Eurographics Conference on Visualization (EuroVis) (2013)
M. Hlawitschka and T. Weinkauf (Editors)

Neighbor Embedding by Soft Kendall Correlation

Marc Strickert & Eyke Hüllermeier

Computational Intelligence, Philipps Universiät Marburg, Germany

Abstract

Correlation-based embedding of complex data relationships in a Euclidean space is studied. The proposed soft formulation of Kendall correlation allows for gradient-based optimization of scatter point neighborhood relationships for reconstructing original data neighbors. The approach is able to handle asymmetric data relations provided in the form of a general scoring matrix. Scale and shift invariance properties of correlation help circumventing typical embedding distortion artefacts in dimension reduction and data embedding scenarios.

1. Introduction

Data embedding techniques for casting complex source data defined in a relational way into assessable Euclidean spaces have attracted attention during the last years. Distance matrix reconstruction is a well-known goal of common multi-dimensional scaling approaches [FC11, BT12], but they cannot handle asymmetric score relationships. If proper conversion from scores to dissimilarity data is provided along with an effective neighborhood size for local density estimation, then modern packages like the neighbor retrieval visualizer with data density estimation allow for dealing with asymmetric relationships [VPN*10]. Also other prominent techniques like Isomap and stochastic neighbor embedding allow for creating low-dimensional embedding spaces for visual inspection of generic dissimilarity data relationships [TdSL00, HR02, vdMH08], but most applications found in the literature still refer to dimension reduction of Euclidean data. Standard methods in MATLAB and R, mdscale and isoMDS, respectively, based on isotonic regression, are restricted to symmetric dissimilarity matrices [VR02], and a recent method for rank-preserving embedding was again only tested on dimension reduction problems [OLWV10].

A focus of the present work is the non-parametric reconstruction of general object-related similarity profiles in the input and embedding spaces. Some years ago, Pearson correlation was introduced as global (matrix-wide, matrix-conditioned) linear similarity measure between source dissimilarity matrix and target distance matrix [SSVS09]. Global matrix-conditioned comparisons may bare some problems though. For example, if score calculations depend on the size of the compared structures, large structures might misleadingly yield larger scores than smaller, more tightly matching structures. Here, row-conditioning is studied which naturally induces pairwise neighborhood order comparisons between object-specific source and target scores. The maximization of the count-based non-linear Kendall correlation between these pairs of score profiles is a natural optimization target for locating the variable embedding points. Since Kendall correlation is a non-differentiable measure with complex optimization structure, a soft formulation of Kendall's τ is proposed and employed here for gradient-based optimization of correlation-based MDS.

2. Correlation-based neighbor reconstruction

Let n data items be represented by a $n \times n$–matrix containing pairwise comparison scores S. Then another matrix D^X is sought such that its contained pairwise similarities of points reconstructed in a low-dimensional Euclidean space is in best correspondence with the source score matrix. Correlation is a natural measure of correspondence for which common trends show up as high values, and its maximization can be utilized as essential neighbor embedding operation. Formally, each object is described by its relationship to $(n-1)$ neighbors, thus, the averaged correlations r of source and embedding neighborhoods should be maximized along all rows i of the corresponding matrices:

$$\bar{\mathrm{r}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{r}(-S_i, D_i^X) \rightarrow \max . \tag{1}$$

The $n \times n$-matrix D^X actually contains Euclidean distances

$$D_{ij}^X = \mathbf{d}(X^i, X^j) = \left(\sum_{k=1}^{d} (X_k^i - X_k^j)^2\right)^{\frac{1}{2}} \tag{2}$$

DOI: 10.2312/PE.EuroVisShort.EuroVisShort2013.073-077

of d-dimensional adjustable data-representing points $X^l \in \mathbb{R}^d$, and signs of original scores S are flipped to make smaller values appear as smaller 'distances'. This way, row-wise matching targets are established between ordinal neighborhood ranks of original and embedded objects.

