Probability Index of Metric Correspondence as a measure of visualization reliability

Magdalena Wiercioch[1], Marek Śmieja[1], and Jacek Tabor[1]

Faculty of Mathematics and Computer Science, Jagiellonian University
Lojasiewicza 6, 30-348 Kraków, Poland
{magdalena.wiercioch,marek.smieja,jacek.tabor}@ii.uj.edu.pl

Abstract. This paper proposes a metric for measuring the quality of dimensionality reduction called *Probability Index of Metric Correspondence* (PIMC). PIMC quantifies how well a low-dimensional representation of high-dimensional input data reflects its original form. In other words, PIMC is an unsupervised technique which assigns a probability so that the projection of input data preserves the order of distances between every two pairs of elements. Moreover, we present an application of PIMC to alter the Treemap visualization method designed by B. Shneiderman. Introduced modification employs a greedy strategy such that the objects arrangement in plane is chosen based on the highest value of PIMC. The index was employed to assess existing visualization methods and proposed modification of Treemap on several real life datasets including a set of high dimensional chemical compounds. Experimental evaluation indicates that PIMC is a promising tool for quantifying the visualization reliability as well as can improve the performance of existing projection methods.

1 Introduction

The projection and visualization techniques of high dimensional data play a crucial role in computer graphics, machine learning and data analysis [14], [8]. Although numerous algorithms have been introduced, there is no unified methodology how to assess the obtained results. The validity is usually identified by perceptual research. In this contribution we propose the Probability Index of Metric Correspondence (PIMC) which allows for a numerical assessment of projections. As a corollary its application to modify the Treemap visualization method is presented [21].
A growing demand for efficient visualization and validation techniques is motivated by real life examples [15], [12]. In chemoinformatics the appropriate low dimensional representation of chemical compounds enables to search for drugs acting on various diseases with use of computer only (Computer Aided Drug Design - CADD) [1]. Since the most popular representation of compounds contains as many as 4860 coordinates [11], the visualization of chemical spaces is of great importance.
Proposed index allows for an unsupervised visualization, where the labeled data are not required, and relies on comparing the distances between elements in input and output spaces. From a practical standpoint, the construction of PIMC is motivated by the following observation:

Points that are close in the original space are expected to be close in "reduced" space as well.

In order to apply the aforementioned rule in practice let d_n be a distance in n-dimensional space and let d_k refer to a distance between responses in k-dimensional visualization space, where $n > k$. PIMC focuses on quantifying the probability that:

$$\text{if } \mathrm{d}_n(x,y) < \mathrm{d}_n(w,z) \text{ then } \mathrm{d}_k(x,y) < \mathrm{d}_k(w,z),$$

where x, y, w, z are dataset elements.

If the condition is satisfied for all objects, then the visualization is accurate. It means the metric structures of input data were properly preserved. In the opposite case when none of the elements comply with the rule, the projection yields an arrangements of points such that objects that were 'close' to each other are far away now and vice-versa. If the condition holds for one-half the number of objects, the visualization is not reliable at all and it is random.
We demonstrate how to modify Shneidersman's Treemap visualization algorithm [21] to maximize the visualization reliability in terms of PIMC value. Treemap is a space-filling method which maps a hierarchical structure of data on 2-dimensional space by splitting the plane into hierarchical regions. Our extension, which we call PIMC Treemap (PTM), relies on making such a division which maximizes a PIMC at each step. To achieve a simple and computationally efficient tool a greedy approach has been employed.
To demonstrate the utility of PIMC, we examine five visualization techniques: Principal Component Analysis [10], Factor Analysis [6], Independent Component Analysis [2], Treemap [21] and PIMC Treemap on several datasets including well-known UCI examples as well as life science datasets of high dimensional chemical compounds. Experimental results show that the proposed PIMC is a promising method for multivariate data visualization optimization and projection evaluation. The usefulness of PIMC in improving existing visualization algorithms is especially evident in the case of high dimensional data of chemical compounds, where PTM method outperforms standard Treemap and achieves comparable performance to the widely used standard techniques.
The paper is organized as follows. Next section gives a brief review of related visualization measures and methods. Section 3 introduces a PIMC measure while section 4 contains a description of Treemap and its proposed modification: PIMC Treemap. Experiments are included in Section 5. The conclusion is given in Section 6. Appendix A contains an alternative analytical version of introduced modification in Treemap method.

