# Credible Visualizations for Planar Projections

Alfred Ultsch, Michael Thrun

DataBionics Research Group
University of Marburg
Hans-Meerwein-Straße 22
D-35032 Marburg, Germany
{Ultsch,Thrun}@mathematik.uni-marburg.de

**Abstract**
Planar projections, i.e. projections from a high dimensional data space onto a two dimensional plane, are still in use to detect structures, such as clusters, in multivariate data. It can be shown that only the subclass of focusing projections such as CCA, NeRV and the ESOM are able to disentangle linear non separable data. However, even these projections are sometimes erroneous. U-matrix methods are able to visualize these errors for SOM based projections. This paper extends the U-matrix methods to other projections in form of a so called generalized U-matrix. Based on previous work, an algorithm for the construction of generalized U-matrix is introduced, that is more efficient and free of parameters which may be hard to determine. Results are presented on a difficult artificial data set and a real word multivariate data set from cancer research.

**Keywords:** Self-organizing Maps, U-matrix, Projections, Knowledge Discovery, Data Science.

## 1 Introduction

A common practice to identify structures in high dimensional data is to use projections from the high-dimensional data space into two dimensions (planar projections). Such projections cannot preserve all distances of the high dimensional space in the output space. Nevertheless, planar projections are in widespread use to show “cluster”-structures in data, see, for example [Everitt et al., 2001, pp. 31-32; Hennig et al., 2015, pp. 119-120, 683-684; Mirkin, 2005, p. 25; Ritter, 2014, p. 223]). The unavoidable errors of the projections are invisible in the typical representation of the projections’ results, usually depicted as two dimensional scatter plots. For example, the Chainlink data [Ultsch, 1995] consist of two natural clusters (two rings) that are well separated in terms of data distances and also in terms of data density. However, many popular projection methods overlay these clusters. Figure 1 shows, for example, the result of a Sammon

mapping projection [Sammon, 1969]. It can be seen, that at two locations the distinct clusters are superimposed (see Figure 1 right panel).

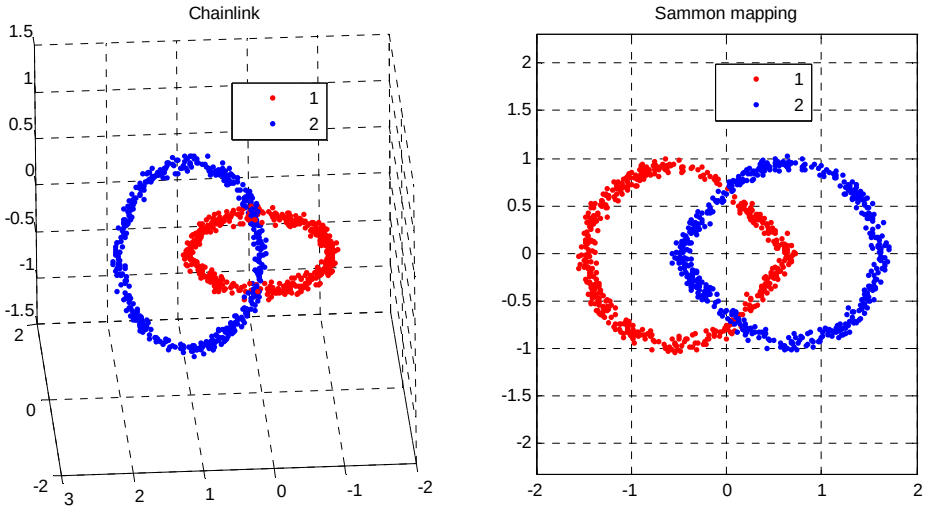


Figure1: left panel: Chainlink data, right panel: Sammon mapping of Chainlink data

The second type of error of the projections is the disruption of data points which belong to one cluster into several different clusters. This can be demonstrated using a data set from cancer research [Haferlach et al., 2010]. The data consist of microarray expression data on d = 12197 genes from n= 554 patients having four different diagnoses with respect to Leukemia: Healthy, AML, APL and CLL. In previous research it has been shown that the expression patterns of the genes correspond clearly to the four diagnostic classes [Ultsch/Lötsch, 2016]. More specifically, it was demonstrated that four clearly defined and distinguishable clusters are present in the data respectively in the distance structures of the data set. However, even modern projection methods, such as the Neighborhood Retrieval Visualizer (NeRV) [Venna et al., 2010] show cluster disruption errors. Figure 2 shows this projection. One of the clusters (APL) (green points in Figure 2 is disrupted.