If the source matrix already would already contain Euclidean distances and squared Euclidean distance was used for minimization instead of r, Eq. 1 results in classical multidimensional scaling [Gow66]. Rather than seeking a diagonal in the Shepard diagram, i.e. least squares distance reconstructions, straight lines with any slope (distance scaling) and intercept (distance shift) are allowed in correlation-based optimization to effectively by-pass common distance concentration problems [vdMH08]. If source data were dissimilarity measures and r Pearson correlation, this would be related to high-throughput multidimensional scaling (HiTMDS) [SSVS09]. Recently, a soft ordinal ranking approach was established and combined with Pearson correlation, leading to a soft version of Spearman rank correlation [SB13]. Here, we propose a rather direct approximation of Kendall correlation that requires order relationships only between pairs of objects in the neighborhood of source and embedding space.

2.1. Soft Kendall correlation

The Kendall correlation coefficient r_τ compares two vectors $w = (w_k)$ and $u = (u_k)$, $k = 1 \ldots n$, of identical dimension n by assessing the local ordering of all pairs (i, j) of their elements and counting the number of concordant (C_{ij}) and discordant (D_{ij}) outcomes [Ken38]:

C_{ij}: $(w_i > w_j \wedge u_i > u_j) \vee (w_i < w_j \wedge u_i < u_j)$
D_{ij}: $(w_i > w_j \wedge u_i < u_j) \vee (w_i < w_j \wedge u_i > u_j)$

Let $\#C$ be the number of true values for C_{ij} and $\#D$ be the number of true values for D_{ij} for all $i, j \in \{1, \ldots, n\}$. Ignoring attribute pairs for $i = j$, the maximum number of mutually exclusive concordant and discordant pairs is $\#C + \#D = (n \cdot n - n)/2$. Thus, a normalized difference of concordant and discordant pair counts is used to quantify trends of positive or negative correlation:

$$\mathrm{r}_\tau(w,u) = \frac{\#C - \#D}{\frac{1}{2} \cdot n \cdot (n-1)} \in [-1;1]\,. \tag{3}$$

This common definition of Kendall τ correlation does not consider ties, and, for simplicity, we assume their absence in the following.

The proposed formulation of a soft version of Kendall τ is based on products of differences $\tilde{p}_{ij} = (w_j - w_i) \cdot (u_i - u_j)$ which are negative for concordant attribute pairs, and positive for discordant pairs.

Let R be a matrix constructed from all pairs

$$R(\tilde{p}) = \begin{pmatrix} \mathsf{R}(\tilde{p}_{11}) & \ldots & \mathsf{R}(\tilde{p}_{1n}) \\ & \ldots & \\ \mathsf{R}(\tilde{p}_{n1}) & \ldots & \mathsf{R}(\tilde{p}_{nn}) \end{pmatrix}. \tag{4}$$

Then for the step function $\mathsf{R}(x) = 1/2 \cdot (\mathsf{sign}(z) + 1)$, providing zero for negative arguments, 1/2 for zero and else one, Kendall correlation $\mathrm{r}_\tau(w,u) = \hat{\mathrm{r}}_{\tau,\kappa}(w,u)$ is recovered for

$$\hat{\mathrm{r}}_{\tau,\kappa}(w,u) = 1 - \frac{\left(\sum_{i=1}^n \sum_{j=1}^n \mathsf{R}(\tilde{p}_{ij})\right) - q}{\frac{1}{2} \cdot n \cdot (n-1)} \tag{5}$$

with $q = 0$. Basically, twice the relative amount of discordant pairs is subtracted from the highest possible correlation value. This approach is valid, because untied data induces an complementary amount of concordant values. A non-zero value of q is needed for the differentiable sigmoidal approximation of the step function to be plugged into Eq. 5:

$$\mathsf{R}(p_{ij}) = \mathsf{sgd}_\kappa(p_{ij}) = \frac{1}{1 + e^{\kappa \cdot p_{ij}}} \quad \text{with} \tag{6}$$

$$p_{ij} = \frac{\tilde{p}_{ij}}{\sigma_w \cdot \sigma_u} = \frac{w_i - w_j}{\sigma_w} \cdot \frac{u_i - u_j}{\sigma_u} \tag{7}$$

The larger κ, the steeper the transition from zero to one gets in the sigmoid sgd_κ, that is, $\lim_{\kappa \to \infty} \hat{\mathrm{r}}_{\tau,\kappa}(w,u) = \mathrm{r}_\tau(w,u)$. In Eq. 5 a value of $q = n/2$ is required to compensate for the fact that zero differences occurring for $i = j$ get counted as partial discordance by $\mathsf{sgd}_\kappa(0) = 1/2$, irrespective of κ.

