Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 21 (2008) 160–169
Neural Networks
www.elsevier.com/locate/neunet

2008 Special Issue

On multidimensional scaling and the embedding of self-organising maps☆

Hujun Yin*

School of Electrical and Electronic Engineering, University of Manchester, Sackville Street, Manchester, M60 1QD, UK

Received 9 August 2007; received in revised form 30 November 2007; accepted 11 December 2007

Abstract

The self-organising map (SOM) and its variant, visualisation induced SOM (ViSOM), have been known to yield similar results to multidimensional scaling (MDS). However, the exact connection has not been established. In this paper, a review on the SOM and its cost function and topological measures is provided first. We then examine the exact scaling effect of the SOM and ViSOM from their objective functions. The SOM is shown to produce a qualitative, nonmetric scaling, while the local distance-preserving ViSOM produces a quantitative or metric scaling. Their relationship with the principal manifold is also discussed. The SOM-based methods not only produce topological or metric scaling but also provide a principal manifold. Furthermore a growing ViSOM is proposed to aid the adaptive embedding of highly nonlinear manifolds. Examples and comparisons with other embedding methods such as Isomap and local linear embedding are also presented.

Keywords: Self-organising maps; Multidimensional scaling; Principal curve and surface; Dimensionality reduction; Data visualisation

1. Introduction

A great challenge in the information era is to analyse a vast amount of data in order to extract useful information and to discover meaningful patterns and rules. Clustering, classification and projection of multidimensional data are common practices for this purpose in many fields, ranging from high-throughput bioinformatics, web information extraction, decision support and marketing, to data and knowledge management. Seeking a suitable and meaningful representation of a data space has always been a key objective of data analysis and pattern recognition. Projecting and abstracting data onto its underlying subspaces can help reduce the number of features, identify latent variables, detect intrinsic structures, and facilitate the visualisation of variable interactions. With the ever- increasing quantity and complexity of the data (or pattern features) used for pattern recognition tasks, more sophisticated methods are required and are being developed. A great deal of research has been devoted to this emerging topic, mainly on improving and extending the classical methods such as principal component analysis (PCA) and multidimensional scaling (MDS).

PCA has for long been considered to be the workhorse, and widely used for reducing the number of variables and visualising data in scatter plots or linear subspaces. Singular value decomposition and factor analysis are often adopted to perform the task due to various advantages, such as direct operation on the data matrix, stable results even when the data matrix is ill-conditioned, and decomposition at both the feature and data levels. The linearity of PCA however limits its power for practical, complex and increasingly large data sets, as it cannot capture nonlinear relationships defined by beyond second order statistics. Extension to nonlinear projection, in principle, can tackle the problems better; yet a unique solution is still to be defined (Malthouse, 1998). Various nonlinear methods have been proposed, such as the auto-associative networks (Kramer, 1991), generalised PCA (Karhunen & Joutsensalo, 1995), kernel PCA (Schölkopf, Smola, & Müller, 1998), principal curve and surface (Hastie & Stuetzle, 1989), and local linear embedding (LLE) (Roweis & Saul, 2000).

MDS is another popular methodology that projects data onto a low (often two) dimensional plane by preserving as closely as possible the inter-point distances (or pair-wise dissimilarities) (Cox & Cox, 1994). Metric MDS generalises classical MDS by minimising a stress function. The mapping is generally nonlinear and can reveal the overall structure of the data. Sammon mapping (Sammon, 1969) is a widely known example and uses an inter-point distance in the data space

☆ An abbreviated version of some portions of this article appeared in Yin (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

* Tel.: +44 (0) 161 306 8714; fax: +44 (0) 161 307 4784.
E-mail address: h.yin@manchester.ac.uk.

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.12.027

as the weighting for the stress. In contrast to metric MDS, nonmetric MDS finds a monotonic relationship (instead of metric one) between the dissimilarities of the data points in the data space and those of their corresponding coordinates in the low-dimensional space. A more general weighting scheme has been proposed recently and the resulting MDS is called the generalised MDS (Bronstein, Bronstein, & Kimmel, 2006). Isomap (Tenenbaum, de Silva, & Langford, 2000) applies scaling on geodesic instead of Euclidean distances. MDS techniques are generally point-to-point mappings and do not provide a generalising mapping function or manifold.

Neural networks present alternative approaches to nonlinear data projection and dimension reduction. They can provide (implicit) generalising mapping functions. Early examples include feed-forward neural network based mapping (Mao & Jain, 1995) and radial-basis-function based MDS (Lowe & Tipping, 1996). The self-organising map (SOM) (Kohonen, 1982, 1997) has become a widely used method for data visualisation. Self-organisation is a fundamental pattern recognition process, in which intrinsic inter- and intra-pattern relationships within the sensory data are learnt. Kohonen's SOM is a simplified, abstracted version of Willshaw and von der Malsburg's retinotopic mapping model (Willshaw & von der Malsburg, 1976). Modelling and analysing such mappings are important to understanding how the brain perceives, encodes, recognises, and processes the patterns it receives and thus, if somewhat indirectly, are beneficial to machine-based pattern recognition. The SOM is a topology-preserving vector quantisation. The topology-preserving property is utilised to extract and visualise relative mutual relationships among the data. Many variants and extensions have since been proposed, including the visualisation induced SOM (ViSOM) (Yin, 2002a). The ViSOM regularises the inter-neuron distances within a neighbourhood so as to preserve (local) distances on the map. The SOM and some variants have been linked with the principal curve and surface (e.g. Ritter, Martinetz, and Schulten (1992) and Yin (2002b)). It has also been widely observed that SOMs produce a similar effect to MDS. In fact it has been argued that the SOM is closer to MDS than it is to the principal curve/surface (Ripley, 1996). However, the exact connection between SOMs and MDS has not been established. Initial analysis has shown such connections (Yin, 2007). This paper further elaborates the relationship between the SOM (and ViSOM) and MDS, and to what extent they are analogous, by analysing the underlying objective functions of the mappings. The connections with principal manifolds are also analysed. A growing variant of local distance preserving ViSOM is then proposed for embedding nonlinear manifolds.