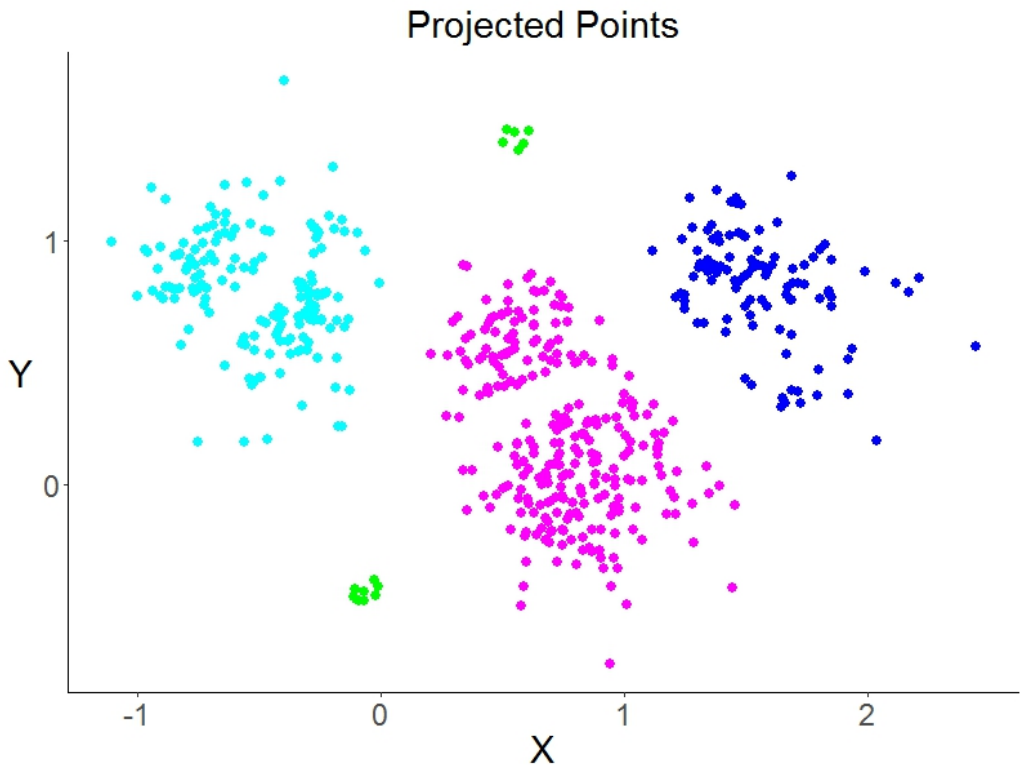


Figure 2: Neighborhood Retrieval Visualizer (NeRV) projection of the Leukemia data. The diagnosis class for APL (green) is misrepresented as two clusters.

Using the emergent SOM (ESOM) as projection method delivers a U-matrix respectively a U*-matrix [Ultsch, 2003]. These 3D visualizations on top of the output plane (gird) show in the third dimension where the output space is distorted. More specifically, if large distances are represented on a small region of the output space, these matrices show large heights. This can be used either to identify cluster structures in the data, or to show errors in the projection [Lötsch et al.]. Figure 3 shows the U* matrix of the Chainlink and the Leukemia data set. Here the cluster structures of the high dimensional data is correctly represented. However, the U-matrix methods are only defined for ESOM based projection methods.

