ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Soft Rank Neighbor Embeddings

Marc Strickert[1] and Kerstin Bunte[2]

1 - Philipps Universität Marburg - Computational Intelligence Group, DE
2 - University of Bielefeld - CITEC Center of Excellence, DE;
Aalto University - Department of Information and Computer Science, FI

Abstract. Correlation-based multidimensional scaling is proposed for reconstructing pairwise dissimilarity or score relationships in a Euclidean space. Pearson correlation between pairs of objects in source and target space can be directly maximized by gradient methods, while gradient optimization of Spearman rank correlation profits from a numerically soft formulation introduced in this work. Scale and shift invariance properties of correlation help circumventing typical distance concentration problems.

1 Introduction

During the last decade, data embedding techniques are intensively studied for converting source data, defined in a relational way by pairwise scores or dissimilarities, into approximated relationships of a typically Euclidean space. Low-dimensional embedding spaces resulting from prominent techniques like Isomap and stochastic neighbor embedding allow for a substitutional visual inspection of original data relationships [1, 7]. Meaningful embeddings aim at an optimum reconstruction of the original relationships. Distance reconstruction is a well-known goal, but it cannot handle non-metric data like asymmetric score relationships. Therefore, a more general goal is the reconstruction of object-related similarity profiles in the input and embedding spaces. Pearson correlation for global (matrix-wide, matrix-conditioned) rather than for local (object-specific, row-conditioned) similarity was introduced for implementing high-throughput multidimensional scaling (HiT-MDS) [6]. Global comparisons may bare some problems, e.g., if score calculations depend on the size of structures compared; then, large structures might misleadingly yield larger scores than smaller, more tightly matching structures. In addition to turn symmetric matrix-conditioning into asymmetric row-conditioning, locally weighted rank correlation is introduced. Since neighborhood rankings of embedded Euclidean points are typically asymmetric, asymmetric similarity profiles can be reconstructed.

Rank correlation bears the challenge of optimizing discrete order relationships. Non-metric MDS based on isotonic regression (isoMDS) or approximate distance–rank mappings are possibilities to achieve efficient rank-based reconstructions [8, 4]. However, these are matrix-conditioned and/or restricted to symmetric dissimilarity data. For circumventing remaining non-differentiability properties of the generally well-suited fuzzy Kendall rank correlation [5] a soft formulation of Spearman's rank correlation is proposed and employed here for gradient-based optimization of correlation-based MDS (cbMDS).

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2 Soft Rank Embeddings

Instead of strict order optimization a soft version of the Spearman rank correlation is maximized between pairwise similarity relations of n objects in the input space $\boldsymbol{S}$ and their reconstructions $\boldsymbol{D}^{\boldsymbol{X}}$ as d-dimensional embedding points $\boldsymbol{X}$:

$$\bar{\mathrm{r}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{r}(-\boldsymbol{S}_i, \boldsymbol{D}_i^{\boldsymbol{X}}) = \max \tag{1}$$

The $n{\times}n$-matrix $\boldsymbol{D}^{\boldsymbol{X}}$ contains Euclidean distances $\boldsymbol{D}_{ij}^{\boldsymbol{X}} = \left(\sum_{k=1}^{d}(\boldsymbol{X}_k^i - \boldsymbol{X}_k^j)^2\right)^{1/2}$ of adjustable data-representing points $\boldsymbol{X}^i \in \mathbb{R}^d$, and signs of scores $\boldsymbol{S}$ are flipped to make smaller values express smaller 'distances'.