2 Related work

Generally, the task of building metrics which addresses the problem of evaluation in different fields has been widely discussed [22], [25]. Challenges connected with visualization assessment have been discussed in previous works. Sanyal et al. [20] show uncertainty comparing techniques for 1D and 2D datasets. Streit et al. provide a look at this area of research conducting analysis based on a probabilistic model [24]. It shows there is a great majority of different types of measures to consider in an evaluation.
On the other hand, many authors have proposed various multidimensional data visualization techniques [19], [5], too. The commonly used method is Principal Components Analysis (PCA) [10] which aims at exposing the covariance structure of a set of features. Another tool utilized in visualization is Self-organizing map (SOM) [9]. It is a two-layer neural network which allows to rearrange the data taking a similarity into account. Typically SOM can be viewed as a two-dimensional hexagonal grid. Mutidimensional Scalling (MDS) is a popular non-linear technique that finds a set of vectors in k dimensional space such that the matrix of Euclidean distances among them corresponds as closely as possible to some function of the input dissimilarity matrix [3]. Another approach provides Factor Analysis (FA) [13] which attempts to represent a set of observed variables in terms of a number of common factors plus a factor which is unique to each variable. In particular, it is used to reduce many variables to a more manageable number. Moreover, Independent Component Analysis (ICA) [7] is a method in which the goal is to find a linear representation of nongaussian data so that the components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction. Finally, Shneiderman considers a method for drawing a tree structures that makes maximal use of specified rectangle area [21].

3 Visualization validity measurement

In this section we introduce a measure for visualization assessment called Probability Index of Metric Correspondence (PIMC). Before presenting its definition let us first establish a notation.

We assume that X is n-dimensional input space and $\mathrm{d}_n(x, y)$ is a distance between $x, y \in X$. We consider a visualization mapping:

$$\pi_k : X \in x \to \pi_k(x) \in Y,$$

which projects input data onto k-dimensional space Y, where $k < n$. The distance between projections $\pi_k(x)$ and $\pi_k(y)$ is denoted by

$$\mathrm{d}_k(x, y) := \mathrm{d}_k(\pi_k(x), \pi_k(y)).$$

Generally, PIMC checks whether the relation of distances between every two pairs of points from the input space is preserved in the output space. Intuitively, points that are close in the original space should also be close in the visualization plane. The result is quantified as a probability of the event that the output space preserves the input distance order. The formal statement is given in the following definition:

Definition 1. *The Probability Index of Metric Correspondence (*PIMC*) of a visualization mapping π_k on X is defined by:*

$$\mathrm{PIMC}(X, \pi_k) := P(\{x, y, z, w \in X : sign(\mathrm{d}_n(x, y) - \mathrm{d}_n(w, z)) = sign(\mathrm{d}_k(x, y) - \mathrm{d}_k(w, z))\}) \tag{1}$$

where $P(\cdot)$ is a probability function and $sign(\cdot)$ denotes a sign function.

It is worth to mention that the formula (1) of PIMC can be written less formally by the notation:

$$P(\mathrm{d}_n(x, y) < \mathrm{d}_n(w, z) \implies \mathrm{d}_k(x, y) < \mathrm{d}_k(w, z)). \tag{2}$$

Clearly, it is impossible to preserve all the distances between objects from high dimensional space into low dimensional space, i.e. $d_n(x, y) = d_k(x, y)$. Therefore, PIMC focuses on preserving the order of input data which is a less restrictive condition.
PIMC assumes its maximum of 1 in case of ideal agreement when all distances are preserved. For completely not valid placement of elements located in reduced space PIMC gives value of 0 and forms the inverse distance relation. Therefore, objects that were at a large distance from one another, are now closely separated, whereas nearby points are far from each other. If PIMC is equal to 0.5, the arrangement of objects is totally random which characterizes the worst visualization methods.

Remark 1. In practice, the crucial problem is to effectively compute or approximate PIMC. If the number of dataset elements is not high the calculation can be performed by taking all pairs of tuples into account. However, when the cardinality of X is high or even infinite, the approximation has to be applied. Since PIMC is based on a probability, one can use a sample of elements of X to estimate this index. The more elements are considered, the higher accuracy of PIMC is obtained.

Table 1. Coordinates of objects in 5-D space X.

Object	x_1	x_2	x_3	x_4	x_5
A	4.8	0.5	3.1	2.6	1.3
B	3	2.2	3.9	2.8	3.5
C	1.3	5	4.1	3.8	0.6

The following examples give the intuition behind PIMC.

Example 1. Let us consider a 5-D space X with three points whose coordinates are shown in Table 1. Let consider three different transformations defined by:

Table 2. Comparison of distances in the input and output space after apllying the visualization mappings.

Transformation	$\|AC\|$	$\|BC\|$	$\|AB\|$
original distances	5.952	4.492	3.413
T_1	5.701	3.276	2.476
T_2	3.569	3.895	2.377
T_3	1.237	2.879	3.007

- $T_1 : (x_1, x_2, x_3, x_4, x_5) \to (x_1, x_2)$
- $T_2 : (x_1, x_2, x_3, x_4, x_5) \to (x_2 - x_3, x_5)$
- $T_3 : (x_1, x_2, x_3, x_4, x_5) \to (x_4, x_3 + x_5)$.