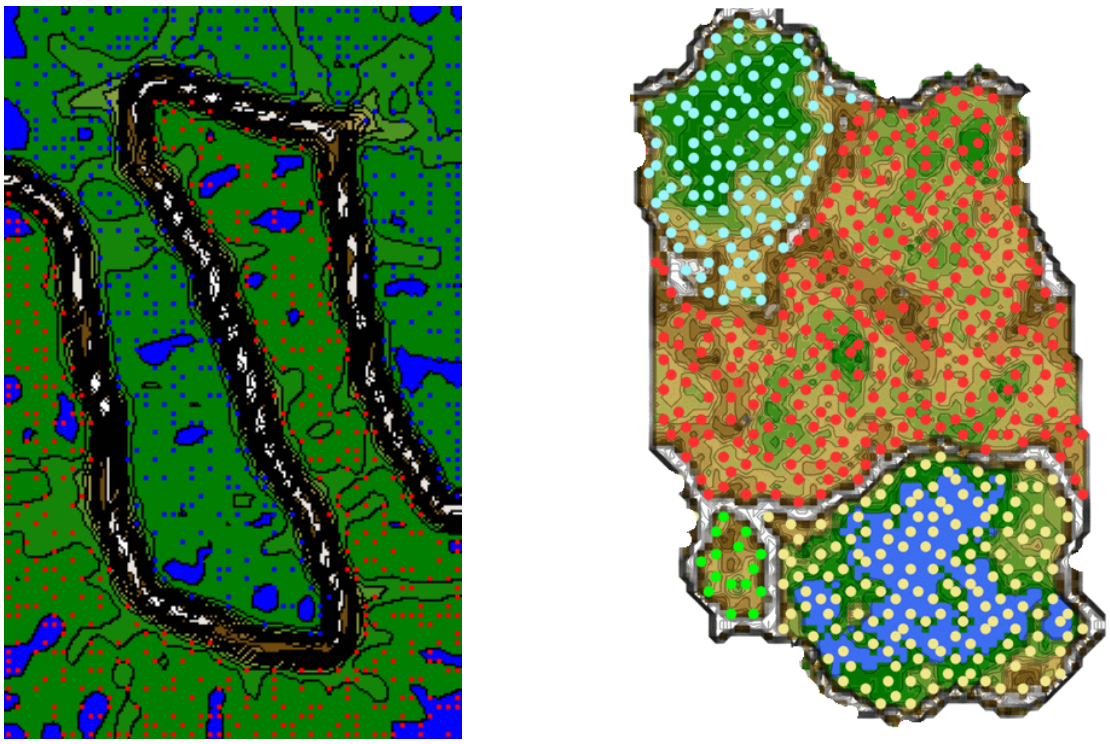

Figure 3: U*matrix of an ESOM projection of the Chainlink data (left panel) and the Leukemia data (right panel).

In this paper a generalization of these techniques to arbitrary planar projections is proposed. The following chapter reviews the common projection methods and shows their principal limits. In chapter three the generalized U-Matrix is described. In chapter four a suitable coloring scheme (hypsometric tints) is proposed. In the results chapter the method is applied to the most common focusing projection methods. The method is demonstrated on the linearly non separable data set Chainlink and the Leukemia data which contains very well defined clusters.

## 2 Projection methods

Projection methods can be categorized in general as linear and nonlinear. Linear methods perform orthonormal rotations of the data's coordinate system. Linear projections for planar projections choose the two directions, which are optimal with respect to a predefined criterion. Typical examples of these linear projections are principal component analysis PCA (variance criterion), independent component analysis ICA (non-nor-

mality criterion) and Projection Pursuit PP (user defined criterion). Since all linear projections are orthonormal rotations of the data coordinates, clusters that are linear non separable, such as the Chainlink data, cannot be separated. With this type of projections it is unavoidable that at some locations distant data are erroneously superimposed in the output space.

Nonlinear projection methods can be subdivided in global optimizing projections and focusing projections. Global optimizing projection methods use a global error measure, also known as stress- or Liapunov- function, which is optimized. The measure is based on the distances in data and output space for determining the projection. For example, Multidimensional Scaling (MDS) [Kruskal, 1964] and Sammon mapping [Sammon, 1969] use as error functions the (weighted) differences of the distances in data space and output space. These types of projections are also unable to disentangle linear non separable data set, see, for example, Figure 1.

Focusing projections start with the preservation of a global error measure which is then more and more localized (focused). Curvilinear Component Analysis (CCA) [Demartines/Hérault, 1995], stochastic Neighborhood embedding (SNE) [Hinton/Roweis, 2002], the Neighborhood Retrieval Visualizer (NeRV) [Venna et al., 2010] and ESOM/U-matrix [Ultsch, 2003] are typical examples of these methods.

