Nonparametric Dimensionality Reduction Quality Assessment based on Sortedness of Unrestricted Neighborhood

Davi Pereira-Santos¹, Tácito T. A. T. Neves², Andre C.P.L.F. de Carvalho¹, and F. V. Paulovich³

¹Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil
²NCEx – Federal University of Alagoas (UFAL), Arapiraca, Brazil
³Eindhoven University of Technology: Eindhoven, Noord-Brabant, NL

Abstract

High-dimensional data are known to be challenging to explore visually. Dimensionality Reduction (DR) techniques are good options for making high-dimensional data sets more interpretable and computationally tractable. An inherent question regarding their use is how much relevant information is lost during the layout generation process. In this study, we aim to provide means to quantify the quality of a DR layout according to the intuitive notion of sortedness of the data points. For such, we propose a straightforward measure with Kendall τ at its core to provide values in a standard and meaningful interval. We present sortedness and pairwise sortedness as suitable replacements over, respectively, trustworthiness and stress when assessing projection quality. The formulation, its rationale and scope, and experimental results show their strength compared to the state-of-the-art.

CCS Concepts

• Human-centered computing → Visualization design and evaluation methods; Visual analytics; • Computing methodologies → Dimensionality reduction and manifold learning;

1. Introduction

Visualization techniques provide useful tools for data exploration and decision-making. If a data set has up to two (or three) attributes (or dimensions), scatter plots can be good alternatives for exploring intuitive notions, such as spatial distribution, distance, and neighborhood. However, some sort of dimensionality reduction is usually necessary for data sets with more dimensions. In the realm of multidimensional visualization strategies, Dimensionality Reduction (DR) techniques have proven their usefulness [NA19]. Despite the advances in the field, it is inevitable that the mapping process, from a high-dimensional to a visually low space, incurs in loss of information due to the high intrinsic dimensionality [FO71] of many data sets. Therefore, DR techniques are often subject to a trade-off between interpretability and reliability, raising the question of how to effectively assess the impact of the implicit loss of information from applying a DR technique.

In the current literature, different quality measures have been proposed Section 2. Despite their popularity, the existing measures have limitations, especially the adoption of arbitrary/empirical thresholds to interpret the quality of a layout and the strong dependence on parametrization. In this work, we address these problems by proposing the concept of sortedness, able to provide a value within a meaningful interval of how well-preserved is a data set structure after a mapping process. Sortedness is the level of agreement between a data set and its DR layout regarding the order of the points as measured by the Kendall τ correlation index in relation to a given reference point [Vig15]. Adopting a correlation index provides more meaningful values, i.e. from a standard interval, and consistency to the measure by making it commensurable across different tasks and domains. We also present variants.

Our experimental results show evidence that sortedness is better suited to assess DR layout quality in general, and also in some specific scenarios, than the current practice of adopting metric/non-metric Kruskal stress, trustworthiness, and other related measures [KNO03, Krn64b, Krn64a]. Additionally, sortedness is able to indirectly assess the preservation of features usually relied upon by the human interpretation bias during data visualization, and provide a measure for machine learning pipelines.

2. Related Work

In this section, we present DR techniques as they are a frequently used type of data transformation where quality assessment is of interest, and relevant evaluation measures from the literature.

Dimensionality Reduction techniques aim to find a low-dimensional representation of high-dimensional data [EMK*21, NA19]. Existing methods map data points into graphical elements to preserve pairwise distances or neighborhoods. We can classify the techniques as global or local according to the type of distance intended to be preserved, respectively: all points; and, small neighborhoods. Both aspects are considered in our proposal.
Global techniques fail to preserve neighborhood relationships, especially when sparse high-dimensional spaces are considered [PNML08]. Kruskal [Kru64a] presented the Multidimensional Scaling (MDS) technique to map points from a high-dimensional space to a low-dimensional space by optimization. It minimizes the quadratic difference between the dissimilarities established in the original space and the calculated distances in the transformed space. This quantity is known as stress, adopted as the reference in this study. Least Squares Projection (LSP) [PNML08] introduces a different bias into the data transformation which first builds a neighborhood graph between the points. Next, it selects a subset of points to project. Remaining points are handled through interpolation by solving a system of linear equations. Many other global techniques exist. Each approach introduces its specific bias into the data transformation.