For gradient-based optimization the gradient of the correlation in Eq. (1) with respect to the i-th point locations $\boldsymbol{X}^i$ is provided as the product of the Jacobians of the inner and outer functions:

$$\boldsymbol{J}_{\bar{\mathrm{r}}|-\boldsymbol{S}}(\boldsymbol{X}^i) = \frac{1}{n} \cdot \boldsymbol{J}_{\mathrm{r}|-\boldsymbol{S}^i}(\boldsymbol{D}_i^{\boldsymbol{X}})\, \boldsymbol{J}_{\boldsymbol{D}_i^{\boldsymbol{X}}}(\boldsymbol{X}^i)\,. \tag{2}$$

The notation $\boldsymbol{J}_{\bar{\mathrm{r}}|-\boldsymbol{S}}(\boldsymbol{X}^i)$ refers to the Jacobian of function $\bar{\mathrm{r}}$ given fixed inverted scores $-\boldsymbol{S}$ with respect to $\boldsymbol{X}^i$. Eq. (2) contains distance matrix derivatives

$$\boldsymbol{J}_{\boldsymbol{D}_i^{\boldsymbol{X}}}(\boldsymbol{X}^i) = \begin{pmatrix} \partial \boldsymbol{D}_{i1}^{\boldsymbol{X}}/\partial \boldsymbol{X}_1^i & \dots & \partial \boldsymbol{D}_{i1}^{\boldsymbol{X}}/\partial \boldsymbol{X}_d^i \\ & \dots & \\ \partial \boldsymbol{D}_{in}^{\boldsymbol{X}}/\partial \boldsymbol{X}_1^i & \dots & \partial \boldsymbol{D}_{in}^{\boldsymbol{X}}/\partial \boldsymbol{X}_d^i \end{pmatrix} \text{ with } \frac{\partial \boldsymbol{D}_{ij}^{\boldsymbol{X}}}{\partial \boldsymbol{X}_k^i} = \frac{\boldsymbol{X}_k^i - \boldsymbol{X}_k^j}{\boldsymbol{D}_{ij}^{\boldsymbol{X}}}\,. \tag{3}$$

2.1 Attribute-weighted Pearson correlation

Linear correlations between two vectors $\boldsymbol{w}$ and $\boldsymbol{u}$ are measured by the Pearson correlation coefficient $\mathrm{r}_{\mathrm{P}}^{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{u}) \in [-1, 1]$ implementing $\boldsymbol{\lambda}$-weighted attributes:

$$\mathrm{r}_{\mathrm{P}}^{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{u}) = \frac{\sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 \cdot (\boldsymbol{w}_i - \mu_{\boldsymbol{w}}) \cdot (\boldsymbol{u}_i - \mu_{\boldsymbol{u}})}{\sqrt{\left(\sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 \cdot (\boldsymbol{w}_i - \mu_{\boldsymbol{w}})^2\right) \cdot \left(\sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 \cdot (\boldsymbol{u}_i - \mu_{\boldsymbol{u}})^2\right)}} =: \frac{\mathscr{B}}{\sqrt{\mathscr{C} \cdot \mathscr{D}}}\,. \tag{4}$$

The gradient with respect to the second argument vector $\boldsymbol{u}$ is given by [6]

$$\frac{\partial \mathrm{r}_{\mathrm{P}}^{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{J}_{\mathrm{r}_{\mathrm{P}}^{\boldsymbol{\lambda}}|\boldsymbol{w}}(\boldsymbol{u}) = \mathrm{r}_{\mathrm{P}}^{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{u}) \cdot \boldsymbol{\lambda} \circ \left(\frac{\boldsymbol{u} - \mu_{\boldsymbol{u}}}{\mathscr{B}} - \frac{\boldsymbol{w} - \mu_{\boldsymbol{w}}}{\mathscr{D}} \right) \quad \text{with} \tag{5}$$

$$\mu_{\boldsymbol{w}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \boldsymbol{w}_i\,, \quad \mu_{\boldsymbol{u}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \boldsymbol{u}_i \quad \text{and entrywise product } (\circ)\,. \tag{6}$$

