Publication 2

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks - ICANN 2001*, Vienna, Austria, August 21–25, pp. 485–491. Springer, Berlin, 2001.

Neighborhood preservation in nonlinear projection methods: An experimental study

Jarkko Venna and Samuel Kaski

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
{jarkko.venna,samuel.kaski}@hut.fi

Abstract. Several measures have been proposed for comparing nonlinear projection methods but so far no comparisons have taken into account one of their most important properties, the trustworthiness of the resulting neighborhood or proximity relationships. One of the main uses of nonlinear mapping methods is to visualize multivariate data, and in such visualizations it is crucial that the visualized proximities can be trusted upon: If two data samples are close to each other on the display they should be close-by in the original space as well. A local measure of trustworthiness is proposed and it is shown for three data sets that neighborhood relationships visualized by the Self-Organizing Map and its variant, the Generative Topographic Mapping, are more trustworthy than visualizations produced by traditional multidimensional scaling-based nonlinear projection methods.

1 Introduction

Nonlinear projection methods map a set of multivariate data samples into a lower-dimensional space, usually two-dimensional, in which the samples can be visualized. Such visualizations are useful especially in exploratory analyses: They provide overviews of the similarity relationships in high-dimensional data sets that would be hard to acquire without the visualizations.

The methods differ in what properties of the data set they try to preserve. The simplest methods, such as the principal component analysis (PCA) [3], are based on linear projection. A more complex set of traditional methods, that are based on multidimensional scaling (MDS) [10], try to preserve the pairwise distances of the data samples as well as possible. That is, the pairwise distances after the projection approximate the original distances. In a variant of nonlinear MDS, nonmetric MDS [8], only the rank order of the distances is to be preserved. Another variant, Sammon mapping [9], emphasizes the preservation of local (short) distances relative to the larger ones.

The Self-Organizing Map (SOM)[6, 7] is a different kind of a method that fits a discretized nonlinear surface to the distribution of the input data. The data samples are projected to the map display by finding the closest location on the discretized surface. In the end of the fitting procedure each discrete grid point on the surface is located in the centroid of all data projected onto it and its neighbors on the surface. The procedure implicitly defines a mapping from the input space onto the map grid that aims at keeping close-by grid points close-by in the original data space.

There exist several related algorithms and variants. In this study we will only consider the generative topographic mapping (GTM) [1], a probabilistic variant of the SOM. The GTM is a generative model of the probability distribution of the data, defined by postulating a latent space, a mapping from the latent space to the input space, and a noise model. The latent space corresponds to the SOM grid, and samples can be projected to it by finding their maximum a posterior locations.

In literature the methods have been compared using several measures. Since the methods are obviously good for different purposes it is not sensible to search for the overall best method but to use the "goodness measures" to characterize in which kinds of tasks each method is good at.

Most of the methods try to preserve the pairwise distances of the data samples [2], even though the emphasis of the methods varies. The SOM does not belong to this group, however. Several measures have been used to compare the SOM to other nonlinear projection methods and to measure the properties of the SOM, but according to our knowledge one key property has not been measured thus far: Whether the data points mapped close-by on the displays are generally close-by in the input space as well. The goal of this study is to test empirically which methods are best with respect to this key property of trustworthiness of the neighborhood relationships in the resulting displays.

2 Neighborhood Preservation

Preservation of neighborhoods or proximity relationships has often been mentioned as the key property of SOMs, and several different kinds of more or less heuristic measures of neighborhood preservation have been proposed. In this paper the main purpose is not to compare these methods or to develop yet another method. Our goal is to point out that, according to our knowledge, a key aspect related to neighborhood preservation in nonlinear projection methods has not been measured so far, and to compare the methods with respect to that aspect that we call the trustworthiness of the neighborhoods in the visualization.

For discrete data we define *neighborhoods* of data vectors as sets consisting of the k closest data vectors, for relatively small k. Topological concepts have been defined in terms of "arbitrarily small" neighborhoods but for discrete data we have to resort to finite neighborhoods. When the data are projected, the neighborhood is preserved if the set of the k nearest neighbors does not change.