Local techniques are intended to preserve the neighborhood relationships. They help to identify groups and define their boundaries, especially for high-dimensional data sets [FFDP15]. The Piecewise Laplacian-based Projection (PLP) [PEP11] addresses the problem of sensitivity to the positioning of control points, found in techniques such as LSP, by splitting the data set into smaller subsets. The Local Affine Multidimensional Projection (LAMP) [JCC+11] allows a user-controlled redefinition of the mapping matrix based on a first mapping of control points. This kind of application would benefit directly from our proposed intuitive measure to guide the user through the interactive process. Conversely, some techniques have probabilistic bias. An example is Stochastic Neighbor Embedding (SNE) [HR02]. Each mapped point is positioned next to a group of its original neighbors with a given size. Probability distributions representing the chance of each point choosing another as a neighbor are defined, with higher probabilities assigned to closer points. Finally, local approaches have an inherent type of bias which affects the quality of the DR layout in a very different way than global techniques. Our proposed measures address such diversity by considering the unrestricted neighborhood ordering which encompasses many notions that are intuitive to the human bias, differently from the stress measure which was created having MDS in mind, and from measures limited to a certain number of neighbors.

Evaluation measures help to assess the quality of maps created by DR techniques. Most of them aggregate quality measures based on distance, neighborhood, or cluster segregation measures. However, the values returned by them are not as interpretable as those provided by our proposed measure, sortedness. A popular group of measures proposed to quantify the preservation of distances after the DR process is the set of stress functions. One of the most known is the Kruskal [Kru64a] stress function, represented in this text by \( \sigma_1 \), which measures the difference between the distances calculated in the original space and those calculated in the projected space. The stress value ranges from 0 to 1, when normalized. The smaller the value, the higher the quality. When the computed dissimilarities are not metric, e.g., based on ranking positions, the non-metric Kruskal [Kru64a] stress function, represented by \( \sigma_n \), can be used. Unreasonably good stress values can be observed for very disordered DR results [PNML08]. Furthermore, similar stress functions can lead to different perceptions of quality. When the target characteristic to be preserved is related to the neighborhood, one can adopt the Neighborhood Preservation (NP) [PM08]. This measure evaluates how many nearest neighbors established in the original space remain as nearest neighbors in the projected space. Its range is also in the interval \([0, 1]\), with values closer to 1 representing better neighborhood preservation. While this measure is interpretable, it lacks the standard meaning of our proposed interval based on a correlation index. Other measures are dependent on a parameter. Trustworthiness [KNO∗03], represented in this text by \( T_k \), depends on the number of neighbors \( k \). It measures the precision of the low-dimensional neighborhoods regarding false positives. Continuity [KNO∗03] measures the recall of the low-dimensional neighborhoods regarding false negatives. Overall, it is very similar to \( T_k \).

3. Proposed Method

In this section, we introduce the concept of sortedness and its respective numeric representation. Sortedness means how well-preserved is the data set structure after a transformation regarding the order of the points. The value is calculated by a function that returns the level of agreement between two sets of points as measured by the Kendall \( \tau \) correlation index in relation to a reference point [Ken38]. This is the most commonly used statistics to understand the correlation between two different scores for the same set of items [Vig15]. We present in the next subsections the function to evaluate the local sortedness (i.e., for a given point), which can potentially replace trustworthiness, and its reciprocal version, which considers neighborhood in the same perspective as adopted in the hubness concept [TRMI13] (Section 3.1). The measures do not depend on a strong parameter like trustworthiness and can be considered non-parametric. We also present a function to evaluate the global (pairwise) sortedness which is sensitive to distortions beyond neighborhood ordering changes (Section 3.2) while still ignoring irrelevant perturbations, which is a potential issue with the Kruskal stress formula I - see Figure 1 in Section 4. Additionally, the pairwise sortedness is generalized to accept a weighting scheme that turns it into a local measure. The other proposed local variants differ by depending on a ranking of neighbors according to their distance to a reference point. The global variant, on the other hand, depends on a ranking of all pairwise distances. The weighting scheme is based on a generalization of the Knight algorithm which has complexity \( O(n\log n) \) provided the distances are already ranked [Knu66]. We adopted the Euclidean distance in this work. All measures are provided as an open-source Python package [PSN23].