The paper is organised as follows. Section 2 provides a review on the SOM and its statistical properties. Issues surrounding its convergence and cost functions are clarified. Then the distance-preserving ViSOM is described in Section 3. In Section 4, after a general framework and definitions of MDS including metric, nonmetric and generalised MDS are given, the scaling effects of the SOM and ViSOM are examined and demonstrated. Their links to the principal curve/surface are also analysed, followed by a growing ViSOM for embedding nonlinear manifolds. Examples and experiments are presented in Section 5. Brief conclusions are given in the last section.

2. Self-organising maps: A review

2.1. The SOM algorithm

External stimuli are received by various sensory or receptive fields (e.g. visual-, auditory-, motor-, or somato-sensory), coded or abstracted by the living neural networks, propagated through axons, and projected onto the cerebral cortex, often to distinct parts of the cortex. The different areas of the cortex (cortical maps) respond to different sensory inputs, though many functions and actions require collective responses from various areas. Topographically ordered mappings are widely observed in the cortex. The main structures (primary sensory areas) of the cortical maps are established genetically in a predetermined manner (Kohonen, 1984). More detailed areas (associative areas) between the primary sensory areas, however, are developed through self-organisation gradually during life, and in a topographically meaningful fashion. Therefore, studying such topographic projections, which had been ignored during the early period of neural information processing research (Kohonen, 1986), is undoubtedly fundamental to the understanding and construction of the dimension-reduction mapping for effective representation of sensory information and feature extraction.

Von der Malsburg and Willshaw first developed in mathematical form the self-organising topographic mappings, mainly from two-dimensional presynaptic sheets to two-dimensional postsynaptic sheets, based on retinatopic mapping: the ordered projection of the visual retina to the visual cortex (von der Malsburg & Willshaw, 1973; Willshaw & von der Malsburg, 1976). Kohonen (1982) abstracted this self-organising learning model and proposed a much simplified mechanism which ingeniously incorporates the Hebbian learning rule and lateral interactions. This simplified model can emulate the self-organisation effect. Although the SOM algorithm was more or less proposed in a heuristic manner, it is an abstract and generalised model of the self-organisation or unsupervised learning process.

In the SOM, a set of neurons, often arranged in a 2-D rectangular or hexagonal grid or map, is used to form a discrete, topological mapping of an input space, $\boldsymbol{X} \in \mathbf{R}^n$. At the start of the learning, all the weights $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ are initialised to either small random numbers or some specific values (e.g. the principal subspace), where $\mathbf{w}_i$ is the weight associated to neuron i and is a vector of the same dimension, n, of the input, M is the total number of neurons. Denote by $\mathbf{r}_i$ the discrete vector defining the position (coordinates) of neuron i on the map grid. Then the algorithm iterates the following steps.

- At each time t, present an input, $\mathbf{x}(t)$, select the winner,
$$v(t) = \arg\min_{k \in \Omega} \|\mathbf{x}(t) - \mathbf{w}_k(t)\|. \quad (1)$$
- Update the weights of the winner and its neighbours,
$$\Delta\mathbf{w}_k(t) = \alpha(t)\eta(v, k, t)[\mathbf{x}(t) - \mathbf{w}_k(t)]. \quad (2)$$

- Repeat until the map converges.

Here $\alpha(t)$ is the learning rate and $\eta(v, k, t)$ is the neighbourhood function and Ω is the set of neuron indexes. Although one can use a top-hat type of neighbourhood function, a Gaussian, $\eta(v, k, t) = \mathrm{e}^{-\frac{d_{vk}^2}{2\sigma(t)^2}}$, is often used in practice with $\sigma(t)$, representing the effective range of the neighbourhood and $d_{vk} = \|\mathbf{r}_v - \mathbf{r}_k\|$, the distance between neurons v and k on the map grid.

2.2. Convergence and cost function

The SOM was proposed to model the sensory-to-cortex mapping, and thus is an unsupervised, associative memory mechanism. Such a model is also related to vector quantisation (VQ) in coding terms. The SOM has been shown to be an asymptotically optimal VQ (Yin & Allinson, 1995). More importantly, with the neighbourhood learning, the SOM is an error tolerant VQ and a Bayesian VQ (Luttrell, 1990, 1994). Convergence and ordering has only been formally proven in the trivial one- dimensional case. A complete proof of both convergence and ordering in multidimensional cases is still outstanding, though there have been several attempts that have made progress(for example, Ritter and Schulten (1988), Erwin, Obermayer, and Schulten (1992a), Erwin, Obermayer, and Schulten (1992b), Lo and Bavarian (1991), Yin and Allinson (1995) and Lin and Si (1998)). Especially, it was shown that there was no cost function that the SOM would follow *exactly* (Erwin et al., 1992a, 1992b). Such an issue is also linked to the claimed lack of an exact cost function that the algorithm optimises.