Table 2 presents original distances between objects and their projections.
After applying transformation 1, we obtain $\text{PIMC}(X, T_1) = 1$ since the metric structures of input data have been preserved, i.e. $|BC| < |AC|$, $|AB| < |AC|$ and $|AB| < |BC|$. The second mapping causes $\text{PIMC}(X, T_2) = 0.67$ ($|BC| \not< |AC|$). Nevertheless, PIMC gives value of 0 for the last transformation T_3 because none of the metric structures were preserved.

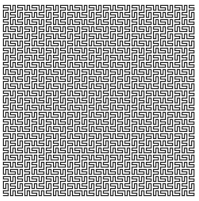


Fig. 1. Visualization of $[0, 1] \times [0, 1]$ by a line segment $[0, 1]$ defined by inverse Peano Curve transformation.

Example 2. Figure 1 presents Peano Curve [17] which continuously maps a line segment $[0, 1]$ onto a square $[0, 1] \times [0, 1]$. Let us consider a projection $\pi_1 : [0, 1] \times [0, 1] \to [0, 1]$ defined by the inverse Peano Curve transformation. We have $\text{PIMC}([0, 1] \times [0, 1], \pi_1) = 0.88$ that is close to optimal.

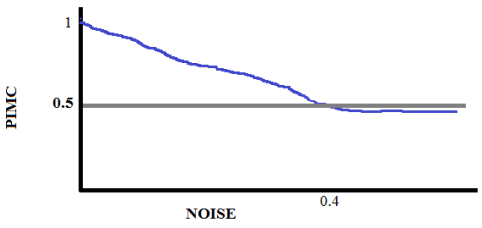


Fig. 2. The impact of noise on PIMC value.

Example 3. In order to examine the influence of the noise in the visualization mapping on the PIMC value let an identity function $\text{id} : [0, 1] \times [0, 1] \to [0, 1] \times [0, 1]$ represents a trivial projection. Clearly,

$\mathrm{PIMC}([0,1]\times[0,1],\mathrm{id}) = 1$. We modify this transformation by adding a normally distributed noise to every point of the image, i.e., let

$$\pi_2(x_1, x_2) = (x_1 + r_1, x_2 + r_2),$$

where $r_i \sim \mathcal{N}(0, \sigma)$, for $\sigma > 0, i = 1, 2$. The relation between σ and corresponding $\mathrm{PIMC}([0,1]\times[0,1],\pi_2)$ presented in the Figure 2 shows that the more noise is given, the smaller value of PIMC is. PIMC stabilizing at 0.5 which means a random visualization.

4 PIMC Treemap

B. Schneiderman proposed a visualization Treemap method which builds a hierarchy of regions. In this section we show how this method can be modified in order to maximize a visualization reliability. We start with a short description of Treemap and then present its enchanced form which we call PIMC Treemap (PTM).

Treemap assumes that an input data is represented as a tree structure. Its general idea is to map this hierarchical structure in a 2-D space Y. Each tree node corresponds to the rectangle area in Y; in particular a tree root corresponds to the entire Y. We traverse a tree in preorder visiting[1]. When visiting a tree node the corresponding rectangle region is divided into smaller rectangles according the the following rules:

- the number of partitions is determined by the number of child nodes in the tree,
- splitting direction is connected with actual tree level, i.e. horizontally at odd levels and vertically at even,
- the size of constructed regions is proportional to the number of elements of corresponding child node.

As a result visualization space is split into rectangles representing the input data.

Recursive Treemap algorithm based on a preorder visiting and corresponding **Split** function is presented in the following pseudocodes:

Treemap

Input:
R $\subset$ Y: {visualization rectangle area}
node: {pointer to a node in a tree structure of data X}
level: {level in a tree to indicate cuts to be made vertically and horizontally}
Method:
if node is null **then**
return
end if
(R_1, R_2) = **Split**(R, node, level)
Treemap(R_1, node->left, level + 1)
Treemap(R_2, node->right, level + 1)

Split

Input:
R=$[x_1, x_2] \times [y_1, y_2] \subset$ Y: {visualization rectangle area}
node: {pointer to a node in a tree structure of data X}
level: {level in a tree to indicate cuts to be made vertically and horizontally}

[1] Note that other visiting ordering are also possible.