3.1. Local Neighborhood Sortedness

Let \( \rho : \mathbb{R}^d \to \mathbb{R}^2 \) be a transformation function (e.g., a dimensionality reduction), and \( X \subset \mathbb{R}^d \) a \( d \)-dimensional data set. The sortedness \( \lambda : \mathbb{R}^d \to [-1;1] \) of a given point \( x \in X \) projected as \( \hat{x} \) onto a resulting set \( \hat{X} \subset \mathbb{R}^2 \) by \( \rho \) is defined by Equation (1).

\[
\lambda_r(x) = \tau_w(r_X(x), r_{\hat{X}}(\hat{x}))
\] (1)

where \( \tau_w(a, b) \) is the weighted Kendall \( \tau \) correlation index between rankings \( a \) and \( b \) [Vig15]. Such variant of the Kendall \( \tau \) index is weighted by the function \( w(i) = (i + 1)^{-1} \) defined for each rank \( 0 \leq i < |X| \); and, \( r_X(x) \) is the ranking of all other points in \( X \) according to...
Change in neighborhood order
Change in reciprocal neighborhood order
Change in pairwise distance relations
Change in scale

Change in reciprocal neighborhood order
Change in neighborhood order

Change in pairwise distance relations
Change that keeps neighborhood and order
Rigid motion (rotation, translation, reflection)
Change in scale

Figure 1: Overview of measures: each measure is sensitive to narrower subsets of transformations/DR artifacts. $\sigma_1$ and $\lambda_\tau_\omega$ are sensitive to strong changes in the data, e.g., neighborhood order: $\sigma_1$ is sensitive to small, including irrelevant, changes.

3.3. Interpretation

All three functions benefit from the meaningful values provided by the Kendall $\tau$ rank correlation index. Notable values are 1.0, 0.5, 0.0, and -1.0, meaning respectively: perfect correspondence; half of the sortedness is preserved; meaningless transformation (equivalent to random projection); and, worst possible transformation (total unsortedness) which should only happen if it is actively/accidentally engineered to provide the exact opposite of the expected result. These meaningful values contrast with the current practice, which relies upon arbitrary values, e.g., adopting $\sigma_4 > 0.200$ as an indication of a low-quality DR, among other arbitrary values for increasing quality (0.100, 0.050, 0.025, and 0.000) [Kru64a]. Therefore, our measures are superior from the interpretability perspective. Each measure is sensitive to a different set of distortion types. Relevant types for this text are illustrated by Figure 2. The distortion degree varies from minor disturbances to major changes in the ordering. The more distorted the DR layout, the more misleading it will be to human interpretation. Rigid motion changes the point of view keeping the data topology.

4. Experiments

The following subsections present an experimental comparison of the proposed and literature measures to illustrate their properties. We applied global and local changes to a set of randomly generated two-dimensional points to simulate projection artifacts. The set is intentionally small to accentuate the effect of such changes. Therefore, both input and output data sets are two-dimensional, allowing to isolate the effects of interest from possible projection biases.

4.1. Subset Randomization

The lesser the correspondence between the data set and the DR layout, the lower its quality for visual interpretation. To simulate this, we uniformly sampled 1000 random points bounded by a 100x100 square. We randomized the location of an increasingly larger subset of the points to show how this progressive change in the ordering of the points affects the measures. Figure 3 shows meaningful near
zero end values for all proposed variants, while $T_3$ and $\sigma_1$ end values are 0.5 and 0.4, respectively. A more detailed comparison of the local variants is provided by boxplots in Figure 4 where each point in the plot represents the local measure for a data point. Non-global curves are calculated by averaging local values. The boxplot illustrates the stability of the measures across all points. Pairwise sortedness has better overall stability when the point of interest is moved away from its original neighbors. Here, such translated point is part of a set of uniformly sampled 25 collinear points within the interval $[0;50]$ along the $x$ axis. Each natural number in the axis corresponds to an offset in Figure 6. On one hand, it shows that $\Lambda_{\lambda_1}$ and $\lambda_2$ are the least sensitive due to its hit-or-miss nature which is limited to $k$ neighbors.

4.2. Gaussian Distortion

The application of Gaussian noise to each point with increasing amplitude is a way to highlight which measure is affected earlier or later as a result of its degree of sensitivity. We sampled 17 points bounded by a 100x100 square, and translated each point in a fixed random direction. The first affected measure is $\Lambda_2$, followed by the pairwise variants ($\Lambda_{\lambda_2}$), and finally, the $\Lambda_{\lambda_1}$ variants as shown in Figure 5. The last affected measure is $T_3$, which means it was the least sensitive to Gaussian noise. This shows how sensitive is $\sigma_1$ to even the smallest change in the data while $T_3$ is the least sensitive due to its hit-or-miss nature which is limited to $k$ neighbors.