Recent work by various researchers has shed light on this intriguing issue surrounding the SOM. In Yin and Allinson (1995), the Central Limit Theorem is extended and used to show that with the diminishing neighbourhood, the weight vectors are asymptotically Gaussian distributed and will converge in the mean-square sense to the means of the Voronoi cells. The authors have also proved that the initial state has a diminishing effect on the final weights when the learning parameters meet the convergence conditions. Such an effect has been recently verified in de Bolt, Cottrell, and Verleysen (2002) using Monte-Carlo bootstrap cross-validation. The ordering was not considered. Luttrell (1990) first related a hierarchical noise tolerant coding theory to the SOM. When the transmission channel noise is taken into consideration, a two-stage optimisation has to be done, not only to minimise the representation distortion (as in the VQ), but also to minimise the distortion caused by the channel noise. He revealed that the SOM can be interpreted as such a coding algorithm. The neighbourhood function acts as the model for the channel noise distribution and should not go to zero. Such a noise tolerant VQ has the following objective function (Luttrell, 1990),

$$D_2 = \int \mathrm{d}\mathbf{x} p(\mathbf{x}) \int \mathrm{d}\mathbf{n} \pi(\mathbf{n}) \|\mathbf{x} - \mathbf{w}_k\|^2 \tag{3}$$

where $\mathbf{n}$ is the noise variable and $\pi(\mathbf{n})$ is the noise distribution. Durbin and Mitchison (1990) and Mitchison (1995) have linked the SOM and this noise tolerant VQ with minimal wiring of cortex-like maps.

When the codebook (map) is finite, the noise can be considered as discrete; the cost function can thus be re-expressed as,

$$D_2 = \sum_i \int_{V_i} \sum_k \pi(i, k) \|\mathbf{x} - \mathbf{w}_k\|^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{4}$$

where V_i is the Voronoi region of cell i. When the channel noise distribution is replaced by a neighbourhood function (analogous to inter-symbol dispersion), this gives the cost function of the SOM algorithm. The neighbourhood function can be interpreted as a channel noise model. Such a cost function has been discussed in the SOM community (e.g. Kohonen (1991), Lampinen and Oja (1992), Heskes (1999), Ripley (1996) and Yin (1996)). The cost function is, therefore,

$$E(\mathbf{w}_1, \ldots \mathbf{w}_M) = \sum_i \int_{V_i} \sum_k \eta(i, k) \|\mathbf{x} - \mathbf{w}_k\|^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{5}$$

This leads naturally to the SOM update algorithm using the sample, or stochastic, gradient descent method. That is, for each Voronoi region, the sample cost function is,

$$\hat{E}_i(\mathbf{w}_1, \ldots \mathbf{w}_M) = \int_{V_i} \sum_k \eta(i, k) \|\mathbf{x} - \mathbf{w}_k\|^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{6}$$

The optimisation for the weights $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ can be sought using the sample gradients. The sample gradient for $\mathbf{w}_j$ is,

$$\frac{\partial \hat{E}_i(\mathbf{w}_1, \ldots \mathbf{w}_M)}{\partial \mathbf{w}_j} = 2\eta(i, j)(\mathbf{x} - \mathbf{w}_j) \tag{7}$$

which leads to the SOM updating rule, Eq. (2). Note that, although the neighbourhood function $\eta(i, j)$ is only implicitly related to $\mathbf{w}_j$, it does not contribute to the weight optimisation, nor does the weight optimisation lead to its adaptation (neighbourhood adaptation is controlled by a pre-specified scheme, unrelated to the weight adaptation); thus the neighbourhood can be omitted from taking the partial differential. *This point has caused problems in interpreting the cost function of the SOM in the past.*

It has however been argued that this energy function is violated at boundaries of Voronoi cells where the input $\mathbf{x}$ has exactly the same smallest distance to two neighbouring neurons. Thus this energy function holds mainly for the discrete cases where the probability of such boundary input points is close to zero, or the local (sample) cost function $\hat{E}_i$ should be used instead in deciding the winner (Heskes, 1999). When a spatial-invariant neighbourhood function is used (as is often the case), assigning the boundary input to either cells will lead to the same local sample cost or error; therefore any input data on the boundary can be assigned to either Voronoi cells that have the same smallest distance to it (for example, using the first-come-first-served manner). Only when the neurons concerned lie on the borders of the map, does such violation occur due to unbalanced neighbourhoods of the neurons. The result is a

slightly more contraction towards to the centre (inside) of the map for the border neurons (Kohonen, 1991). Using either the simple distance or the local distortion measure as the winning rule will result in border neurons being contracted towards the inside of the map, especially when the map is not fully converged or when the effective range of the neighbourhood function is large. With the local distortion rule, this boundary effect is greater as greater local error is incurred at the border neurons due to their few neighbouring neurons than inside neurons.

To follow the cost function exactly, the winning rule should be modified to follow the local sample cost function $\hat{E}_i$ (or the local distortion measure) instead of the simplest nearest distance, that is,

$$v = \arg\min_i \sum_k \eta(i,k)\,\|\mathbf{x}-\mathbf{w}_k\|^2 . \tag{8}$$

When the neighbourhood function is symmetric (as is often the case), and when the data density function is smooth, this local distortion winning rule is almost the same as the simplest nearest distance rule for most non-boundary nodes, especially if the number of nodes is large. On the map borders, however, differences exist due to the imbalance of the nodes presented in the neighbourhood. Such differences however become negligible to the majority of the neurons, when a large map is used and when the neighbourhood function shrinks to its minimum scale.

2.3. Topological order measures

The quality of the mapping in terms of topographic or topological preservation is measured for its topological ordering, in addition to the overall quantisation error. Such a measure is not unique (unless the input and map dimensions are the same) and there is no clear definition of order (Goodhill & Sejnowski, 1997). Among several proposed measures, Bauer and Pawelzik (1992) proposed a measure called the topology product to measure the topological ordering of the map:

$$P = \frac{1}{M^2-M} \times \sum_i \sum_j \log\left(\prod_{l=1}^{j} \frac{d^D(\mathbf{w}_i,\mathbf{w}_{\zeta^O(l,i)})}{d^D(\mathbf{w}_i,\mathbf{w}_{\zeta^D(l,i)})}\frac{d^O(i,\zeta^O(l,i))}{d^O(i,\zeta^D(l,i))}\right)^{\frac{1}{2k}} \tag{9}$$

Here d^D and d^O represent the distance measures in the input (or data) space and on the output (or map) space respectively. $\zeta^D(l,i)$ and $\zeta^O(l,i)$ represent the l-th neighbour of node i in the data (D) and map(O) spaces, respectively.