Two kinds of errors are possible. Either new data may enter the neighborhood of a data vector in the projection, or some of the data vectors originally within the neighborhood may be projected further away on the 2D graphical display. The latter kinds of errors result from *discontinuities* in the mapping, and they have been measured extensively when quantifying the neighborhood preservation of SOMs (see e.g. [5]). As a result of discontinuities not all of the proximities in the original data are visible after the projections.

We argue that the former kinds of errors are even more harmful since they reduce the *trustworthiness* of the proximities or neighborhood relationships that are visible on the display after the projection: Some data points that seem to be close to each other may actually be quite dissimilar. According to our knowledge this property of the projection methods has not been measured previously; for SOMs the accuracy has been measured as "quantization errors" in the input space, but not in the output space.

If the data manifold is higher-dimensional than the display, then both kinds of errors cannot be avoided and all projection methods must make a tradeoff. In this paper we will concentrate on measuring the new property of trustworthiness. The discontinuities in the mapping will be measured by the preservation of neighborhoods to show the tradeoffs.

In principle, the errors could be measured simply as the average number of data items that enter or leave the neighborhoods in the projection. We have used slightly more informative measures: Trustworthiness of the neighborhoods is quantified by measuring how far from the original neighborhood the new data points entering a neighborhood come. The distances are measured as rank orders; similar results have been obtained with Euclidean distances as well (unpublished). Using the notation in Table 1 the measure for trustworthiness of the projected result is defined as

$$M_1(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in U_k(\mathbf{x}_i)} (r(\mathbf{x}_i, \mathbf{x}_j) - k) \; , \qquad (1)$$

where the term before the summation scales the values of the measure between zero and one[1].

Preservation of the original neighborhoods is measured by

$$M_2(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in V_k(\mathbf{x}_i)} (\hat{r}(\mathbf{x}_i, \mathbf{x}_j) - k) \; . \qquad (2)$$

3 Test Setting

Three datasets were used: UCI[2] 9-dimensional Glass identification database (Glass) having 214 data vectors; UCI 34-dimensional ionosphere database (Ionosphere) having 350 data vectors; and 20-dimensional Phonetic[3] dataset having 1962 data vectors.

The resolution and flexibility of SOM and GTM can be freely selected, and need be fixed for the comparison. The number of grid points on the map or latent

[1] For clarity we have only included the scaling for neighborhoods of size $k < N/2$

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] Included in LVQ_PAK, available at http://www.cis.hut.fi/research/software.shtml

Table 1. Symbols used in defining the error measures

$\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, N$ data vector

$C_k(\mathbf{x}_i)$ the set of those k data vectors that are closest to $\mathbf{x}_i$ in the original data space

$\hat{C}_k(\mathbf{x}_i)$ the set of those k data vectors that are closest to $\mathbf{x}_i$ after projection

$U_k(\mathbf{x}_i)$ the set of data vectors $\mathbf{x}_j$ for which $\mathbf{x}_j \in \hat{C}_k(\mathbf{x}_i) \wedge \mathbf{x}_j \notin C_k(\mathbf{x}_i)$ holds

$V_k(\mathbf{x}_i)$ the set of data vectors $\mathbf{x}_j$ for which $\mathbf{x}_j \notin \hat{C}_k(\mathbf{x}_i) \wedge \mathbf{x}_j \in C_k(\mathbf{x}_i)$ holds

$r(\mathbf{x}_i, \mathbf{x}_j), i \neq j$ the rank of $\mathbf{x}_j$ when the data vectors are ordered based on their Euclidean distance from the data vector $\mathbf{x}_i$ in the original data space

$\hat{r}(\mathbf{x}_i, \mathbf{x}_j), i \neq j$ the rank of $\mathbf{x}_j$ when the data vectors are ordered based on their distance from the data vector $\mathbf{x}_i$ after projection

space, which governs the resolution of the projection, was set about equal to the number of data points. The flexibility or stiffness of the SOM is determined by the final radius σ of the neighborhood function that governs adaptation during the learning process. The radius was selected according to the goodness measure given in [4].[4] We used the neighborhood function $h(d) = 1 - d^2/\sigma^2$, for $d <= \sigma$ and $h(d) = 0$ otherwise. The argument d refers to the distance on the map grid, the unit being the distance between neighboring map nodes.