Method:
if level is odd **then**
{split rectangle vertically}
$\mathrm{R}_1 = [\mathrm{R}.x_1, \mathrm{R}.x_1 + \frac{\text{sizeof(node->left)}}{\text{sizeof(node)}} \cdot (\mathrm{R}.x_2 - \mathrm{R}.x_1)] \times [\mathrm{R}.y_1, \mathrm{R}.y_2]$
$\mathrm{R}_2 = [\mathrm{R}.x_1 + \frac{\text{sizeof(node->left)}}{\text{sizeof(node)}} \cdot (\mathrm{R}.x_2 - \mathrm{R}.x_1), \mathrm{R}.x_2] \times [\mathrm{R}.y_1, \mathrm{R}.y_2]$
else
{split rectangle horizontally}
$\mathrm{R}_1 = [\mathrm{R}.x_1, \mathrm{R}.x_2] \times [\mathrm{R}.y_1, \mathrm{R}.y_1 + \frac{\text{sizeof(node->left)}}{\text{sizeof(node)}} \cdot (\mathrm{R}.y_2 - \mathrm{R}.y_1)]$
$\mathrm{R}_2 = [\mathrm{R}.x_1, \mathrm{R}.x_2] \times [\mathrm{R}.y_1 + \frac{\text{sizeof(node->left)}}{\text{sizeof(node)}} \cdot (\mathrm{R}.y_2 - \mathrm{R}.y_1), \mathrm{R}.y_2]$
end if
return $(\mathrm{R}_1, \mathrm{R}_2)$

Proposed extension of Treemap introduces a modification in **Split** function. Unlike Shneiderman's method, our algorithm (PIMC Treemap - PTM) aims at maximizing the value of PIMC by selecting the best variant of rectangle area division. The decision of whether to divide rectangle horizontally or vertically is nondeterministic and depends on value of PIMC calculated for each (four in total) versions of objects arrangement. The division which gives the highest PIMC value is selected. The procedure runs recursively by dividing the rectangle area proportionally to the number of elements included in child nodes.
The main difficulty of the above modification is that the exact positions of elements in the visualization space is not known when a given node is processed. More precisely, since a tree is traversed recursively from the top to the bottom, the locations of elements represented by the leaves are not determined when a splitting criterion is calculated for a given node. In consequence, an approximation form of PIMC has to be used in order to select an optimal variant of a split. Our reasoning is based on the fact that the resulting arrangement of points in a visualization plane is expected to be close to the uniform. Therefore, we assume that the elements are equally distributed over the area of associated rectangle. By such assumption, an object position is picked randomly inside the rectangle that contains it. Moreover, note that for big data, it takes too much time to verify all dataset elements. For that reason, we check the condition from Definition 1 for a fixed number of quadruples. An alternative analytical approach for a calculation of split criterion (which avoids sampling) is studied in Appendix A where Central Limit Theorem is employed.

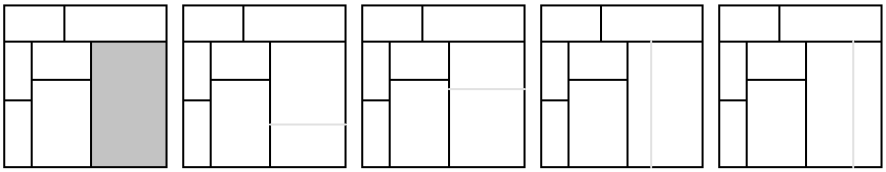


Fig. 3. Process of splitting. After a few iterations the rectangle marked with grey is going to be split 3(a) into one of the four splitting configurations 3(b), 3(c), 3(d), 3(e).

Figure 3 shows the example of rectangle partition process. Let us say the rectangle was supposed to be split in the ratio 1:2 (see Figure 3(a)). Thus, we consider four variants of its partition: two vertically aligned (see Figures 3(d) and 3(e)) and two horizontally aligned (see Figures 3(b) and 3(c)). The decision which option is best in current situation depends on the value of PIMC.
Our modification of **Split** function is as follows:

Split

Input:
R=$[x_1, x_2] \times [y_1, y_2] \subset$ Y: {visualization rectangle area}
node: {pointer to a node in a tree structure of data X}
Method:
{(R_1, R_2), (R_3, R_4), (R_5, R_6), (R_7, R_8)} = **possibleSplitting**(R) {see Figure 3}
(P_1, P_2) = **argmax** { PIMC(X, Y with R = $R_i \cup R_{i+1}$): i = 1,3,5,7 }
return (P_1, P_2)

Both Treemap and PIMC Treemap run on data represented by a tree. If data does not have a natural tree structure one can apply a hierarchical clustering algorithm [16] to form a binary tree for a data. In this case visualization results depend strictly on the performance of clustering.