The basic principle of these methods are inspired by Kohonen's SOM algorithm. During the construction of these type of projections - usually called the learning phase - the optimization criteria for the projection shifts from global optimization to local distance preservation (focusing). This is formulated using a neighborhood function. This function encompasses at the beginning of the learning phase many or all data points and focusses during the learning phase to more local points respectively distances.

This leads to non-continuous projections which are in principle capable to disentangle linearly non separable structures and should be able to preserve cluster structures which are clearly defined in data space. However, by the stochastic nature of the learning phase, and in particular the focusing process, there is no guarantee for an error free projection, even for very clearly defined clusters. The generalized U-matrix method defined below, allows at least to identify where the errors of the projections are located.

## 3 The generalized U-matrix

In this section we introduce an U-matrix technique that is generally applicable for arbitrary projection methods. It can be used to visualize both distance- and density-based structures. The basic idea was proposed in [Ultsch/Mörchen, 2006] and consist of three steps: I) discretization of the output plane to a grid structure and assignment of the projected points to units on the grid, i.e. Best Matching Units (BMUs), II) using the grid structure in a modified ESOM algorithm and III) construction of a U- respectively U*-matrix on top of the trained ESOM to show the distance respectively density structure of the output space. In step II) the principal modification compared to the classical SOM algorithm was the clamping of the positions of the projected points (BMUs) to their grid positions [Ultsch/Mörchen, 2006].The method proposed here has two advantages over the previous proposal: usage of a toroid output grids [Ultsch, 1999] and omission of epoch wise learning (sweeps through the data set) by using a special defined neighborhood function.

For the discretization of the output plane to a grid structure let ($x_i$, $y_i$) $\in$ O ( output space) denote the coordinates of the point $X_i \in R^n$ , dx denote the range of all $x_i$, dy denote the range of the $y_i$, L the number of lines of a SOM grid, C the number of columns of a SOM grid, NN the minimal number of neurons such that the SOM is able to be emergent [Ultsch, 2007]. Using $L * C \geq NN$ and the preservation of the axis relationship: $\frac{L-1}{C-1} \approx \frac{dy}{dx} = \Delta$, L (respectively C) can be determined as $L \geq -\frac{1+\Delta}{2} + \sqrt{\left(\frac{1+\Delta}{2}\right)^2 + NN * \Delta}$ [Thrun, 2017].

After the discretization of the projected points $p \in O$ to points on a discrete grid, the points, respectively their grid positions (neurons) and the corresponding input data vectors (weights) are called the best-matching units (BMUs) $bmu \in B \subset \mathbb{R}^2$ of the high-dimensional data points j, analogous to the SOM algorithm with $fgrid: O \to B, p \mapsto bmu.$ The topology of the grid is toroidal. I.e., the borders of the grid are cyclically connected [Ultsch, 1999]. Based on symmetry considerations, a simplified ESOM (sESOM) algorithm is introduced using a special neighborhood function $h: M \times M \times \mathbb{R}^+ \to [0,1]$ which is defined as $h = \begin{cases} 1 - \frac{d(j,l)^2}{\pi R^2}, \text{iff } \frac{d(j,l)^2}{\pi R^2} < 1 \\ 0, \text{else} \end{cases}$ [Thrun, 2017].

Let $M = \{m_1, \dots, m_n\}$ be a set of neurons (all grid positions) with the corresponding weights $W = \{w_1, \dots, w_n\}$ , $\subset \mathbb{R}^n$ where dim(W)=dim(I) and #W =#M, then in sESOM,

learning is achieved in each step by modifying the weights in a neighborhood as follows: $\Delta w(R) = 1 * h(bmu(j), m_i, R) * (j - w(m_i))$

In contrast to [Ultsch/Mörchen, 2006], the algorithm does not require any input parameters, the algorithm for sESOM is given in the following pseudocode:

