd}^{kl}_\kappa - 1).$$

The Jacobian of the soft rank $\boldsymbol{J}(\mathrm{rnk}(\boldsymbol{u}))$ is constructed by the above derivatives in corresponding to the proper summation indices in (11). Since n summations are carried out for which the Jacobian involves derivatives for all variables $\boldsymbol{u}_{1\ldots n}$, $\boldsymbol{J}(\mathrm{rnk}(\boldsymbol{u}))$ is an $n\times n$ matrix. The complete gradient vector of the soft Spearman rank correlation is given by

$$\frac{\partial r_\rho(\boldsymbol{w},\boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{J}_{r_P|\mathrm{rnk}(\boldsymbol{w})}(\mathrm{rnk}(\boldsymbol{u}))\boldsymbol{J}_{\mathrm{rnk}(\boldsymbol{u})}(\boldsymbol{u}). \tag{C.1}$$

Appendix D. Derivative of $\hat{r}_{\tau,\kappa}$

The sigmoid-based formulation (18) allows for gradient ascend optimization of Kendall τ with respect to the adaptive vector $\boldsymbol{u}$ assuming a fixed vector $\boldsymbol{w}$:

$$\frac{\partial \hat{r}_{\tau,\kappa}(\boldsymbol{w},\boldsymbol{u})}{\partial \boldsymbol{u}_k} = -\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial \mathrm{sgd}_\kappa\left(\frac{\boldsymbol{w}_i-\boldsymbol{w}_j}{\sigma_{\boldsymbol{w}}}\cdot\frac{\boldsymbol{u}_i-\boldsymbol{u}_j}{\sigma_{\boldsymbol{u}_k}}\right)}{\partial \boldsymbol{u}_k}, \tag{D.1}$$

where

$$\frac{\partial}{\partial \boldsymbol{u}_k}\mathrm{sgd}_\kappa\left(\frac{\boldsymbol{w}_i-\boldsymbol{w}_j}{\sigma_{\boldsymbol{w}}}\cdot\frac{\boldsymbol{u}_i-\boldsymbol{u}_j}{\sigma_{\boldsymbol{u}}}\right) = \frac{\partial}{\partial \boldsymbol{u}_k}\mathrm{sgd}_\kappa(z(\boldsymbol{u}_i,\boldsymbol{u}_j)) = \frac{\partial \mathrm{sgd}_\kappa(z)}{\partial z}\cdot\frac{\partial z(\boldsymbol{u}_i-\boldsymbol{u}_j)}{\partial \boldsymbol{u}_k}$$

$$\text{with } z(\boldsymbol{u}_i-\boldsymbol{u}_j) = \tilde{w}_{ij}\cdot\frac{\boldsymbol{u}_i-\boldsymbol{u}_j}{\sigma_{\boldsymbol{u}}} \quad\text{and}\quad \tilde{w}_{ij} = \frac{\boldsymbol{w}_i-\boldsymbol{w}_j}{\sigma_{\boldsymbol{w}}}.$$

The derivative of the sigmoid and the standard deviation is given by

$$\frac{\partial \mathrm{sgd}_\kappa(z)}{\partial z} = \kappa\cdot \mathrm{sgd}_\kappa(z)\cdot(\mathrm{sgd}_\kappa(z)-1)$$

$$\frac{\partial \sigma_{\boldsymbol{u}}}{\partial \boldsymbol{u}_k} = \frac{\boldsymbol{u}_k-\mu_{\boldsymbol{u}}}{(n-1)\sigma_{\boldsymbol{u}}}.$$

For the derivative of $z(\boldsymbol{u}_i,\boldsymbol{u}_j)$ we distinguish three cases:

$$\frac{\partial z(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_k} = \left(\frac{1}{\sigma} - \frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial \sigma_{\boldsymbol{u}}}{\partial \boldsymbol{u}_k}\right)\cdot \tilde{w}_{kl}$$

$$\frac{\partial z(\boldsymbol{u}_l,\boldsymbol{u}_k)}{\partial \boldsymbol{u}_k} = \frac{\partial z(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_k} \quad \text{because } (z(\boldsymbol{u}_l,\boldsymbol{u}_k) = z(\boldsymbol{u}_k,\boldsymbol{u}_l))$$

$$\frac{\partial z(\boldsymbol{u}_m,\boldsymbol{u}_l)}{\partial \boldsymbol{u}_k} = -\frac{\boldsymbol{u}_m-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial \sigma_{\boldsymbol{u}}}{\partial \boldsymbol{u}_k}\cdot \tilde{w}_{ml}$$

assuming $k\neq l$, $k\neq m$ and $l\neq m$. The standard deviation causes non-vanishing derivatives for the frequent last case in which the kth attribute does not appear in the difference part of z.