5 Experiments

We have evaluated PIMC on 24 examples retrieved from UCI repository and 7 real-life datasets of chemical compounds. We considered 3 visualization techniques mentioned in related work section: PCA, FA and ICA. Moreover, we also compared the performance between Treemap (TM) with proposed PIMC Treemap (PTM). TM and PTM methods used hierarchical clustering algorithm with complete linkage functions to obtain tree structure of data [23]. We used scikit-learn [18] implementations of ICA and FA as well as R version of PCA. The codes of TM and PTM was written in C# and is available for the publicity from `http://ww2.ii.uj.edu.pl/~wiercioc/pimc` together with datasets of chemical compounds used in the experiments.

5.1 UCI datasets

Firstly, we checked how five (previously mentioned) algorithms behave when low dimensional data is considered. Results of two dimensional projection of UCI data sets are shown in Table 3. It can be seen that PCA, FA and ICA projections provided comparable PIMC values. Furthermore, TM do not appear to be successful since index is close to 0.6 (which means almost random projection). The above experiment reveals the weaknesses of TM which can be partially overcome by application of PIMC. Proposed PTM gave significant improvement with respect to TM but was not able to provide such good results as other visualization techniques. In our opinion such a poor visualization result is mainly caused by the hierarchical clustering performed at the initial stage of the algorithm.

5.2 Chemical compounds

Previous experiment focuses on relatively low dimensional data sets. It is worth noting that dimensionality reduction is especially expected for high dimensional data. For this reason, we focused on real-world examples including various datasets of selected chemical compounds.
The compounds are usually represented by fingerprints, i.e. binary strings which value of 1/0 at given position means presence/absence of specified property. Since different properties of compounds can be taken into account, a lot of fingerprint representations were introduced. In the experiments we used eight different fingerprints which dimensions are reported in the Table 4.
In the space of chemical fingerprints various notions of distances can be applied. According to recent experimental results [23], Buser metric and complete linkage method have been applied. The spatial relationship between compounds in two-dimensional space was measured according to Euclidean metric.

Table 3. PIMC values after applying PCA, FA and ICA, TM and PTM on UCI datasets.

dataset	#instances	#attributes	PCA	FA	ICA	TM	PTM
Ecoli	336	8	0.89	0.89	**0.9**	0.59	0.67
Yeast	1484	8	**0.8**	0.79	0.79	0.56	0.6
Abalone	4177	8	**0.98**	0.96	0.97	0.55	0.63
Balance-scale	625	4	0.68	**0.7**	0.69	0.57	0.7
Ionosphere	351	34	0.72	0.74	**0.75**	0.6	0.75
Breast-cancer	286	9	0.63	0.64	**0.65**	0.59	0.68
Iris	150	4	**0.93**	**0.93**	0.91	0.65	0.85
Wine	178	13	**0.88**	0.87	**0.88**	0.58	0.67
Glass	214	10	**0.9**	0.91	0.88	0.56	0.65
Image Segmentation	2310	19	0.89	0.89	**0.9**	0.58	0.65
Haberman	306	3	**0.98**	**0.99**	0.98	0.59	0.7
Zoo	101	17	**0.95**	**0.95**	0.94	0.56	0.68
Statlog	690	14	0.92	**0.93**	0.92	0.62	0.74
Seeds	210	7	**0.99**	**0.99**	**0.99**	0.68	0.8
Parkinsons	197	23	0.98	0.97	**0.99**	0.66	0.82
Madelon	4300	500	0.89	**0.91**	0.9	0.69	0.84
Hill-Valey	606	101	**0.85**	0.84	0.84	0.66	0.81
Arcene	900	10000	**0.81**	0.79	0.8	0.58	0.68
Dorothea	1950	100000	**0.78**	**0.78**	**0.78**	0.59	0.7
Housing	506	14	0.83	0.83	**0.84**	0.58	0.75
Pima Indians Diabetes	768	8	0.85	0.86	**0.88**	0.64	0.8
Page Blocks	5473	10	0.87	**0.89**	0.87	0.63	0.79
Skin Segmentation	245057	4	0.91	0.91	**0.93**	0.63	0.85
Fertility	100	10	0.82	**0.84**	0.83	0.63	0.77
AVERAGE			0.86	0.87	0.87	0.61	0.73

In the first experiments we focused on a set of compounds acting on 5-HT$_1 A$ receptor extracted from ChEMBL database [4]. It is one of the proteins responsible for the regulation of Central Nervous System. From the results summarized in Table 4 it can be noticed that, ICA gave the highest PIMC values on average. However, PTM performs better for Klekota Roth fingerprint which is considered as the most relevant representation by chemists. Furthermore, the best accuracy for all fingerprints was obtained for Graph Only with use of our method. This indicates that PTM might be useful in analysis of multidimensional data.