4.3. Single Point Translation

A user trying to interpret a point far from its original neighbors would inevitably draw wrong conclusions. This experiment illustrates how each measure is affected when the point of interest is moved away from its original position. Here, such translated point is part of a set of uniformly sampled 25 collinear points within the interval $[0;50]$ along the $x$ axis. Each natural number in the axis corresponds to an offset in Figure 6. On one hand, it shows that $\Lambda_{\lambda_3}$ values are penalized by less than 0.2 in the complete translation (step=50) due to the measure global nature. On the other hand, local measures are directly related to the translated point which makes
them more sensitive to its translation. Despite its focus on weight-
ing pairwise distance relationships in the neighborhood, $A_{\tau}$ has a
noticeable decrease in value from 1.0 to around 0.4. This shows the
point displacement still penalizes the measure as one would expect
from a local measure, instead of exclusively taking into account
the remaining perfect ordering of the neighbors among themselves.
The penalization is fully present in the case of the $A_{\tau}$ variants
due to their use of rankings always relative to the point of interest. This
is illustrated by the extreme case of $A_{\lambda} = -1$ in the plot. The value
$A_{\lambda} = 0$ is consistent with the translation having completed half of
the linear data set length: half of the points still keep their relative
position; the other half is ordered in the exact opposite direction of
the expected.

4.4. Global Distortion

We also investigated the effect of changing the overall data struc-
ture, while keeping most local structures preserved. This represents
changes that affect, e.g., the visual interpretation of how large struc-
tures are related inside the DR layout. We created a data set with
three non-overlapping clusters with 100 random points each. The
global distortion consisted in swapping the position of two clusters.
Figure 7 shows the effect of such change for increasing clus-
ter sizes for different measures. Notice that the curve of the global
variant ($A_{\tau}$) is the most penalized by the cluster swap, while the
local variants are mainly focused on the fact that the local neighbor-
hood is preserved for the majority of the points. On the other hand,
trustworthiness ($T_2$) is completely unable to detect changes
beyond the scope of $k$ neighbors. $A_{\tau}$ behaves similarly to $A_{\lambda}$, with
the advantage of providing a meaningful value. Interestingly, the
curve stays around 0.5, meaning half of the ordering is lost. Notice
that 0 represents a total absence of correlation, and 1 represents a
perfect correlation. This suggests that, for the global variant, clus-
tering is as important as local ordering in this experiment.

Conversely, the local variants (mean values) are penalized by a de-
crease between 0.1 and 0.15 (right half of the chart) due to the
global change. Among the local variants, $A_{\tau}$ and $A_{\lambda}$ are respec-
tively the most and the less sensitive to global changes, specially
for smaller cluster sizes.

5. Conclusion

We presented the concept of sortedness as a measure of how or-
dered a DR layout is when compared to the original data points.
The simplest variant quantifies the correlation between the orig-
inal and projected neighborhoods of a given point. When un-
weighted, sortedness provides an interval of values more meaning-
ful than non-metric stress while still presenting the same behavior,
i.e., one increases monotonically with the other. When weighted,
sortedness is a suitable non-parametric replacement for trustwor-
thiness. In practice, we consider it non-parametric as the weighting
function does not need to be changed. The weighting function rec-
ommended in the literature already has the desired behavior of de-
creasing the importance of each neighbor according to its proxim-
ity [Vig15]. Additionally, we presented the reciprocal counterpart
of sortedness as a slightly more sensitive measure that considers
neighborhoods in the same way adopted in the concept of hubness.

Pairwise sortedness is the most sensitive variant. It quantifies the
correlation between the rankings of pairwise distances of a DR lay-
out when compared to the original data. This measure can be a re-
placement for Kruskal stress formula I. The former is less sensitive
to irrelevant changes in the points location than the latter, and, like
al all proposed variants, the interval of values provided by Kendall $\tau$
is more interpretable. When unweighted, it provides a global mea-
sure. When weighted, pairwise sortedness is a local (point-wise)
measure that can be averaged across all points if a balance between
locality and globality is desired.

6. Acknowledgments

This work was supported by Wellcome Leap 1kD Program; São
Paulo Research Foundation (FAPESP) - grant 2020/09835-1; Cana-
dian Institute for Health Research (CIHR) Canadian Research
Chairs (CRC) stipend [award number 1024586]; Canadian Founda-
tion for Innovation (CFI) John R. Evans Leaders Fund (JELF)
[grant number 38835]; Dalhousie Medical Research Fund (DMRF)
COVID-19 Research Grant [grant number 603082]; and the Cana-
dian Institute for Health Research (CIHR) Project Grant [award
number 177968].
References


© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.