References

[1] D. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, H. Theisel, Selecting coherent and relevant plots in large scatterplot matrices, Comput. Graph. Forum 31 (6) (2012) 1895–1908, http://dx.doi.org/10.1111/j.1467-8659.2012.03069.x.

[2] L. van der Maaten, E. Postma, H. van den Herik, Dimensionality Reduction: A Comparative Review, Technical Report TiCC 2009-005, Tilburg University, NL, 2009 ⟨homepage.tudelft.nl/19j49/⟩.

[3] N. Halko, P.-G. Martinsson, Y. Shkolnisky, M. Tygert, An algorithm for the principal component analysis of large data sets, ArXiv e-prints http://arxiv.org/abs/1007.5510.

[4] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.

[5] B. Scholkopf, A. Smola, K.-R. Müller, Kernel principal component analysis, in: Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999, pp. 327–352.

[6] M. Li, J.T.-Y. Kwok, Making large-scale Nystro m approximation possible, in: Proceedings of the International Conference on Machine Learning (ICML), 2010, pp. 631–638.

[7] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[8] J. Tenenbaum, V. da Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[9] Y. Guo, J. Gao, P.W. Kwan, Kernel Laplacian eigenmaps for visualization of non-vectorial data, in: Proceedings of AI 2006: Advances in Artificial Intelligence, Springer, Berlin, Heidelberg, 2006, pp. 1179–1183, ⟨http://link.springer.com/chapter/10.1007/11941439_144⟩.

[10] S. France, J. Carroll, Two-way multidimensional scaling: a review, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 41 (5) (2011) 644–661, http://dx.doi.org/10.1109/TSMCC.2010.2078502.

[11] G. Hinton, S.T. Roweis, Stochastic neighbor embedding, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Neural Information Processing Systems 15 (NIPS), vol. 15, MIT Press, Cambridge, MA, USA, 2002, pp. 857–864.

[12] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[13] K. Bunte, S. Haase, M. Biehl, T. Villmann, Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences, Neurocomputing 90 (2012) 23–45, http://dx.doi.org/10.1016/j.neucom.2012.02.034.

[14] J.A. Lee, E. Renard, G. Bernard, P. Dupont, M. Verleysen, Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation, Neurocomputing 112 (2013) 92–108, http://dx.doi.org/10.1016/j.neucom.2012.12.036.

[15] B. Mokbel, S. Gross, M. Lux, N. Pinkwart, B. Hammer, How to quantitatively compare data dissimilarities for unsupervised machine learning? in: N. Mana, F. Schwenker, E. Trentin (Eds.), Artificial Neural Networks in Pattern Recognition, Lecture Notes in Computer Science, vol. 7477, Springer, Berlin, Heidelberg, 2012, pp. 1–13, ⟨http://dx.doi.org/10.1007/978-3-642-33212-8_1⟩.

[16] A. Gisbrecht, W. Lueks, B. Mokbel, B. Hammer, Out-of-sample kernel extensions for nonparametric dimensionality reduction, in: European Symposium on Artificial Neural Networks (ESANN), i6doc, Louvain-La-Neuve, Belgium, 2012, pp. 531–536.

[17] N. Bushati, J. Smith, J. Briscoe, C. Watkins, An intuitive graphical visualization technique for the interrogation of transcriptome data, Nucleic Acids Res. 39 (17) (2011) 7380–7389, http://dx.doi.org/10.1093/nar/gkr462 ⟨http://nar.oxfordjournals.org/content/39/17/7380.abstract⟩.

[18] M. Strickert, N. Sreenivasulu, B. Usadel, U. Seiffert, Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue, BMC Bioinformatics 8 (165) (2007), http://dx.doi.org/10.1186/1471-2105-6-76.