The variable substitution in Eq. 7 helps equalizing different domains in w and u by their inverse standard deviations. This involves the variance; for the example of u this yields:

$$\begin{aligned} \sigma_u^2 &= \frac{1}{n-1} \cdot \sum_{k=1}^n (u_k - \mu_u)^2 = \frac{1}{n-1} \cdot \sum_{k=1}^n \left(u_k - \frac{1}{n} \cdot \sum_{l=1}^n u_l \right)^2 \\ &= \frac{1}{2 \cdot n \cdot (n-1)} \cdot \sum_{k=1}^n \sum_{l=1}^n (u_k - u_l)^2 \end{aligned} \tag{8}$$

Thus, individual attribute differences in Eq. 7 are scaled by their corresponding overall quadratic means $\sqrt{\sigma_u^2}$ and $\sqrt{\sigma_w^2}$, excluding zero contributions from $l = k$.

Gradients of soft Kendall τ

The sigmoid-based formulation allows for gradient-based optimization of Kendall τ, and the derivative of Eq. 5 gets

$$\frac{\partial \hat{\mathrm{r}}_{\tau,\kappa}(w,u)}{\partial u_k} = -\frac{2}{n \cdot (n-1)} \cdot \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \mathsf{sgd}_\kappa\left(\frac{w_i - w_j}{\sigma_w} \cdot \frac{u_i - u_j}{\sigma_u}\right)}{\partial u_k}\,. \tag{9}$$

This equation can be used as part of a gradient ascend scheme to adapt a vector u so as to maximize its soft Kendall correlation with a fixed vector w.

The derivative summands in Eq. 9 are transformed into

$$\begin{aligned} &\frac{\partial}{\partial u_k} \mathsf{sgd}_\kappa \left(\frac{w_i - w_j}{\sigma_w} \cdot \frac{u_i - u_j}{\sigma_u} \right) = \frac{\partial}{\partial u_k} \mathsf{sgd}_\kappa \left(z(u_i, u_j) \right) \\ &= \frac{\partial \mathsf{sgd}_\kappa(z)}{\partial z} \cdot \frac{\partial z(u_i,u_j)}{\partial u_k} \quad \text{with} \\ &z(u_i,u_j) = \tilde{w}_{ij} \cdot \frac{u_i - u_j}{\sigma_u} \quad \text{and} \quad \tilde{w}_{ij} = \frac{w_i - w_j}{\sigma_w}\,. \end{aligned} \tag{10}$$

The sigmoid possesses the simple derivative

$$\mathrm{sgd}_\kappa(z)/\partial z = \kappa \cdot \mathrm{sgd}_\kappa(z) \cdot (\mathrm{sgd}_\kappa(z) - 1) \,. \tag{11}$$

Using the derivative of the standard deviation,

$$\frac{\partial \sigma_u}{\partial u_k} = \frac{u_k - \mu_u}{(n-1)\cdot \sigma_u} \,, \tag{12}$$

three cases must be distinguished for the derivative of z:

$$\begin{aligned}
\frac{\partial z(u_k,u_l)}{\partial u_k} &= \left(\frac{1}{\sigma_u} - \frac{u_k - u_l}{\sigma_u^2}\cdot\frac{\partial \sigma_u}{\partial u_k}\right)\cdot \tilde{w}_{kl} \,, \\
\frac{\partial z(u_l,u_k)}{\partial u_k} &= \frac{\partial z(u_k,u_l)}{\partial u_k} \quad \big(z(u_l,u_k) = z(u_k,u_l)\big) \,, \\
\frac{\partial z(u_m,u_l)}{\partial u_k} &= -\frac{u_m - u_l}{\sigma_u^2}\cdot\frac{\partial \sigma_u}{\partial u_k}\cdot \tilde{w}_{ml} \,.
\end{aligned} \tag{13}$$

For these cases $k \neq l$, $k \neq m$ and $l \neq m$ are assumed. The standard deviation causes non-vanishing derivatives for the frequent last case in which the derived k-th attribute does not appear in the difference part of z.