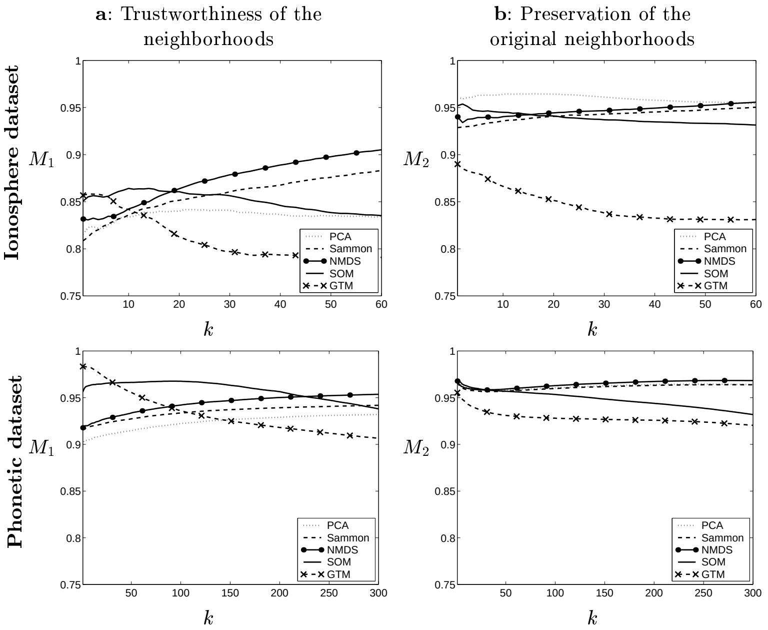
The flexibility of the GTM is governed by the number and width of the basis functions that map the latent grid to the input space. For each data set the values that gave the best log-likelihood on a validation set were chosen.

Note that since the map grids of SOM and GTM are discrete, several data points may be projected onto the same points and hence there will be ties in the rank ordering of the distances between the projected data points. To make these methods comparable to the rest, we assume that in the case of ties all rank orders are equally likely, and compute averages of the error measures.

4 Results

Trustworthiness of the visualized neighborhood relationships. When interpreting the visualizations it is particularly important that the small neighborhoods, having a small value of k, are trustworthy. They are the most salient relationships between the data items. In Fig. **1a** the trustworthiness of the projected proximities is measured for the three data sets. The SOM and GTM preserve proximities (at small values of k up to 5-10% of the size of the datasets tested) better than the other methods, whereas the MDS-based methods are mostly better at preserving the global ordering (large values of k). NMDS is the best of the traditional methods, being consistently better than Sammon mapping and PCA. The relative order of the SOM and the GTM differs on different datasets;

[4] Because the measure works better if there is more data per grid point, the number of grid points on the map was set to about 1/10 of the number data points (but to at least 50) for the selection of the stiffness. The results were scaled back to the larger map.



a: Trustworthiness of the neighborhoods

b: Preservation of the original neighborhoods

Fig. 1. Trustworthiness of the neighborhoods after projection (**a**) and preservation of the original neighborhoods (**b**) for two data sets as a function of the neighborhood size k. The results on the third data set (Glass) were similar.

the explanation is probably related to the different methods used in selecting their stiffness (cf. the discussion about stiffness and trustworthiness below).