[19] J. Kruskal, Nonmetric multidimensional scaling: a numerical method, Psychometrika 29 (2) (1964) 115–129, http://dx.doi.org/10.1007/BF02289694.

[20] R.E. Barlow, H.D. Brunk, The isotonic regression problem and its dual, J. Am. Stat. Assoc. 67 (337) (1972) 140–147 ⟨http://www.jstor.org/stable/2284712⟩.

[21] M. Strickert, E. Hüllermeier, Neighbor embedding by soft Kendall correlation, in: M. Hlawitschka, T. Weinkauf (Eds.), Workshop Proceedings of Eurographics Conference on Visualization (EuroVis), 2013, pp. 1–5.

[22] M. Strickert, K. Bunte, Soft rank neighbor embeddings, in: M. Verleysen (Ed.), European Symposium on Artificial Neural Networks (ESANN), i6doc, Louvain-La-Neuve, Belgium, 2013, pp. 77–82.

[23] J. Venna, S. Kaski, Local multidimensional scaling, Neural Netw. 19 (6–7) (2006) 889–899.

[24] S. Lespinats, B. Fertil, P. Villemain, J. Hérault, RankVisu: mapping from the neighborhood network, Neurocomputing 72 (13–15) (2009) 2964–2978, http://dx.doi.org/10.1016/j.neucom.2009.04.008.

[25] V. Onclinx, J.A. Lee, V. Wertz, M. Verleysen, Dimensionality reduction by rank preservation, in: IJCNN, 2010, pp. 1–8. http://dx.doi.org/10.1109/IJCNN.2010.5596347.

[26] K. Bunte, M. Biehl, B. Hammer, A general framework for dimensionality reducing data visualization using explicit mapping functions, Neural Computation 24 (3) (2012) 771–804. http://dx.doi.org/10.1162/NECO_a_00250.

[27] J. Lee, M. Verleysen, Nonlinear Dimension Reduction, Springer, Berlin, Heidelberg, 2007.

[28] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization, J. Mach. Learn. Res. 11 (2010) 451–490.

[29] T. Müller, S. Rahmann, M. Rehmsmeier, Non-symmetric score matrices and the detection of homologous transmembrane proteins, Bioinformatics 17 (Suppl. 1) (2001) S182–S189 ⟨http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S182.short⟩..
[30] P. Cornbleet, M. Shea, Comparison of product moment and rank correlation coefficients in the assessment of laboratory method-comparison data, Clin. Chem. 24 (6) (1978) 857–861 ⟨http://www.clinchem.org/content/24/6/857.abstract⟩..
[31] G.B. Rabinowitz, An introduction to nonmetric multidimensional scaling, Am. J. Polit. Sci. 19 (2) (1975) 343–390 ⟨http://www.jstor.org/stable/2110441⟩.
[32] R.N. Shepard, Multidimensional scaling, tree-fitting, and clustering, Science 210 (4468) (1980) 390–398, http://dx.doi.org/10.1126/science.210.4468.390.
[33] K. Kampa, E. Hasanbelliu, J. Principe, Closed-form Cauchy–Schwarz PDF divergence for mixture of Gaussians, in: The 2011 International Joint Conference on Neural Networks (IJCNN), 2011, pp. 2578–2585. http://dx.doi.org/10.1109/IJCNN.2011.6033555.
[34] V.D. Silva, J.B. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, in: Advances in Neural Information Processing Systems (NIPS) 15, MIT Press, Cambridge, MA, USA, 2003, pp. 705–712.
[35] T.-Y. Liu, Learning to Rank for Information Retrieval, Springer, Berlin, Heidelberg, 2011.
[36] W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, in: R. Beals, A. Beck, A. Bellow, A. Hajian (Eds.), Contemporary Mathematics – Conference in Modern Analysis and Probability, vol. 26, American Mathematical Society, Providence, RI, Ann Arbor, MI, and Washington, DC, 1984, pp. 189–206. http://dx.doi.org/10.1090/conm/026.
[37] D. Christensen, Fast algorithms for the calculation of Kendall's τ, Comput. Stat. 20 (1) (2005) 51–62, http://dx.doi.org/10.1007/BF02736122.
[38] F.J. García-Fernández, M. Verleysen, J.A. Lee, I. Díaz, Stability comparison of dimensionality reduction techniques attending to data and parameter variations, in: M. Aupetit, L. van der Maaten (Eds.), EuroVis 2013 Workshop on Visual Analytics using Multidimensional Projections (VAMP), 2013, pp. 1–5.
[39] J. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72 (7–9) (2009) 1431–1443.
[40] B. Mokbel, W. Lueks, A. Gisbrecht, B. Hammer, Visualizing the quality of dimensionality reduction, Neurocomputing 112 (2013) 109–123.
[41] AT & T Laboratories Cambridge, Olivetti faces database, 1994, Original 112×92 images provided as 64×64 at http://www.cs.nyu.edu/~roweis/data.html (access: July 2013).
[42] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (1) (2003) 365–370, http://dx.doi.org/10.1093/nar/gkg095.
[43] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, A. Bairoch, ExPASy: the proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res. 31 (13) (2003) 3784–3788, http://dx.doi.org/10.1093/nar/gkg563.
[44] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410, http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
[45] E. Pekałska, R.P. Duin, The Dissimilarity Representation for Pattern Recognition: Foundations and Applications, Series in Machine Perception and Artificial Intelligence, vol. 64, World Scientific Publishing, London, Singapore, 2005.