Setting $\boldsymbol{w} = -\boldsymbol{S}_i$ and $\boldsymbol{u} = \boldsymbol{D}_i^{\boldsymbol{X}}$ the gradient in Eq. (5) can be plugged into Eq. (2) which allows to maximize the correlation between negative similarity scores and pairwise Euclidean distances of complex input space and low-dimensional embedding, respectively. Localized influence of attributes, i.e. neighbors, is achieved for $\boldsymbol{\lambda}_i \neq 1$. Rather than seeking a diagonal in the Shepard diagram, i.e. least squares distance reconstructions, straight lines with any slope (distance scaling)

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

and intercept (distance shift) are allowed here to effectively by-pass the distance concentration problem. This is the basic idea of High-Throughput Multidimensional Scaling (HiT-MDS) [6]. Non-Euclidean input data relationships, though, are more reliably handled by looking at order relationships, as discussed next.

2.2 Soft weighted 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 being then applied to the Pearson correlation in Eq. (4):

$$\mathrm{r}_\rho(\boldsymbol{w},\boldsymbol{u}) = \mathrm{r}_\mathrm{P}^{\boldsymbol{\lambda}}(\mathsf{rnk}(\boldsymbol{w}),\mathsf{rnk}(\boldsymbol{u})) \tag{7}$$

Instead of utilizing a sorting operation, the ranking of vector elements in $\boldsymbol{u}$ can be alternatively achieved by summing up rows of the indicator matrix $\boldsymbol{R}$:

$$\mathsf{rnk}(\boldsymbol{u}) = \begin{pmatrix} \sum_{i=1}^n \mathsf{R}(\boldsymbol{u}_1,\boldsymbol{u}_i) \\ \dots \\ \sum_{i=1}^n \mathsf{R}(\boldsymbol{u}_n,\boldsymbol{u}_i) \end{pmatrix} \text{ for } \boldsymbol{R}(\boldsymbol{u}) = \begin{pmatrix} \mathsf{R}(\boldsymbol{u}_1,\boldsymbol{u}_1) & \dots & \mathsf{R}(\boldsymbol{u}_1,\boldsymbol{u}_n) \\ & \dots & \\ \mathsf{R}(\boldsymbol{u}_n,\boldsymbol{u}_1) & \dots & \mathsf{R}(\boldsymbol{u}_n,\boldsymbol{u}_n) \end{pmatrix} . \tag{8}$$

For the Heaviside step function $\mathsf{R}(\boldsymbol{u}_k,\boldsymbol{u}_l) = H(\boldsymbol{u}_k - \boldsymbol{u}_l)$, providing zero for negative arguments and else one, correct ranks are obtained for vector elements $\boldsymbol{u}_k$ in the absence of ties. Using the standard deviation $\sigma_{\boldsymbol{u}}$ and its derivative

$$\sigma_{\boldsymbol{u}} = \Big(\frac{1}{n-1}\cdot\sum_{i=1}^n(\boldsymbol{u}_i-\mu_{\boldsymbol{u}})^2\Big)^{1/2} \quad \text{with} \quad \frac{\partial\sigma_{\boldsymbol{u}}}{\partial\boldsymbol{u}_t} = \frac{\boldsymbol{u}_t-\mu_{\boldsymbol{u}}}{(n-1)\cdot\sigma_{\boldsymbol{u}}} \tag{9}$$

the step function $H(\boldsymbol{u}_k - \boldsymbol{u}_l)$ can be replaced by a differentiable sigmoid

$$\mathsf{R}(\boldsymbol{u}_k,\boldsymbol{u}_l) = \mathsf{sgd}_\kappa^{kl} + \frac{1}{2} = \mathsf{sgd}_\kappa\Big(\frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}}\Big) + \frac{1}{2} = \frac{1}{1+e^{\kappa\cdot(\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}}} + \frac{1}{2} \tag{10}$$

with mid-tied ranks being approximated for $\kappa\to\infty$. Thus, large κ are preferred, but $5<\kappa<100$ is numerically sensible. Derivatives for Eqns. 8 and 10 are