*function (B, I)*
  *for all* $bmu(j) \epsilon$ *B:*
    *assign the positions* $m_j \in M$ *with random weightings* $w_j \epsilon$ *W on the grid*
    *assign to each* $bmu(j) = m_j$ *the weighting* $w_j = j \in I$
  *end for* $bmu(j)$
  *for R=Rmax to 1 do*
    *for all* $j \epsilon$ *I:*
    $bmu(j) = argmin_{m \in M} \{D(j, w(m))\}$
      $\Delta w(R, bmu(j)) = h(bmu(j), m_i, R) * (j - w(m_i))$
        *for all* $w(m_k) \in h(bmu(l), m_i, R)$
          $w(m_k) = w(m_k) + \Delta w(R, bmu(l))$
        *end for* $w(m_k)$
    *end for* $j \epsilon$ *I*
    *for all* $bmu(j) \epsilon$ *B:*
      *assign to each* $bmu(j) = m_j$ *the weighting* $w_j = j \in I$
  *end for R*
*end function*

This implements a stepwise iteration from the maximum radius Rmax which is given by the grid size (Rmax = C/6) stepwise with one per step and down to 1. Additionally, the search for a new best-matching unit still is used and these prototypes may change during one iteration. The predefined prototypes are reset to the weights of their corresponding high-dimensional data points after each iteration.

The calculation of the generalized U-matrix is calculated as follows: let *N(j)* be the eight immediate neighbors of $m_j \in M$ (Moore neighborhood), and let $w_j \in W$ be the prototype corresponding to $mj$; then, the average of all distances between $w_j$ and the other prototypes $w_i$ is called the U-height corresponding to the position $m_j$: $u(j) = \frac{1}{n} \sum_{i \in N(j)} D(w_i, w_j)$, $n = |N(j)|$

In general, any U-matrix approximates the Abstract U-matrix (AU-matrix). The AU-matrix has the Voronoi cells of the data points as floor and the distances between the data points as walls [Lötsch/Ultsch, 2014]. Therefore, the U-matrix on top of the sE-SOM is rescaled such that the resulting generalized U-matrix is proportional to the AU-matrix. As a consequence the neurons located on the Voronoi cell borders have exactly the distances in the data space as heights in the generalized U-matrix. In addition to the U-matrix, [Ultsch, 2003], a P-matrix allows a visualization the high-dimensional density [Ultsch]. P-heights on top of the receptive fields are displayed. The P-height $p(m_i)$ for a position $m_i$ is a measure of the density of data points in the vicinity of $w(m_j)$: $p(m_j) = |\{i \in I | D(i, w(m_j)) < r > 0, r \in \mathbb{R}\}|$. The P-height is the number of data points within a hypersphere of radius r. The radius for the density estimation can be determined using ABC analysis [Ultsch/Lötsch, 2015] and Pareto Density Estimation [Ultsch, 2005], [Thrun 2017]. In summary the P-matrix and the U*-matrix can be calculated for generalized U-matrices on arbitrary projections. The P-matrix gives an indication of the densities in data space. The U*-matrix allows a combined distance and density based visualization of the projection of the structures in the data space.

## 4 Hypsometric Tints

Hypsometric tints are surface colors that represent ranges of elevation [Patterson/Kelso, 2004]. For the visualization of U-matrix and P matrix specific color scales are combined with contour lines. The color scale is chosen to display various valleys, ridges and basins: blue colors indicate small distances (sea level), green and brown colors indicate middle distances (low hills), and white colors indicate large distances (high mountains covered with snow and ice). Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters (e.g. Fig. 8). The landscape consists of receptive fields, which correspond to certain U*-height intervals with edges delineated by contours. First, the range of U*-heights is split up into intervals, which are assigned uniformly and continuously to the color scale described above through robust normalization [Milligan/Cooper, 1988]. In the next step, the color scale is interpolated based on the corresponding CIELab color space [Colorimetry, 2004]. The largest possible contiguous areas corresponding to receptive fields in the same U*-height intervals are outlined in black to form contours. Consequently, a receptive field corresponds to one color displayed in one particular location

in the U*-matrix visualization within a height-dependent contour. Let u(j) denote the U*-heights, and let q01 and q99 denote the first and 99-th percentiles, respectively, of the U*-heights; then, the robust normalization of the U*-heights u(j) is defined by $u(j) = \frac{u(j)-q01}{q99-q01}$. The number of intervals in is defined by $\frac{1}{in} = \frac{q01}{q99}$. The resulting visualization consists of a hierarchy of areas of different height levels represented by corresponding colors. To the human eye, the visualization using the generalized U-matrix tool is a 3D landscape. One can visually interpret the presented data structures in an intuitive manner. It enables the layman to interpret projection results.