2.2. Correlation-based embedding

For optimizing a point set X with ranks of Euclidean neighborhood relationships best matching the ranks of original data relationships, the dependence $u = D_{ij}^X = \mathbf{d}(X^i, X^j)$ is connected by the chain rule of differentiation with a given correlation measure:

$$\frac{\partial \mathrm{r}(w,u)}{\partial X_k^j} = \frac{\partial \mathrm{r}(w, \mathbf{d}(X^i,X^j))}{\partial \mathbf{d}(X^i,X^j)} \cdot \frac{\partial \mathbf{d}(X^i,X^j)}{\partial X_k^j} \,. \tag{14}$$

The first factor is taken from Eq. 9, and second one being the derivative of the Euclidean distance in Eq. 2 is just

$$\frac{\partial \mathbf{d}(X^i,X^j)}{\partial X_k^j} = -\frac{X_k^i - X_k^j}{\mathbf{d}(X^i,X^j)} \,. \tag{15}$$

These terms can be used for substituting back $w = -S_i$ and $u = D_i^X$ for computing the gradient of Eq. 1 as

$$\frac{\partial \bar{\mathrm{r}}}{\partial X_k^j} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathrm{r}(-S_i, D_i^X)}{\partial D_i^X} \cdot \frac{\partial D_i^X}{\partial X_k^j} \,. \tag{16}$$

In practice, the gradient used in optimization can be calculated as product of Jacobian matrices constructed from the partial derivatives. As illustrated in the provided code implementation of the derivatives, the Jacobians reveal sparsity and symmetry structure that can be effectively exploited for boiling down the gradient calculation to the order of $\mathcal{O}(n^3)$ that is also required for evaluating the average soft Kendall correlation in Eq. 1. The matrix-based implementation can be easily transferred to graphics processing units for enabling gradient ascend on the correlation-based embedding using memory-limited quasi-Newton l-BFGS gradient optimization that usually provides good convergence [NW99].

3. Experiments

Generally, the soft-rank embedding approach behaves much like non-metric MDS based on isotonic regression, that is, Euclidean neighborhood relationships can be reliably recovered. In addition, it offers the flexibility to directly operate on asymmetric score data. Two experiments are reported for demonstrating these novel features. The first contains only few data points for conveying the ideas of asymmetric relationships and soft neighborhood correlation. The second refers to a complex protein data base.

3.1. Asymmetric amino acid transition matrix

Asymmetric score data naturally occurs in the domain of bioinformatics. For example, protein sequence alignments depend on the properties of the amino acid transitions. While common comparison tasks are carried out with the symmetric block substitution matrix BLOSOM62, specialized analyses profit from asymmetric models. Especially, the homology in transmembrane proteins was found to be faithfully captured by the SLIM family of score matrices leading to intra-membrane domains [MRR01]. Here, the asymmetric sub-matrix of five transitions shown in Figure 1 is considered for illustration.

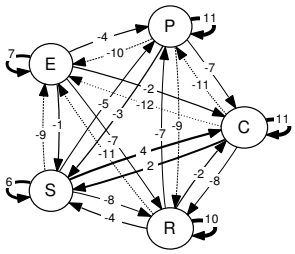

Figure 1: *Transition scores for the amino acids cysteine (C), glutamic acid (E), proline (P), arginine (R), ans serine (S) from the SLIM161 substitution matrix [MRR01].*

The embedding of the score matrix in 2D Euclidean space leads to Figure 2 with perfect neighborhood preservation, that is $\bar{\mathrm{r}} = 1 - \varepsilon$ in Eq. 1 with $\varepsilon = 10^{-9}$ being attributed to optimizer convergence goals. Note that the embedded neighborhood distances are symmetric, while the corresponding ranks, computed in row-wise manner, are not. Color patches in the top panel of Figure 2 highlight constant plateaus in the change of crisp Kendall correlation when points are moved. The proposed softening procedure used for embedding optimization leads to the bottom color field for which non-vanishing gradients and a good correspondence to the crisp calculation can be stated.