$$\begin{aligned}
\frac{\partial\mathsf{R}(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial\boldsymbol{u}_k} &= \frac{\partial\mathsf{sgd}_\kappa\left((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}\right)}{\partial\boldsymbol{u}_k} = \left(\frac{1}{\sigma_{\boldsymbol{u}}} - \frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial\sigma_{\boldsymbol{u}}}{\partial\boldsymbol{u}_k}\right)\cdot\mathsf{sgd}_\kappa'^{kl} \\
\frac{\partial\mathsf{R}(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial\boldsymbol{u}_l} &= \frac{\partial\mathsf{sgd}_\kappa\left((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}\right)}{\partial\boldsymbol{u}_l} = \left(\frac{-1}{\sigma_{\boldsymbol{u}}} - \frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial\sigma_{\boldsymbol{u}}}{\partial\boldsymbol{u}_l}\right)\cdot\mathsf{sgd}_\kappa'^{kl} \qquad (11) \\
\frac{\partial\mathsf{R}(\boldsymbol{u}_k,\boldsymbol{u}_l)}{\partial\boldsymbol{u}_m} &= \frac{\partial\mathsf{sgd}_\kappa\left((\boldsymbol{u}_k-\boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}\right)}{\partial\boldsymbol{u}_m} = -\frac{\boldsymbol{u}_k-\boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}^2}\cdot\frac{\partial\sigma_{\boldsymbol{u}}}{\partial\boldsymbol{u}_m}\cdot\mathsf{sgd}_\kappa'^{kl} \\
\mathsf{sgd}_\kappa'^{kl} &= \kappa\cdot\mathsf{sgd}_\kappa^{kl}\cdot\Big(\mathsf{sgd}_\kappa^{kl}-1\Big) . \qquad (12)
\end{aligned}$$

The Jacobian of the soft rank $\boldsymbol{J}_{\mathsf{rnk}(\boldsymbol{u})}$ is constructed by the derivatives in Eq. (11) corresponding to the proper summation indices in Eq. (8). Since n summations are carried out for which the Jacobian involves derivatives for all variables $\boldsymbol{u}_{1\dots n}$, $\boldsymbol{J}_{\mathsf{rnk}(\boldsymbol{u})}$ is an $n\times n$ matrix.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Finally, the gradient vector of the soft Spearman rank correlation is

$$\frac{\partial \mathrm{r}_\rho(\boldsymbol{w},\boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{J}_{\mathrm{r}_\mathrm{P}^{\boldsymbol{\lambda}}|\mathsf{rnk}(\boldsymbol{w})}(\mathsf{rnk}(\boldsymbol{u}))\, \boldsymbol{J}_{\mathsf{rnk}(\boldsymbol{u})}(\boldsymbol{u})\,. \tag{13}$$

Substituting $\boldsymbol{w} = -\boldsymbol{S}_i$ and $\boldsymbol{u} = \boldsymbol{D}_i^{\boldsymbol{X}}$ this gradient is plugged into Eq. (2) for optimizing a point set $\boldsymbol{X}$ with ranks of Euclidean relationships best matching the ranks of original data relationships. In practice, the memory-limited quasi-Newton l-BFGS gradient optimization scheme provides rapid convergence.

2.3 Experiments

An artificial and a real-world data set are embedded in a two-dimensional space, the first one for demonstrating the general validity of rank-based reconstruction, the second one showing the potential for complex relationships in protein data.

2.3.1 Artificial 2D data set

For neighborhood-preserving embedding techniques it is an interesting exercise to carry out embeddings of known 2D relationships in a 2D reconstruction space. Deviations from a perfect reconstruction indicate potential bias in the neighborhood model. A set of 236 points is considered that form two overlaid shapes, a logarithmic spiral and a rectangle, and two small clusters of three and four points. Thus, irregular spacing (spiral) and tied distances (rectangle) complicate the embedding task as well as the preservation of the two small clusters.