## 5 Results

The generalized U*matrix visualization is able to visualize distance based errors. A linear projection of linear non separable multivariate clusters, for example the Sammon mapping (see Fig. 4 top left) of Chainlink, cannot be cluster preserving. The same holds for nonlinear projections using a global error criterion such as for example Multidimensional Scaling (MDS). The generalized U-matrix however shows the projection errors in form of large heights surrounding single points resembling volcanos. Focusing projections are the only projection methods that are in principle able to separate linear non separable data sets correctly. Figure 4 bottom panel shows a correct projection (NeRV) of Chainlink. Figure 3 right panel shows a correct projection of the Leukemia data with ESOM projection. However, some projection methods introduce spurious clusters. See, for example CCA of Chainlink in Figure 4 top right panel. Generalized U-matrix visualizations can depict the discontinuities in high-dimensional data sets: clusters lie in valleys and are separated by hills. However, the introduction of spurious gaps between projected points (the disruption of clusters) cannot be seen using this approach.

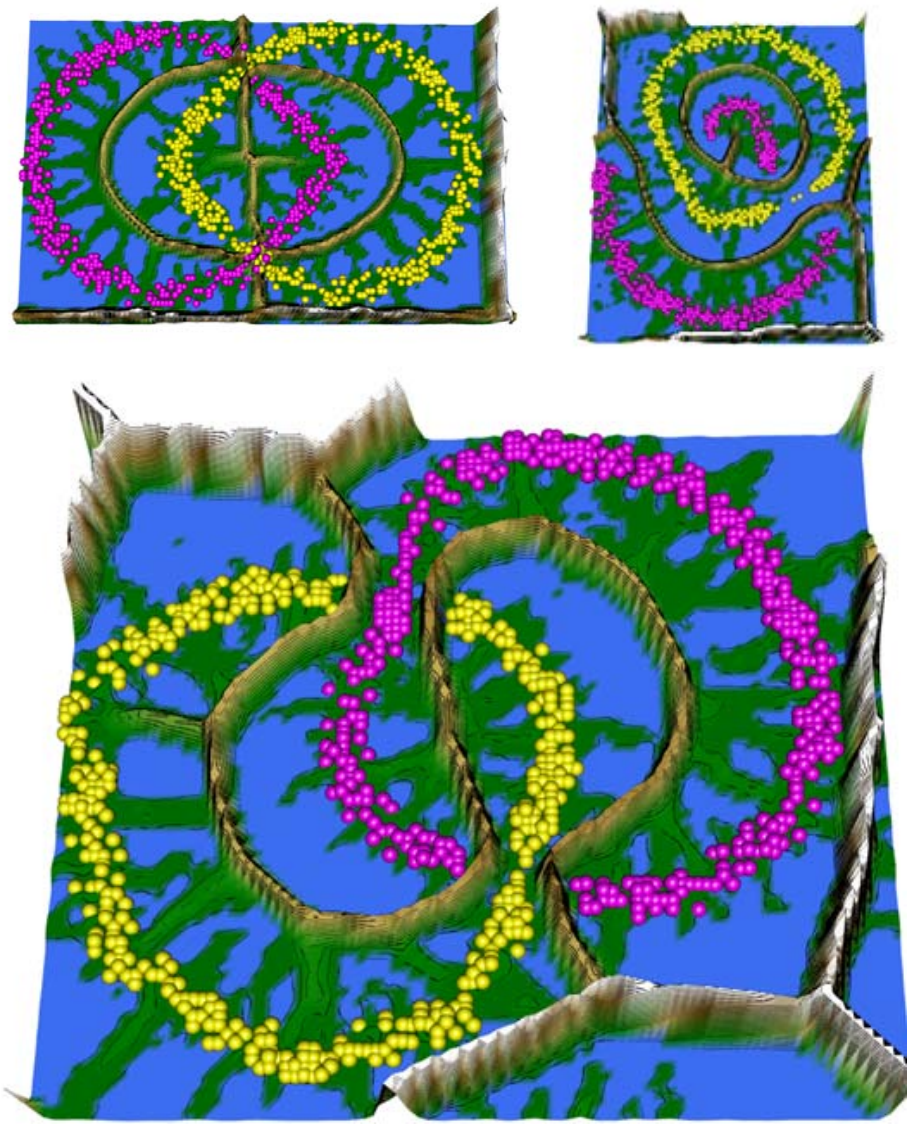

Figure 4: Generalized U-matrix for Chainlink projected with Sammons Mapping (top left), CCA (top right) and NeRV (bottom panel)

Figure 1: Leukemia Data set projected with Sammons Mapping (top left), CCA Top right), t-SNE (bottom left), NeRV (bottom right). All projections show severe error in the representation of the four distinctive clusters.

The projected points in the output space with similar high-dimensional distances lie in valley regions. If the distance-based error around a projected point is high, then the

visualization generates a mountain at this point. Figure 5 is a visualization four diagnostic classes of leukemia. All projections used, Sammons Mapping, CCA, t-SNE and NeRV, show projection errors in this data set.

## 6 Conclusion

Planar projection methods are often used to detect cluster structures in high-dimensional data. However, most of the projection methods fail to preserve even very clearly defined cluster structures. Therefore a tool to assess the quality of such projections is essential. In form of the U-matrix such tools existed for the special case of ESOM projections. This paper proposes a generalization of this methods to any type of projections. The