For comparison with a standard non-metric multidimensional scaling approach (mdscale), scores require conversion into symmetric dissimilarities. A common conversion

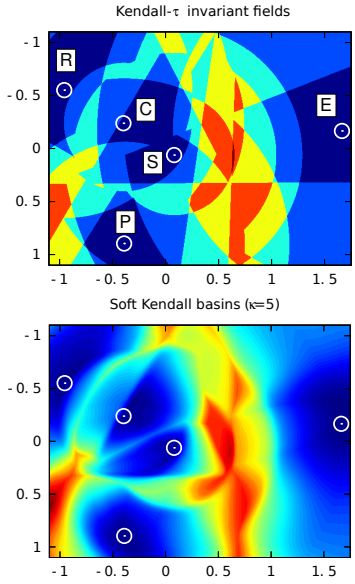

Figure 2: *Embedded sub-matrix of SLIM161 transitions scores and neighborhood order correlation domains. Circles denote 2D locations of amino acid coordinates from soft Kendall correlation embedding. Blue areas indicate regions with little change of Kendall correlation neighborhood structure for varied points, assuming other points to be fixed. Top: crisp Kendall correlation; flat color patches denote* r_τ*-invariant regions where gradient-based point placement methods fail. Bottom: soft Kendall for* $\kappa = 5$*; pronounced basins allow for optimization of the scatter point neighborhood configuration.*

is $D_{ij} = \sqrt{S_{ii} + S_{jj} - S_{ij} - S_{ji}}$ [PD05]. Using mdscale, near-perfect 2D embedding results are obtained for this problem. Not surprisingly, a comparison of this solution to the original asymmetric problem yields an average Kendall neighborhood correlation of only $\bar{r} = 0.6$ in Eq. 1. This example emphasizes that proper tools for asymmetric score data are needed rather than mapping a problem to a different one that can be solved by existing methods.

3.2. SCOP protein data set

In a second example, the SCOP database of structural classification of proteins is visualized [LN04]. It contains p-values of asymmetric pairwise alignments of 4352 proteins. The matrix covers a broad spectrum of protein families with 2888 hierarchically organized unique labels.

Figure 3 shows the embedding result using the new embedding strategy. The average neighborhood reconstruction quality of $\bar{r} = 0.201$ seems to be rather low. However, the complexity of the original data relationships is very high, because many very different proteins are compared that cannot be easily embedded into the limited 2D space. Still, color

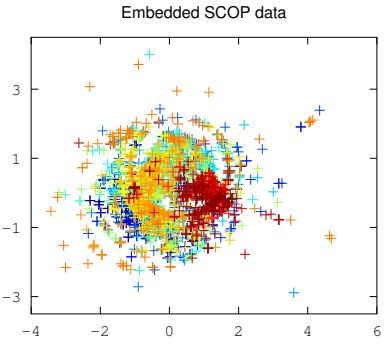

Figure 3: *Embedding of an asymmetric protein similarity matrix. 4352 points correspond to 2888 classes. Similar colors refer to nearby classes in the class hierarchy.*

patches show up when similar functional protein classes map to neighbored point coordinates.

For general performance comparison, an asymmetric version of t-distributed stochastic neighbor embedding [vdMH08, Str12] was used with an effective neighborhood size of 50. Since this parameter controls the balance between local and global neighborhood reproduction, embeddings are necessarily optimized under that constraint. Because of such focused optimization, the average overall Kendall correlation is only $\bar{r} = 0.088$, and the choice of effective neighborhood size has limited effect on that. Like in the previous example, a comparison of methods designed for the different purposes of global and focused neighborhood reconstruction is problematic. After all, one might not want to base comparisons on criteria that methods are not designed for.