Figure 1 shows the embedding quality and behavior, referring to information-theoretic assessment of neighborhood retrieval [2], and example embeddings for this artificial data set. The proposed cbMDS approach, tSNE with perplexity 10 and the weighted cbMDS approach with $\boldsymbol{\lambda}_i = \exp\left(\frac{-\boldsymbol{S}_i^2}{(0.5\cdot n)^2}\right)$ are compared. Every algorithm was run 10 times with random initialization. Panel A1 shows the mean quality (Q), 1 being maximum, and behavior (B), 0 being least biased, for different neighborhood sizes K, and the standard deviation over the 10 runs. Panel A2 to A4 show embeddings of the different techniques. The cbMDS (A2) reconstructions are almost perfect, just with little tension near the center of the spiral, while tSNE (A3) breaks the structures more and more apart the smaller the perplexity is chosen. If cbMDS is forced to favor small neighborhoods by the above $\boldsymbol{\lambda}$-weighting, this leads to the fragmented result in panel A4.

2.3.2 SCOP protein data set

A real-world database is visualized containing structural classification of proteins (SCOP), being online available as supplemental material [3]. It contains p-values of pairwise Smith-Waterman alignments of 4352 proteins using an asymmetric substitution matrix. The 4352×4352 matrix is asymmetric, with smaller entries indicating higher similarity. The matrix covers a broad spectrum of protein families with 2888 hierarchically organized unique labels. For ease of experimenta-