The first ratio in the product measures the match of weight distance sequences within a neighbourhood (upto j) on the map and those in the data space. The second ratio is the ratio of index distance sequences within the neighbourhood on the map and those in the data space. The topographic product measures the product or the match of these two ratios for all the possible neighbourhoods.

Villmann, Der, Herrmann, and Martinetz (1997) proposed a topographic function to measure the "neighbourhood-ness" of weight vectors in data space as well as on the map grid. The neighbourhood-ness of the weight vectors is defined by the adjacent Voronoi cells of the weights. The function measures the degree to which weight vectors are ordered in the data space as to their indexes on the grid, as well as how well the indexes are preserved when their weight vectors are neighbours.

Defining a fully ordered map can be straightforward using the distance relations (Yin, 1996). For example, if all the nearest neighbouring nodes on the map grid have their nearest neighbouring nodes' weights in their nearest neighbourhood in the data space, we can call the map a 1st-order (ordered) map, that is,

$$d(\mathbf{w}_i,\mathbf{w}_j) \le d(\mathbf{w}_i,\mathbf{w}_k), \quad \forall i \in \Omega;\ j \in \Lambda_i^1;\ k \notin \Lambda_i^1 \tag{10}$$

where Ω is the set of grid indexes and Λ_i^1 denotes the 1st order neighbourhood of node i on the map grid.

Similarly, if the map is a 1st-order map, and all the 2nd order neighbouring nodes on the grid have their weights in their 2nd nearest neighbourhoods in the data space, we call the map a 2nd-order (ordered) map. For the 2nd ordered map, the distance relations to be satisfied are:

$$d(\mathbf{w}_i,\mathbf{w}_j) \le d(\mathbf{w}_i,\mathbf{w}_k) \le d(\mathbf{w}_i,\mathbf{w}_l), \quad \forall i \in \Omega;\ j \in \Lambda_i^1;\ k \notin \Lambda_i^1\ \&\ k \in \Lambda_i^2;\ l \notin \Lambda_i^1 \vee \Lambda_i^2 \tag{11}$$

and so forth to define higher order maps with inter-neuron distance hierarchies. An m-th order map is optimal for tolerating the channel noise spreading up to the m-th neighbouring code. Such fully ordered maps, however, may not be always achievable, especially when the mapping is a dimension reduction one. Then the degree (percentage) of the nodes with their weights being ordered can be measured. Together with the probabilities of the nodes being utilised, it can be used to quantify the topology preservation and to what degree and to what order the map can tolerate the channel (inter-symbol) noise. This approach can directly associate the topological order with the error/fault tolerance ability of the map.

Goodhill and Sejnowski (1997) proposed the C measure, a correlation between the similarity of stimulus in the data space and the similarity of their prototypes in the map space, to quantify the topological preservation:

$$C = \sum_i \sum_j F(i,j)G[M(i),M(j)] \tag{12}$$

Here F and G are symmetric similarity measures in the input and map spaces respectively and are problem specific, and $M(i)$ and $M(j)$ are the mapped points of nodes i and j, respectively.

The C measure evaluates the correlation between distance relations of two spaces and is a more generalised topographic measure. Many topographic mapping objectives can be unified under the C measure. It has also been shown that if a mapping that preserves ordering exists, then maximising C will find it. Thus the C measure can also be used as the objective function of the mapping, an important property different from other topology preservation measures and definitions.

The above list of measures is by no means complete. Other measures for ordering exist in the literature. One can always use the underlying cost function equation (5) to directly measure the goodness of the mapping, including the topology preservation. At least one can use a temporal window to take a sample of it (Kohonen, 1991). The (final) neighbourhood function specifies the level of topology (ordering) the mapping is likely to achieve or is required.

3. Visualisation induced SOM (ViSOM)

The SOM has been widely used for data visualisation. However, the inter-neuron distances, when referred to the data space, have to be crudely or qualitatively marked by colours or grey levels on the trained map. The coordinates of the neurons (the resulting of scaling) are fixed on the lower dimensional (often 2-D) grid and do not resemble the distances (dissimilarities) in the data space.

For metric scaling and data visualisation, a direct and faithful display of the datas structure and distribution is highly desirable. The visualisation *induced* SOM (ViSOM) has been proposed to extend the SOM for distance preservation on the map (Yin, 2002a). For the map to capture the data manifold structure directly, (local) distance quantities must be preserved on the map, along with the topology. The map can be seen as a smooth and graded mesh embedded into the data space, onto which the data points are mapped and the inter-point distances are approximately preserved.

In order to achieve that, the updating force, $[\mathbf{x}(t) - \mathbf{w}_k(t)]$, of the SOM algorithm is decomposed into two elements, $[\mathbf{x}(t) - \mathbf{w}_v(t)] + [\mathbf{w}_v(t) - \mathbf{w}_k(t)]$. The first term, $[\mathbf{x}(t) - \mathbf{w}_v(t)]$, represents the updating force from the winner v to the input $\mathbf{x}(t)$, and is similar to the updating force used by the winner v. The second term, $[\mathbf{w}_v(t) - \mathbf{w}_k(t)]$, is a lateral contraction force bringing the neighbouring neuron k to the winner. In the ViSOM, this lateral contraction force is regulated in order to help maintain unified inter-neuron distances locally on the map. The update rule is

$$\Delta\mathbf{w}_k(t) = \alpha(t)\eta(v,k,t)\left([\mathbf{x}(t) - \mathbf{w}_v(t)] + \beta[\mathbf{w}_v(t) - \mathbf{w}_k(t)]\right) \tag{13}$$

where β is a constraint coefficient — the simplest form being $\beta = \delta_{vk}/(d_{vk}\lambda) - 1$, δ_{vk} is the distance of neuron weights in the input space, d_{vk} is the distance of neuron indexes on the map, and λ is a resolution constant.