Since PIMC values reported in Table 4 only provide information about reliability of applied projections, one may be curious about the arrangement of compounds after 2-D mapping. A demonstration of PCA, FA, ICA and PTM projections for Klekota Roth fingerprint is given in Figures 4(a), 4(b), 4(c), 4(d). Note that PTM (as well as TM), contrary to other methods, tries to fill the entire space with datasets elements. This is one of the reasons of lower PIMC values for space-filling projections as Treemap.

Investigated space of compounds has been manually clustered by the experts in the field into 26 chemical groups [26]. We checked the location of one of such distinguished classes called Terminal Amides in the obtained visualization spaces. According to the visual inspection, data which belongs to Terminal Amides was divided into a few subgroups and the elements withing the subgroups were highly concentrated (see Figure 4).

Finally, we have examined visualization methods on six more datasets of chemical compounds, each including active and inactive compounds of receptors described in Table 5. The PIMC performance for investigated visualizations is presented in Table 5. Observe that, for 3 out of 6 receptors PTM has achieved better scores than other methods and was very close on average to the best ICA technique.

Table 4. PIMC values after applying PCA, FA, ICA, TM and PTM on dataset including 5-HT$_{1A}$ receptors ligands.

fingerprint	#instances	#attributes	PCA	FA	ICA	TM	PTM
Klekota Roth	3696	4860	0.71	0.73	0.74	0.6	**0.75**
Estate	3696	79	0.69	**0.71**	0.7	0.55	0.61
Extended	3696	1024	0.67	**0.68**	0. 67	0.56	0.61
Fingerprinter	3696	1024	0.69	**0.7**	**0.7**	0.55	0.62
Graph Only	3696	1024	0.71	0.71	0.73	0.6	**0.77**
MACCS	3696	166	0.72	0.74	**0.76**	0.58	0.64
PubChem	3696	881	**0.73**	0.72	**0.73**	0.56	0.67
Substructure	3696	307	0.72	0.72	**0.74**	0.55	0.62
AVERAGE			0.7	0.71	0.72	0.56	0.66

Table 5. Overview of considered datasets and PIMC values after applying PCA, FA, ICA, TM and PTM on datasets of actives and inactive compounds of six biological receptors.

receptor	role	#actives	#inactives	PCA	FA	ICA	TM	PTM
M_1	modulates few of physiological functions	759	938	0.7	**0.73**	0.71	0.62	0.65
h_1	has an impact on pathophysiological conditions	635	545	0.65	0.68	0.7	0.7	**0.73**
5-HT$_7$	influences on various neurological processes, such as aggression	704	339	0.78	0.79	**0.8**	0.69	0.72
5-HT$_{2A}$	has an impact on central nervous system	1835	851	0.58	0.6	0.6	0.67	**0.69**
5-HT_6	mediates both excitatory and inhibitory neurotransmission	1490	341	0.66	0.7	**0.73**	0.59	0.64
5-HT$_{2C}$	has an impact on central nervous system	1210	926	0.69	0.73	0.7	0.76	**0.78**
AVERAGE				0.68	0.7	0.71	0.67	0.7

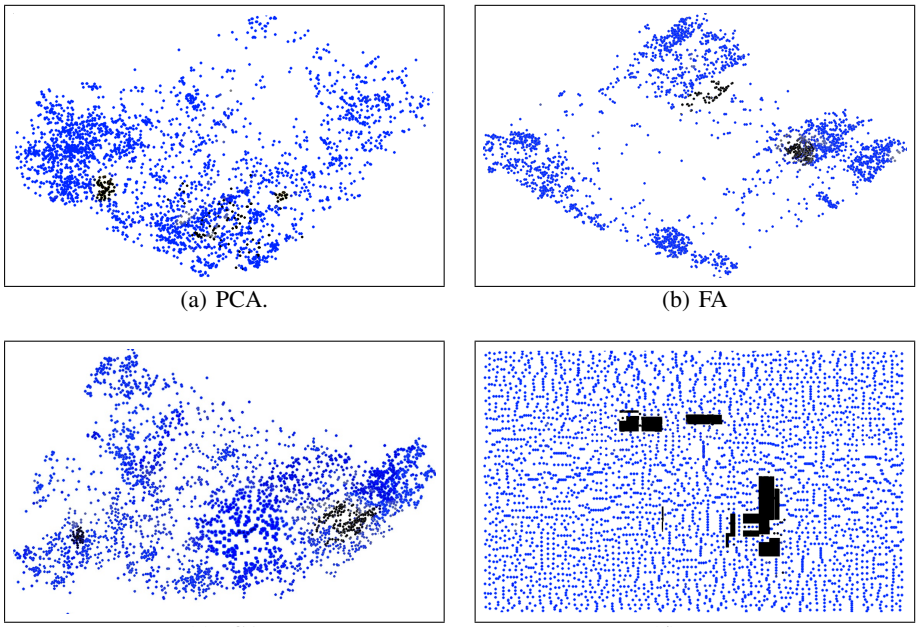
5.3 Results

The following conclusions can be withdrawn from results of our experiments:

- Visualization performed with PCA, FA and ICA gave the highest values of PIMC for UCI datasets. However, since the dimensionality of such data was low, its visualization was not the most desirable task.
- PTM provided significantly better results than the standard TM method for all datasets. It suggests that an appropriate use of PIMC might also increase the performance of other visualization techniques.
- PTM gave comparable results on average to other investigated methods when focusing on high dimensional real-life datasets of chemical compounds. In particular, it was the most accurate for 3 out of 6 receptors.