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

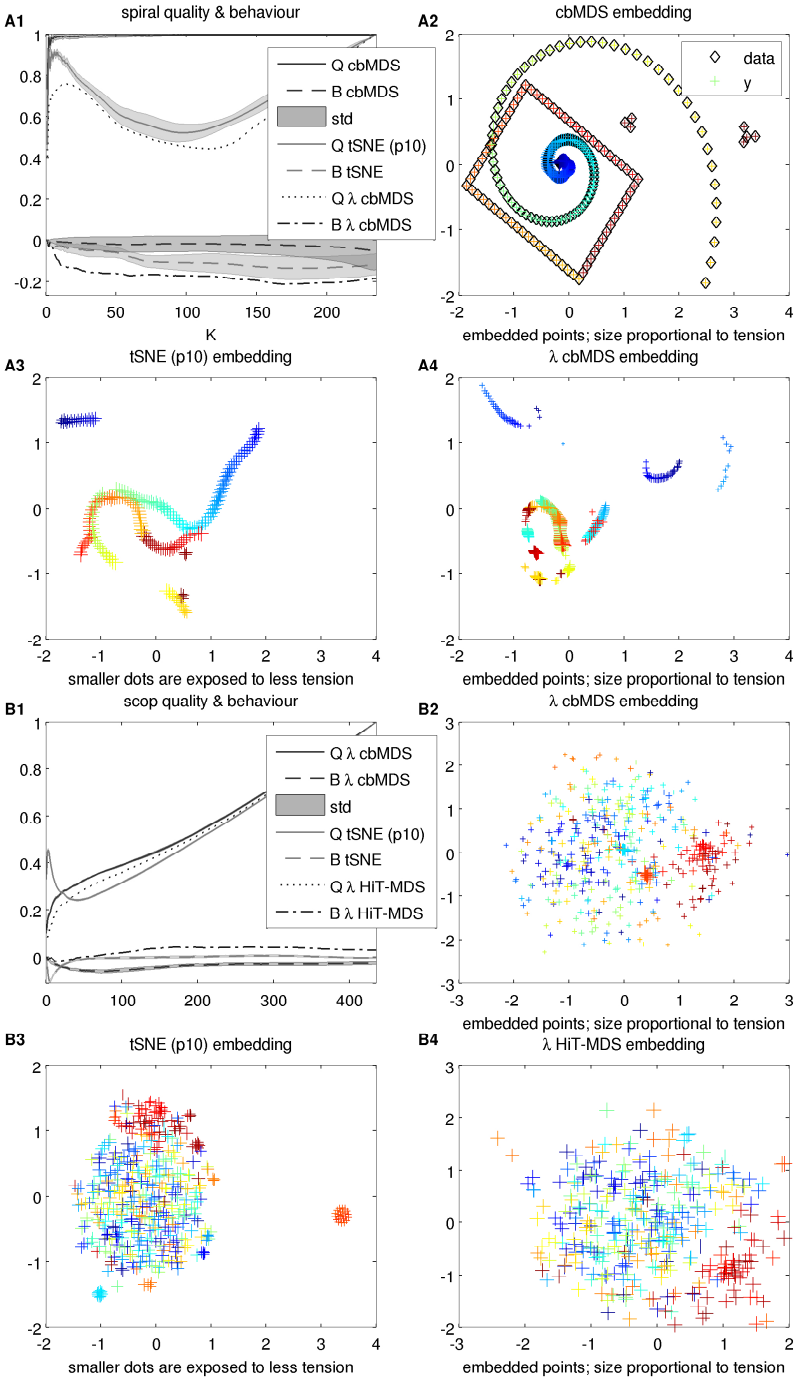

Fig. 1: Quality plots and example embeddings for the two data sets: Artificial 2D (A1-A4) and SCOP (B1-B4). Details are described in the text.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

tion we subset the database by picking each $10th$ alignment pair probability of columns and rows ordered ascendingly by their protein identifier.

The panels B1–4 in Figure 1 show the results of the SCOP experiments. For small neighborhoods tSNE shows highest quality and highlights potential protein clusters; yet, quality drops rapidly in the range of $K = 4$–41. Weighted cbMDS with $\boldsymbol{\lambda}$ set like before has a lower quality in small neighborhoods, but is constantly increasing until it outperforms tSNE at $K > 20$. HiT-MDS exhibits characteristics similar to cbMDS but at higher tension levels.

2.4 Conclusions

A correlation-based relational score embedding scheme has been introduced that maximizes correlations between potentially asymmetric object similarities in the source and embedding space. Using a soft formulation of Spearman rank correlation, gradient-based optimization schemes can be successfully applied for reconstruction of the neighborhood rank order. In terms of co-ranking criteria a comparison with tSNE shows a better overall performance of cbMDS. The visual results for the protein data are more appealing in tSNE, thanks to its good local neighborhood reconstruction, but for the identity reconstruction in 2D tSNE suffers from that inevitable local bias.

Soft rank optimization has a general impact on machine learning problems. Computational demands related to the indicator matrix and its Jacobian can be lowered by computational capabilities of graphics processing units (GPU). Future work is needed to characterize potential invariances when turning distances into ranks and loss of information when transforming hard into soft ranks. A MATLAB/GNU-Octave package with GPU support is online available as package 'cbMDS' at https://mloss.org/.

This work was supported by the "German Science Foundation (DFG)" under grant number HA-2719/4-1 and by the LOEWE center for Synthetic Microbiology SYNMIKRO. Financial support from the Cluster of Excellence 277 Cognitive Interaction Technology (CITEC) is gratefully acknowledged.

References

[1] G. Hinton and S. T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, volume 15, pages 857–864. MIT Press, 2002.

[2] J. A. Lee and M. Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31:2248–2257, October 2010.

[3] L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Journal of Computational Biology*, 10(6):857–868, 2004.

[4] V. Onclinx, J. Lee, V. Wertz, and M. Verleysen. Dimensionality reduction by rank preservation. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8, 2010.

[5] M. D. Ruiz and E. Hüllermeier. A formal and empirical analysis of the fuzzy gamma rank correlation coefficient. *Information Sciences*, 206:1–17, 2012.

[6] M. Strickert, F.-M. Schleif, T. Villmann, and U. Seiffert. Unleashing Pearson correlation for faithful analysis of biomedical data. In M. B. et al. editor, *Similarity-Based Clustering – Recent Developments and Applications*, volume 5400 of *LNCS*, pages 70–91. Springer, 2009.

[7] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[8] W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, $4th$ edition, 2002.