basic idea is to discretize the output plane to a grid and using the SOM training algorithm where the BMUs are clamped to their. This idea is extended to a more efficient and parameter free algorithm. The simplified ESOM (sESOM) algorithm does not require a learning rate, nor the number of training epochs nor a cooling scheme for neighborhoods. This allows to construct the so called generalized U-matrix for arbitrary planar projections. The height information displayed on a generalized U-matrix has a definitive meaning: distances in the input space between the data points. Structure in high dimensional space cannot in general be preserved by projections onto a plane. Different projection methods make different errors. For the most interesting class of focusing projections these errors are usually unpredictable. Therefore a tool for the assessment of the quality of the projection respectively of the location of the errors is a necessity. The generalized U-matrix is able to show these errors. The method will be available in a CRAN package “generalizedUmatrix”.

**Acknowledgements**

The authors thank the Hematology, Oncology and Immunology group of Prof Andreas Neubauer of the University of Marburg and Prof. Thorsten Haferlach of Munich Leukemia Laboratory for the provision of the Leukemia Data set.

## References


[Colorimetry, 2004] **Colorimetry**.C.I.E., Vol. CIE Publication,Central Bureau of the CIE, Vienna, **2004**.

[Demartines/Hérault, 1995] **Demartines, P., & Hérault, J.**: CCA:" Curvilinear component analysis", Proc. 15° Colloque sur le traitement du signal et des images, FRA, 1995, Vol. 199, GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, **1995**.

[Everitt et al., 2001] **Everitt, B. S., Landau, S., & Leese, M.**: *Cluster analysis*, (McAllister, L. Ed. Fourth Edition ed.), London, Arnold, ISBN: 0 340 76119 9, **2001**.

[Haferlach et al., 2010] **Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Béné, M.-C., . . . Mills, K. I.**: Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group, *Journal of Clinical Oncology, Vol. 28*(15), pp. 2529-2537. **2010.**

[Hennig et al., 2015] **Hennig, C., Meila, M., Murtagh, F., & Rocci, R.**: *Handbook of cluster analysis*, New York, USA, CRC Press, ISBN: 9781466551893, **2015**.

[Hinton/Roweis, 2002] **Hinton, G. E., & Roweis, S. T.**: Stochastic neighbor embedding, Proc. Advances in neural information processing systems, pp. 833-840, **2002**.