6 Conclusion

In this paper we have introduced a distance preserving measure called Probability Index of Metric Correspondence (PIMC) to evaluate visualization reliability. We have proposed a modified version of Treemap algorithm for preparation of two-dimensional representation of data. We compared the results of five 2-D projection techniques by measuring the PIMC values. A number of synthetic and real-world datasets were considered. According to experimental results, PIMC seems to be a



(a) PCA. (b) FA

(c) ICA (d) PTM

Fig. 4. The results of visualizations of chemical compounds acting on 5-HT$_{1A}$ receptor ligands. Illustrations contain also the location of Terminal amides subclass in the visualization space.

valuable measurement tool intended to quantify the visualization effect. Furthermore, one of the most important advantages of our index is that it can be used to optimize many kinds of mapping algorithms.

A Alternative calculation of PIMC splitting criterion

In section 4 we proposed a method which approximates PIMC splitting criterion by performing a sampling procedure. In this section we show an alternative approach which avoids sampling and relies on applying Central Limit Theorem (CLT).

First, we will find a distance between segments in 1-dimensional space and then use these calculation in higher dimension.

Distance between segments. Suppose that X and Y denote two independent random variables with uniform distributions on segments $I = [a, b], J = [c, d]$ respectively. We begin with computing a probability density function f of random variable $Z = (X - Y)$.

The segments I and J can be transformed to segments $[-h, h]$ and $[m - p, m + p]$, where $m > 0, h \leq p$ (for the sake of transparency we assume that $b - a < d - c$):

$$h = \frac{b-a}{2}, p = \frac{d-c}{2}, m = \frac{c+d}{2} - \frac{a+b}{2}.$$

Then:

$$f(z) = \frac{1}{h+2p} \begin{cases} 0 & \text{, for } z \leq m-h-p, \\ z-(m-h-p) & \text{, for } m-h-p < z \leq m-p, \\ h & \text{, for } m-p < z \leq m+p, \\ (m+h+p)-z & \text{, for } m+p < z \leq m+p+h, \\ 0 & \text{, for } z > m+h+p. \end{cases}$$

Consequently, a probability density function g of a random variable $Z^2 = d^2(X,Y)$ and its characteristics m_g and σ_g can be calculated.

Distance between rectangles. We now consider two pairs of independent random variables $(X_1, X_2), (Y_1, Y_2)$ with uniform distributions on rectangles $I_1 \times I_2$ and $J_1 \times J_2$ respectively. It is easy to see that, making use of the independence argument, the distance between rectangles can be computed with use of distances between particular segments as:

$$\begin{aligned} d^2((X_1,X_2),((Y_1,Y_2))) &= d^2(X_1,Y_1) + d^2(X_2,Y_2) \\ &= (X_1-Y_1)^2 + (X_2-Y_2)^2 = Z_1^2 - Z_2^2. \end{aligned}$$

PIMC. Let us finally consider a random variable:

$$R = d^2((X_1,X_2),(Y_1,Y_2)) - d^2((W_1,W_2),(Z_1,Z_2)),$$

where $(X_1, X_2), (Y_1, Y_2)$ denote the random variables with uniform distributions on $I_1 \times I_2$, $J_1 \times J_2$, while $(W_1, W_2), (Z_1, Z_2)$ are the analogical variables on $K_1 \times K_2$, $L_1 \times L_2$. Again, making use of the assumption of independence, the mean m_R and standard deviation σ_R of R can be expressed as:

$$m_R = m_{g_{(X_1,Y_1)}} + m_{g_{(X_2,Y_2)}} - m_{g_{(Z_1,W_1)}} - m_{g_{(Z_2,W_2)}},$$

$$\sigma_R = \sqrt{\sigma^2_{g_{(X_1,Y_1)}} + \sigma^2_{g_{(X_2,Y_2)}} + \sigma^2_{g_{(Z_1,W_1)}} + \sigma^2_{g_{(Z_2,W_2)}}}.$$

A number of realizations of R is very high while calculating PIMC. Therefore, given n such samples of R by CLT, we have that a random variable $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ has a normal distribution with mean m_R and standard deviation $\frac{\sigma_R}{\sqrt{n}}$. Consequently, PIMC is calculated as:

$$P(R \leq 0) = \Phi_{N(m_R, \frac{\sigma_R}{\sqrt{n}})}(0),$$

where $\Phi_{N(\cdot,\cdot)}$ is a cumulative normal distribution.
One can also derive formulas for PIMC in an arbitrary N-dimensional space.