Marc Strickert studied Applied Systems Science at the Institute of Ecosystems Research in Osnabrück (1994–2000). In 2004 he received his Ph.D. in Computer Science for work on self-organizing neural networks from the University of Osnabrück and worked as a junior research group leader in the field of bioinformatics at the Leibniz Institute of Crop Plant Research (IPK) in Gatersleben. Between 2010 and 2012 he was a member of the DFG graduate research school Imaging New Modalities at the University of Siegen, and he is currently working in the Computational Intelligence Lab of Eyke Hüllermeier at the University of Marburg, Germany.

Kerstin Bunte received her Ph.D. in Computer Science in December 2011 from the University of Groningen, The Netherlands. Her recent work has focused on Machine Learning techniques and their usability in the field of image processing, supervised dimension reduction, dissimilarity learning, classification and visualization. Research visits have taken her to Rochester (USA) and Bielefeld (Germany). Since 2013 she is working in the Department of Information and Computer Science at Aalto University in Finland. Further information can be obtained from http://www.users.ics.aalto.fi/kbunte/.

Frank-Michael Schleif received his Ph.D. in Computer Science from the University of Clausthal, Germany, in 2006. From 2004 to 2006 he was working for the R&D Department at Bruker Biosciences. From 2006 to 2009 he was a research assistant in the research group of computational intelligence at the University of Leipzig working on multiple bioinformatic projects. In 2010 he joined the chair of theoretical computer science and did research in multiple projects in machine learning and bioinformatics. Since 2014 he is a Marie Curie Fellow in the School of Computer Science at the University of Birmingham, UK. His areas of expertise include machine learning, signal processing, data analysis and bioinformatics. Several long term research stays have taken him to UK, the USA, the Netherlands and Japan. He is a co-editor of the Machine Learning Reports and a reviewer for multiple journals and conferences in the field of machine learning and computational intelligence. He is a founding member of the Institute of Computational Intelligence and Intelligent Data Analysis (CIID) e.V. (Mittweida, Germany), a member of the GI, the DAGM and secretary of the German chapter of the ENNS (GNNS). He is a coauthor of more than 70 papers in international journals and conferences on different aspects of Computational Intelligence, most of which can be retrieved from http://www.techfak.uni-bielefeld.de/~fschleif/.

Eyke Hüllermeier is with the Department of Mathematics and Computer Science at Marburg University (Germany), where he holds an appointment as a full professor and heads the Computational Intelligence Group. He holds M.Sc. degrees in Mathematics and Business Computing, a Ph.D. in Computer Science, and a Habilitation degree, all from the University of Paderborn (Germany). Prior to his current appointment, he held positions at the Universities of Dortmund, Toulouse and Magdeburg. His research interests are focused on methodological foundations of intelligent systems, especially on machine learning and reasoning under uncertainty, as well as applications of AI methods in the life sciences and other fields. Professor Hüllermeier is a Co-Editor-in-Chief of Fuzzy Sets and Systems, one of the leading journals in the field of Computational Intelligence, and serves on the editorial board of several other journals, including Machine Learning, IEEE Transactions on Fuzzy Systems, and International Journal on Approximate Reasoning.