[Kruskal, 1964] **Kruskal, J. B.**: Nonmetric multidimensional scaling: a numerical method, *Psychometrika, Vol. 29*(2), pp. 115-129. **1964.**

[Lötsch et al., 2017] **Lötsch, J., Thrun, M. C., Lerch, F., Schiffmann, S., Tegder, I., Thomas, D., . . . Ultsch, A.**: Machine-learned data structures of lipid marker serum concentrations in multiple sclerosis patients differ from those in healthy subjects, *Journal of Lipid Research, Vol. in revision*, pp., **2017.**

[Lötsch/Ultsch, 2014] **Lötsch, J., & Ultsch, A.**: Exploiting the Structures of the U-Matrix, in Villmann, T., Schleif, F.-M., Kaden, M. & Lange, M. (eds.), Proc. Advances in Self-Organizing Maps and Learning Vector Quantization, pp. 249-257, Springer International Publishing, Mittweida, Germany, **2014**.

[Milligan/Cooper, 1988] **Milligan, G. W., & Cooper, M. C.**: A study of standardization of variables in cluster analysis, *Journal of Classification, Vol. 5*(2), pp. 181-204. **1988.**

[Mirkin, 2005] **Mirkin, B.**: *Clustering: a data recovery approach*, Boca Raton, FL, USA, CRC Press, ISBN: 978-1.58488-534-4, **2005**.

[Patterson/Kelso, 2004] **Patterson, T., & Kelso, N. V.**: Hal Shelton revisited: Designing and producing natural-color maps with satellite land cover data, *Cartographic Perspectives, Vol.* (47), pp. 28-55. **2004.**

[Ritter, 2014] **Ritter, G.**: *Robust cluster analysis and variable selection*, CRC Press, ISBN: 1439857962, **2014**.

[Sammon, 1969] **Sammon, J. W.**: A nonlinear mapping for data structure analysis, *IEEE Transactions on computers, Vol. 18*(5), pp. 401-409. doi doi:10.1109/t-c.1969.222678, **1969.**

[Tasdemir/Merenyi, 2009] **Tasdemir, K., & Merenyi, E.**: Exploiting Data Topology in Visualization and Clustering of Self-Organizing Maps, *IEEE Transactions on Neural Networks, Vol. 20*(4), pp. 549-562. doi 10.1109/tnn.2008.2005409, **2009.**

[Thrun, 2017] **Thrun, M. C.**:*A System for Projection Based Clustering through Self-Organization and Swarm Intelligence,* (Doctoral dissertation), Philipps-Universität Marburg, Marburg**, 2017**.

[Ultsch, 1995] **Ultsch, A.**: Self organizing neural networks perform different from statistical k-means clustering, Proc. Society for Information and Classification (GFKL), Vol. 1995, Basel 8th-10th March **1995**.

[Ultsch, 1999] **Ultsch, A.**: Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series, In Oja, E. & Kaski, S. (Eds.), *Kohonen maps*, (1 ed., pp. 33-46), Elsevier, **1999**.

[Ultsch, 2003] **Ultsch, A.**: *U*-matrix: a tool to visualize clusters in high dimensional data*, Fachbereich Mathematik und Informatik, ISBN, **2003**.

[Ultsch, 2005] **Ultsch, A.**: Pareto density estimation: A density estimation for knowledge discovery, In Baier, D. & Werrnecke, K. D. (Eds.), *Innovations in classification, data science, and information systems*, (Vol. 27, pp. 91-100), Berlin, Germany, Springer, **2005**.

[Ultsch, 2007] **Ultsch, A.**: Emergence in Self-Organizing Feature Maps, Proc. Workshop on Self-Organizing Maps, Vol. WSOM '07, Bielefeld, Germany, **2007**.

[Ultsch/Lötsch, 2015] **Ultsch, A., & Lötsch, J.**: Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data, *PloS one, Vol. 10*(6), pp. e0129767. doi 10.1371/journal.pone.0129767, **2015.**

[Ultsch/Lötsch, 2016] **Ultsch, A., & Lötsch, J.**: Machine-learned cluster identification in high-dimensional data, *Journal of biomedical informatics, Vol. in print*, pp., **2016.**

[Ultsch/Mörchen, 2006] **Ultsch, A., & Mörchen, F.**: U-maps: topograpic visualization techniques for projections of high dimensional data, Proc. Proc. 29th Annual Conference of the German Classification Society, Citeseer, **2006**.

[Venna et al., 2010] **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, Vol. 11*, pp. 451-490. **2010.**