Acknowledgement

We thank Dawid Warszycki for useful discussions about chemical aspects of fingerprints. This research was partially supported by National Centre of Science (Poland) Grants No. 2014/13/N/ST6/01832 and 2014/13/B/ST6/01792.

References

1. Awale, M., van Deursen, R., Reymond, J.L.: Mqn-mapplet: Visualization of chemical space with interactive maps of drugbank, chembl, pubchem, gdb-11, and gdb-13. Journal of Chemical Information and Modeling 53(2), 509–518 (2013)
2. Comon, P.: Independent component analysis, a new concept? Signal Process. 36(3), 287–314 (1994)
3. Cox, T.F., Cox, M.: Multidimensional Scaling, Second Edition. Chapman and Hall/CRC, 2 edn. (2000)

4. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(D1), D1100–D1107 (2012)
5. Gershon, N.D.: Visualization of an imperfect world. IEEE Computer Graphics and Applications 18(4), 43–45 (1998)
6. Hand, D.J.: Analysis of multivariate social science data, second edition by david j. bartholomew, fiona steele, irini moustaki, jane galbraith. International Statistical Review 76(3), 456–456 (2008)
7. Hyvärinen, A., Oja, E.: Independent component analysis: Algorithms and applications. Neural Netw. 13(4-5), 411–430 (2000)
8. Jackowski, K., Krawczyk, B., Wozniak, M.: Improved adaptive splitting and selection: the hybrid training method of a classifier based on a feature space partitioning. Int. J. Neural Syst. 24(3) (2014)
9. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1988)
10. Jolliffe, I.: Principal Component Analysis. Springer Verlag (1986)
11. Klekota, J., Roth, F.P.: Chemical substructures that enrich for biological activity. Bioinformatics 24(21), 2518–2525 (2008)
12. Krawczyk, Bartosz, S.J.W.M.: Data stream classification and big data analytics. Neurocomputing 150, 238–239 (2015)
13. Lawley, D.N., Maxwell, A.E.: Factor analysis as a statistical method. Journal of the Royal Statistical Society. Series D (The Statistician) 12(3), pp. 209–229 (1962)
14. Li, Y., Chen, L.: Big biological data: Challenges and opportunities. Genomics, Proteomics & Bioinformatics 12(5), 187 – 189 (2014), special Issue: Translational Omics
15. Mokbel, B., Lueks, W., Gisbrecht, A., Hammer, B.: Visualizing the quality of dimensionality reduction. Neurocomputing 112, 109–123 (2013)
16. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. Computer Journal 26(4), 354–359 (1983)
17. Peano, G.: Sur une courbe, qui remplit toute une aire plane. Mathematische Annalen 36(1), 157–160 (1890)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research 12, 2825–2830 (2011)
19. Pękalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co., Inc., River Edge, NJ, USA (2005)
20. Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., Moorhead, R.: A user study to compare four uncertainty visualization methods for 1d and 2d datasets. IEEE Transactions on Visualization and Computer Graphics 15(6), 1209–1218 (Nov 2009)
21. Shneiderman, B.: Tree visualization with tree-maps: A 2-d space-filling approach. ACM Transactions on Graphics 11, 92–99 (1991)
22. Sicilia, M.A., Rodríguez, D., García-Barriocanal, E., Sánchez-Alonso, S.: Empirical findings on ontology metrics. Expert Syst. Appl. 39(8), 6706–6711 (Jun 2012)
23. Śmieja, M., Warszycki, D., Tabor, J., Bojarski, A.J.: Asymmetric clustering index in a case study of 5-ht$_{1A}$ receptor ligands. PLoS ONE 9(7), DOI:10.1371/journal.pone.0102069, e102069 (07 2014)
24. Streit, A., Pham, B., Brown, R.: A spreadsheet approach to facilitate visualization of uncertainty in information. Visualization and Computer Graphics, IEEE Transactions on 14(1), 61–72 (Jan 2008)
25. Tang, W., Tsai, F.S., Chen, L.: Blended metrics for novel sentence mining. Expert Syst. Appl. 37(7), 5172–5177 (2010)
26. Warszycki, D., Mordalski, S., Kristiansen, K., Kafel, R., Sylte, I., Chilmonczyk, Z., Bojarski, A.J.: A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds–an application for 5-ht$_{1A}$ receptor ligands. PloS one 8(12) (2013)