The ViSOM regularises the inter-neuron contraction so that local distances between the nodes on the map are analogous to the distances of their weights in the data space. In addition to the SOM objective of minimising the quantisation error, the aim is also to maintain constant inter-neuron distances locally. When the data points are eventually projected onto the trained map, the distance between data points i and j on the map is proportional to the distance of these two points in the data space, at least locally, subject to the quantisation error (the distance between a data point and its neural representative). That is, $d_{ij} \propto \delta_{ij}$ or $\lambda d_{ij} \approx \delta_{ij}$. This makes the visualisation more direct and quantitatively measurable. The resolution of the map can be enhanced by interpolating a (small) trained map or by incorporating the local linear projection (LLP) method (Yin, 2003). Instead of being projected onto the winning node v (or $\mathbf{w}_v$), the data point $\mathbf{x}$ is projected to the sub plane spanned by the two closest edges, as shown in Fig. 1. The projected point is, therefore,

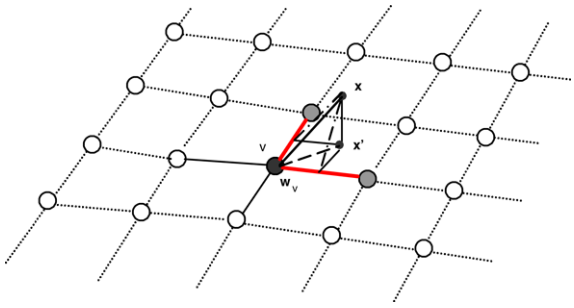


Fig. 1. Local linear projection. $\mathbf{x}$: data point, v: winning node, and $\mathbf{x}'$: the projected point on the map.

$$\mathbf{x}' = \mathbf{w}_v + \max_{v'=v\pm1}\left\{\frac{(\mathbf{x} - \mathbf{w}_v)\bullet(\mathbf{w}_v - \mathbf{w}_{v'})}{\|\mathbf{w}_v - \mathbf{w}_{v'}\|^2}, 0\right\} \tag{14}$$

where '$\bullet$' denotes dot-product.

The size or covering range of the neighbourhood function decreases from an initially large value to a final small one. The final neighbourhood, however, should not contain just the winner. The rigidity or curvature of the map is controlled by the ultimate size of the neighbourhood. The larger this size, the flatter the final map is in the data space. Guidelines for setting these parameters can be found in Yin (2002b).

Several improvements have since been made to the ViSOM. For instance, in Wu and Chow (2005) a probabilistic data assignment is used in both the input assignment and the neighbourhood function, and a second order constraint is adopted. In Estévez and Figueroa (2006) the ViSOM is extended to an arbitrary, neural gas type of map structure.

4. Connection with multidimensional scaling and principal manifolds

4.1. Self-organising maps and multidimensional scaling

MDS is a traditional approach to dimension reduction and data visualisation. MDS aims to project (or embed) high-dimensional data points or map proximities of objects onto a lower, often two dimensional space by preserving as faithfully as possible the inter-point metrics or the proximities (Borg & Groenen, 2005; Cox & Cox, 1994). The projection is generally nonlinear and can reveal the overall structure and mutual relationship of the data set.

Assume that the data is represented by either a set of n-dimensional vectors or given as a dissimilarity matrix. Let δ_{ij} denote the dissimilarity between objects i and j. For a set of n-dimensional data points, the dissimilarity δ_{ij} is often calculated (but not necessarily) by the (Euclidean) distance of data vectors $\delta_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$, $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{R}^n$. When objects are projected onto a lower dimensional space, let $\mathbf{y}_i$ and $\mathbf{y}_j$ be the mapped

points (coordinates) of objects i and j in the new space; then the distance between mapped points is $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$.

There are several types of MDS for different purposes and effects. *Classical MDS* looks for a configuration so that the distances between projected points match the original dissimilarities, i.e. $d_{ij} = \delta_{ij}, \forall i, j$. *Metric MDS* seeks a solution where the dissimilarities are proportional to the distances of projected points, $d_{ij} = f(\delta_{ij}), \forall i, j$; here f is a continuous, monotonic function, or a metric transformation function that transforms the dissimilarities to distance metrics. In practice, such exact dissimilarity-distance matches may not be possible due to data noise and imprecision; the equality is replaced by approximation, i.e. "$\approx$", meaning "as equal as possible" (Borg & Groenen, 2005).

Nonmetric MDS deals with rank order type dissimilarities and seek a configuration so that distances between pairs of mapped points match the dissimilarities order-wise "as well as possible".

An MDS configuration is often sought by minimising the following general cost, or the raw *Stress*, function,

$$S = \sum_{i,j} \left(f(\delta_{ij}) - d_{ij}\right)^2 \tag{15}$$

where f is the metric transformation function. In some cases, the above raw stress is normalised by $\sum_{i,j} d_{ij}^2$ or $\sum_{i,j} \delta_{ij}^2$ to give a relative reading of the overall stress. Other normalisation schemes are also possible. For example, in Sammon mapping (Sammon, 1969), an intermediate normalisation (pair-wise distance of original space) is used to preserve good local distributions and at the same time to maintain a global structure. The Sammon stress is expressed as,

$$S = \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}}. \tag{16}$$

A second-order Newton optimisation method is used to recursively solve the optimal configuration.