4. Conclusions

A relational score embedding scheme has been introduced that maximizes (soft) Kendall correlation between potentially asymmetric object similarities in the source and embedding space. It is demonstrated that gradient-based optimization can be successfully applied for reconstruction of the neighborhood rank order. Note that unlike parametric density estimation approaches there are no embedding distortions when source and target spaces are identical. It is an important topic of future work on soft Kendall correlation to allow for emphasizing local neighborhood reconstruction, because distant source relationships should just be visually distant, while local neighborhood ordering should be better resolved.

A MATLAB/GNU-Octave package with GPU support is online available as package 'cbMDS' at https://mloss.org/.

Acknowledgment

This work was supported by the LOEWE Center for Synthetic Microbiology (SYNMIKRO), Marburg.

References

[BT12] BOLSHOY A., TATARINOVA T.: Methods of combinatorial optimization to reveal factors affecting gene length. *Bioinformatics and Biology Insights 6* (12 2012), 317–327. doi:10.4137/BBI.S10525. 1

[FC11] FRANCE S., CARROLL J.: Two-way multidimensional scaling: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41*, 5 (9 2011), 644–661. doi:10.1109/TSMCC.2010.2078502. 1

[Gow66] GOWER J. C.: Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53* (1966), 325–338. 2

[HR02] HINTON G., ROWEIS S. T.: Stochastic neighbor embedding. In *NIPS* (2002), Becker S., Thrun S., Obermayer K., (Eds.), vol. 15, MIT Press, pp. 857–864. 1

[Ken38] KENDALL M.: A new measure of rank correlation. *Biometrika 30*, 1-2 (1938), 81–93. doi:10.1093/biomet/30.1-2.81. 2

[LN04] LIAO L., NOBLE W. S.: Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Journal of Computational Biology 10*, 6 (2004), 857–868. doi:10.1089/106652703322756113. 4

[MRR01] MÜLLER T., RAHMANN S., REHMSMEIER M.: Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics 17*, suppl 1 (2001), S182–S189. URL: http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S182.short. 3

[NW99] NOCEDAL J., WRIGHT S. J.: *Numerical Optimization*. Springer, 8 1999. 3

[OLWV10] ONCLINX V., LEE J., WERTZ V., VERLEYSEN M.: Dimensionality reduction by rank preservation. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (2010), pp. 1–8. doi:10.1109/IJCNN.2010.5596347. 1

[PD05] PEKALSKA E., DUIN R. P.: *The Dissimilarity Representation For Pattern Recognition: Foundations and Applications*, vol. 64 of *Series in Machine Perception and Artificial Intelligence*. World Scientific Publishing, 2005. 4

[SB13] STRICKERT M., BUNTE K.: Soft rank neighbor embeddings. In *European Symposium on Artificial Neural Networks (ESANN)* (2013), Verleysen M., (Ed.), D-facto Publications, pp. 77–82. 2

[SSVS09] STRICKERT M., SCHLEIF F.-M., VILLMANN T., SEIFFERT U.: Unleashing Pearson correlation for faithful analysis of biomedical data. In *Similarity-Based Clustering – Recent Developments and Applications* (2009), et al. M. B., (Ed.), vol. 5400 of *LNCS*, Springer, pp. 70–91. doi:10.1007/978-3-642-01805-3_5. 1, 2

[Str12] STRICKERT M.: xSNE Stochastic Neighbor Embedding methods with novel neighborhood probabilities 1.1. Machine learning online software repository (MLOSS), 2012. URL: https://mloss.org/software/view/418/. 4

[TdSL00] TENENBAUM J., DA SILVA V., LANGFORD J.: A global geometric framework for nonlinear dimensionality reduction. *Science 290* (2000), 2319–2323. 1

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing Data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 1, 2, 4

[VPN*10] VENNA J., PELTONEN J., NYBO K., AIDOS H., KASKI S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research 11* (2010), 451–490. URL: http://dl.acm.org/citation.cfm?id=1756019. 1

[VR02] VENABLES W., RIPLEY B.: *Modern Applied Statistics with S*, 4*th* ed. Springer, 2002. 1