Preservation of the original neighborhoods. While preservation of the original neighborhoods, measured in Fig. 1**b**, is not as important as the trustworthiness of the projection, it gives insight to the tradeoffs done in the projection. PCA and NMDS are the overall best methods: They perform similarly on all datasets and are consistently among the best methods. For small neighborhoods (small k) the SOM is among the best methods as well.

Flexibility of mapping increases trustworthiness of small neighborhoods. It seems plausible that there is a connection between the stiffness of the mappings and the different kinds of errors. If the mapping is stiff (the linear PCA is the extreme example), then close-by points will be projected close to each other. Unless the data lies within a low-dimensional (linear) subspace of the original space, however, data points originally far away may "collapse" into the same location.

We next investigated how the stiffness of the SOM affects the kinds of errors made.

The more flexible the SOM is (the smaller the σ) the more trustworthy the small neighborhoods are (Fig. 2**a**), as expected. There is a tradeoff in that the trustworthiness of the larger neighborhoods decreases.

The relation between the stiffness and the preservation of the original neighborhoods is not as clear, however, and requires more investigation. For larger neighborhoods it holds that the stiffer the map the better the neighborhoods are preserved (Fig. 2**b**), but for very small neighborhoods the stiffest map ($\sigma = 20$) makes most errors.

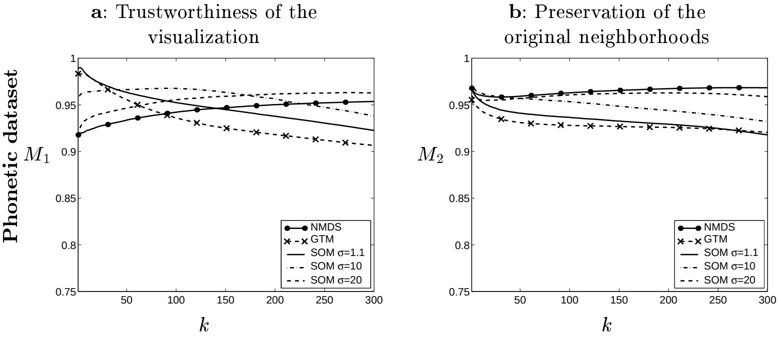


Fig. 2. The effect of the stiffness of the SOM on trustworthiness of the visualized neighborhoods (**a**) and preservation of the original neighborhoods (**b**), as a function of the neighborhood size k. The stiffness was varied by varying the final neighborhood radius σ on a SOM of a fixed size. The stiffness of the GTM shown for comparison was selected by maximum likelihood on a validation set. Data set: Phonetic.

5 Conclusions

The experimental results showed that the neighborhood relationships on the SOM and GTM displays are more trustworthy than on MDS-based displays. That is, if two data samples are close to each other on the visualizations, they are more likely to be close to each other in the original high-dimensional space as well. Moreover, the capacity of the SOM in preserving the original neighborhoods, the other aspect of neighborhood preservation, seems to be comparable to the other methods.

The trustworthiness of the local neighborhoods can still be improved by increasing the flexibility of the mapping, by reducing the radius of the neighborhood function of the SOM. The cost is that larger neighborhoods (longer distances) are not preserved as well.

References

[1] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[2] G. J. Goodhill and T. J. Sejnowski. A unifying objective function for topographic mappings. *Neural Computation*, 9:1291–1303, 1997.

[3] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520, 1933.

[4] S. Kaski and K. Lagus. Comparing self-organizing maps. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Proceedings of ICANN'96, International Conference on Neural Networks*, pages 809–814, Berlin, 1997. Springer.

[5] K. Kiviluoto. Topology preservation in self-organizing maps. *Proceedings of IEEE International Conference on Neural Networks.*, volume 1, pages 294–299, 1996.

[6] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[7] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, Berlin, 1995 (third, extended edition 2001).

[8] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrica*, 29(1):1–26, Mar 1964.

[9] J. W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, May 1969.

[10] W. S. Torgerson. Multidimensional scaling I—theory and methods. *Psychometrica*, 17:401–419, 1952.