For general metric MDS, especially when the original dissimilarities need to be transformed to a distance-like form, f is a monotonic transformation function, e.g. a linear function. For classical MDS and many cases of metric MDS, when the data are given as high dimensional vectors, f is simply the identify function, and the stress becomes

$$S = \sum_{i,j} \left(\delta_{ij} - d_{ij}\right)^2. \tag{17}$$

That is, the metric MDS configuration tries to preserve, as well as possible, the pair-wise distances of original data points on the projected space.

For nonmetric MDS, f is a monotonic function satisfying

$$\text{if } \delta_{ij} \le \delta_{kl}, \text{ then } d_{ij} \le d_{kl}, \quad \forall i, j, k, l. \tag{18}$$

That is, nonmetric MDS produces an ordinal scaling rather than a metric one.

The similarities between SOMs and metric MDS in terms of topographic mapping, mostly the qualitative likeness of the mapping results, have been reported before (e.g. Ripley (1996)). However, some limitations of using the SOM for MDS have also been noted (Flexer, 1997) — the main one being that SOM does not preserve distance. Many applications combine the SOM and MDS for improved visualisation of the SOM projection results. In Ripley (1996), it is argued that the SOM is closer to MDS than to the principal manifold. In Yin (2002b), it is shown that the distance preserving ViSOM approximates a discrete principal manifold and also produces a similar mapping result as compared to the Sammon mapping — a metric MDS.

Let's take a close look at the cost function of metric MDS, Eq. (17). It can be rewritten as,

$$\sum_{i,j} (\delta_{ij} - d_{ij})^2 = \sum_{i,j} (\delta_{ij}^2 + d_{ij}^2 - 2\delta_{ij} d_{ij}). \tag{19}$$

The first term is a constant as data points are fixed, and the second term will be eventually fixed as it is to match the first term. To minimise the above stress is to maximise the third term. The third term plays a dominant role and explains that the mapping is to form a corresponding correlation between inter-point distances in the original and mapped spaces. This is closely related to the C measure.

From the cost function of the SOM, Eq. (5), we can see that the sample cost function – the integrand of Eq. (6) – can be expressed as (for the data contained in Voronoi region i):

$$\sum_k \eta(i,k) \, \|\mathbf{x} - \mathbf{w}_k\|^2 \tag{20}$$

where $\mathbf{x} \in V_i$, Voronoi region of neuron i, and $\mathbf{w}_k$ is the weight vector of neuron k.

As any data that belongs to neuron i will be quantised to, or be represented by, $\mathbf{w}_i$ the weight vector of neuron i, $\mathbf{x}$ can be replaced by $\mathbf{w}_i$ in the above expression, as far as the projection is concerned. Furthermore, $\eta(i,k)$ is a function of $\|\mathbf{r}_i - \mathbf{r}_k\|$ or d_{ik}. Then the above equation can be expressed as,

$$\begin{aligned} \sum_k \eta(i,k) \, \|\mathbf{x} - \mathbf{w}_k\|^2 &= \sum_k f(\|\mathbf{r}_i - \mathbf{r}_k\|) \, \|\mathbf{w}_i - \mathbf{w}_k\|^2 \\ &= \sum_k f(d_{ik}) \delta_{ik}^2. \end{aligned} \tag{21}$$

For the SOM, $f(d_{ik})$ is typically an exponential function of d_{ik}^2 — see Section 2.1. The first term of its Taylor expansion is proportional to d_{ik}^2. This leads the above cost function approximately to

$$\sum_k -(d_{ik}\delta_{ik})^2 \tag{22}$$

where d_{ik} represents the distance between the indexes (mapped data points) of the neurons i and k on the map grid.

Therefore, the SOM preserves the correlation between the pair-wise distances in both data and map spaces, as in MDS. In the standard SOM, the grid is fixed and the distance between nodes, d_{ik}, does not reflect metrically $\|\mathbf{w}_i - \mathbf{w}_k\|$, the distance of their weights in the input space, i.e. δ_{ik}, the distance of the

data points mapped onto these two nodes. Therefore, the scaling is to preserve the orders of the indexes of the neurons with the distances of their corresponding data regions in the input space. As the grid (coordinates) is prefixed and not scalable, the data points will be mapped to the grid positions to achieve maximum correlation. This is a qualitative scaling and does not preserve the distance in the mapped space. This, however, can also be regarded as a kind of generalisation as two spaces are no longer required to be in the same metric space. The SOM thus can be used to relate any two metric spaces by forming such a topological mapping, similarly to the nonmetric MDS condition, Eq. (18). That is, SOM is a nonmetric MDS or ordinal scaling.

In the ViSOM, however, as d_{ik} is proportional to $\|\mathbf{w}_i - \mathbf{w}_k\|$, as shown in Section 3. In the metric scaling sense, d_{ik} is a metric function of δ_{ij}. In other words it shows that the ViSOM is a distance-preserving, metric MDS. The squared distance correlation terms in Eq. (22) have an effect not much different from those non-squared ones of MDS in Eq. (19).

4.2. Nonlinear principal manifold and self-organising maps

Despite a number of recent approaches to extending linear PCA into a nonlinear variant, such as the auto-associative network (Kramer, 1991), generalised PCA (Karhunen & Joutsensalo, 1995), and kernel PCA (Schölkopf et al., 1998), the subject remains largely open and active. The principal curve and principal surface (Delicado, 2001; Hastie & Stuetzle, 1989) constitite the principled nonlinear extension of the PCA. The principal curve is defined as a smooth and self-consistent curve, which does not intersect itself, passing through the middle of the data. Denote $\mathbf{x}$ as a random vector in $\mathbf{R}^n$ with density p and a finite second moment. Let $\psi(\bullet)$ be a smooth unit-speed curve in $\mathbf{R}^n$, parameterised by the arc length ρ (from one end of the curve) over $\Pi \in \mathrm{R}$, a closed interval.

For a data point $\mathbf{x}$, its projection index on ψ is defined as,

$$\rho_\psi(\mathbf{x}) = \sup_{\rho\in\Pi}\{\rho : \|\mathbf{x} - \psi(\rho)\| = \inf_{\vartheta} \|\mathbf{x} - \psi(\vartheta)\|\}. \tag{23}$$

The curve is called a self-consistent principal curve of ρ if

$$\psi(\rho) = E[\mathbf{X}|\rho_\psi(\mathbf{X}) = \rho]. \tag{24}$$

For a finite data set of N points, the density is often not known, and the above expectation is replaced by a smoothing method, such as the kernel smoother

$$\psi(\rho) = \frac{\sum_{t=1}^{N} \mathbf{x}(t)\kappa(\rho, \rho_t)}{\sum_{t=1}^{N} \kappa(\rho, \rho_t)}. \tag{25}$$

The principal component is a special case of the principal curve if the data distribution is ellipsoidal. Although mainly principal curves have been studied, an extension to higher dimensions – for example, principal surfaces or manifolds – is feasible in principle. However, in practice, a good implementation of principal curve/surface relies on an effective and efficient algorithm.

The SOM has been related to discrete principal curve/surface (Ritter et al., 1992). However, differences remain in both the projection and the smoothing processes. In the SOM, data are projected onto the nodes rather than onto the curve/surface. The principal curve or surface performs the smoothing entirely in the data space. The smoothing process in the SOM and ViSOM, as a convergence criterion, is (e.g. Yin (2002b)),

$$\mathbf{w}_k = \frac{\sum_{t=1}^{N} \mathbf{x}(t)\eta(v, k, t)}{\sum_{t=1}^{N} \eta(v, k, t)}. \tag{26}$$

The smoothing is governed by the indexes of the neurons in the map space. The kernel smoothing uses the arc length parameters (ρ, ρ_i) or $\|\rho - \rho_i\|$ exactly, while the neighbourhood function uses the node indexes (k, i) or coordinates $(\mathbf{r}_k, \mathbf{r}_i)$. Arc lengths reflect the curve distances between the data points in the data space. However, the node indexes are integer numbers denoting the nodes or the positions on the map grid, not the positions in the data space. So $\|\mathbf{r}_k - \mathbf{r}_i\|$ does not resemble $\|\mathbf{w}_k - \mathbf{w}_i\|$ in the SOM. In the ViSOM, as the inter-neuron distances on the map represent those in the data space (subject to the resolution of the map and the quantisation error), therefore, $\lambda\|\mathbf{r}_k - \mathbf{r}_i\| \approx \|\mathbf{w}_k - \mathbf{w}_i\|$. Furthermore, the LLP can make this approximation even more precise. The smoothing process in the ViSOM resembles that of the principal curve as shown below:

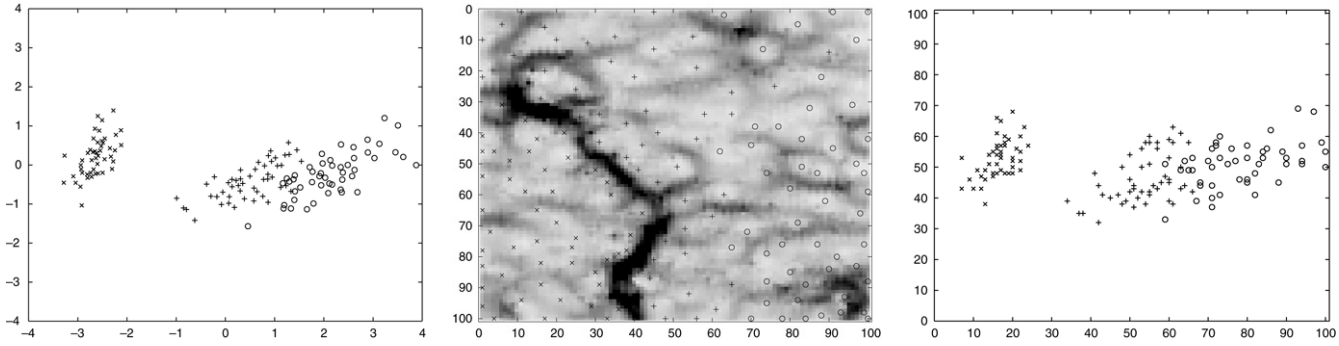
$$\mathbf{w}_k = \frac{\sum_{t=1}^{N} \mathbf{x}(t)\eta(v, k, t)}{\sum_{t=1}^{N} \eta(v, k, t)} \approx \frac{\sum_{t=1}^{N} \mathbf{x}(t)\eta(\mathbf{w}_v, \mathbf{w}_k, t)}{\sum_{t=1}^{N} \eta(\mathbf{w}_v, \mathbf{w}_k, t)}. \tag{27}$$

It shows that the ViSOM is a better approximation to the principal curve/surface than the SOM is. The SOM and ViSOM are similar only when the data are uniformly distributed, or when the number of nodes becomes very large, in which case both the SOM and ViSOM will closely approximate the principal curve/surface.

As the ViSOM is a discrete principal manifold, at the same time it is also a MDS. This implies that the MDS and principal manifold perform the same underlying task at least in the context of data visualisation and dimension reduction. Finding a principal manifold – a smooth curve/surface passing through the middle of the data – may well result in a topographic and metric scaling of the input space onto a lower dimensional manifold. On the other hand, although MDS presents a useful scaling of the data on a low dimensional space for visualisation, the principal manifold can provide the underlying mapping function.

4.3. Growing ViSOM for nonlinear embedding

Although we have shown that SOM and ViSOM are MDS methods and similar to the principal curve/surface, one of the difficulties for SOM-based algorithms is to converge to



(a) Sammon MDS. (b) SOM (with colouring). (c) ViSOM ($\lambda = 0.075$).

Fig. 2. Scaling of Iris data set by various methods. '×': Iris setosa, '+': Iris versicolor, 'o': iris virginica.

highly nonlinear manifolds such as swissroll-like data. Indeed high nonlinearity poses problems for many MDS and data projection methods. Isomap (Tenenbaum et al., 2000) adopts a geodesic distance within a small neighbourhood, instead of a global Euclidean distance in MDS to capture the nonlinearity of the data set. LLE (Roweis & Saul, 2000) uses local linear embedding to approximate global nonlinearity.

For a highly nonlinear manifold, it is often difficult for a map of pre-fixed size to converge correctly, either because the linear sub manifold is far from the true manifold, or due to the complex nature of untangling the map. Here, the core of the ViSOM, that it is local distance-preserving, is used to extract a highly nonlinear manifold. An incremental ViSOM or *growing* ViSOM (gViSOM) is proposed as below for embedding and metric-scaling nonlinear manifolds.

gViSOM algorithm:

(1) Start with a small initial map, say $M_0 \times M_0$, in either rectangular or hexagonal, though the latter is preferred for better nonlinear abilities. Place the initial map onto a linear subspace of either the entire or a local region of the data space. Set the desired resolution and the neighbourhood (locality) size.
(2) Randomly draw a data sample from the data space and find the winning neuron with the shortest distance.
(3) If the sample falls within the neighbourhood, update the weights of the neurons of the neighbourhood using the ViSOM principle; otherwise go back to Step 2.
(4) At regular iteration intervals (for instance, every 1000 iterations), if the growing condition is met (that is, the data is underrepresented by the existing map), grow the map by adding a column or row to the side with the highest activities (measured by the winning frequencies). The added column or row is a linear extrapolation of the existing map. Other growing structures can be used, such as incrementing polygons instead of entire column or row for a free structure of the map and efficient use of neurons.
(5) As in the ViSOM, at regular intervals (every certain number of iterations), refresh the map (neurons) probabilistically.
(6) Check if the map has converged. If not go back to Step (2); if so go to the next step.
(7) Project the data samples onto the map, either to the neurons or by the LLP resolution enhancement method.

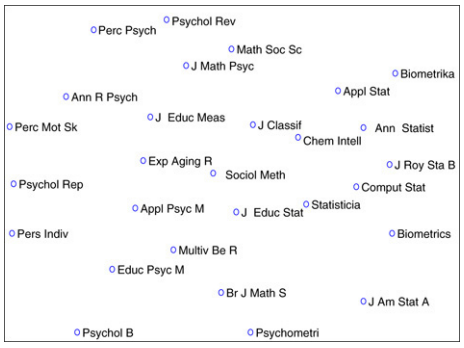
5. Experiments and comparisons

The first example is on the Iris data set, which consists of 150 4-D vectors of three iris categories, each having 50 samples. The results of metric MDS (Sammon mapping), SOM and ViSOM are shown in Fig. 2. A 100×100 hexagonal grid is used for both the SOM and ViSOM and the U matrix colouring is applied to the SOM result. All three methods initialise the mapping on the plane spanned by the first two eigenvectors. As can be seen, ViSOM produces a metric scaling similar to the Sammon mapping; while the SOM provides a qualitative scaling of the data points. The advantage of a metric scaling is that the distribution of the samples or objects is faithfully preserved.

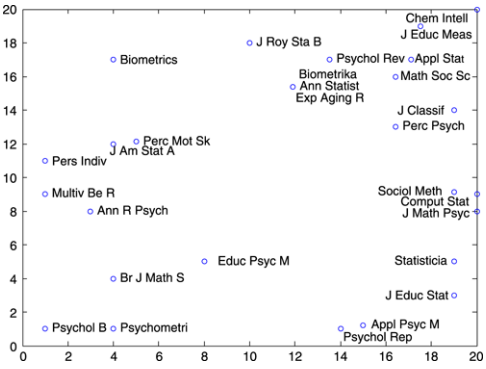
The next experiment is on a journal citation data set used for studying MDS (Groenen & Heiser, 1996). The data set consists of the numbers of citations of 28 journals from the same set of journals in the psychometric literature. The raw data resemble similarity measures. They were scaled by a natural logarithm to provide a more natural effect of the citations, and self-citations were removed. The scaling results of these 28-D vectors by Sammon MDS, SOM and ViSOM are shown in Fig. 3. All three methods have produced some meaningful configurations showing the interrelationships among the journals, with statistical journals being mainly positioned on the left and psychological journals generally on the right. Both SOM and ViSOM maps are in the form of a 20×20 rectangular grid. The ViSOM scaling is more similar to the metric MDS, while the SOM simply confines all the objects topologically onto the grid.

The last experiment is on embedding the highly nonlinear 3-D Swissroll data by the gViSOM, Isomap and LLE. Here 2000 data points were generated according to (Roweis & Saul, 2000). The gViSOM started with a 5×5 grid and finally settled to 18×70. The resolution was set to 1.5. Fig. 4(a) shows the gViSOM embedding in the data space with its final flattened manifold revealed in Fig. 4(b). Typical results of Isomap and LLE, obtained using the code provided by their authors, are shown in Fig. 4(c) and (d), respectively.

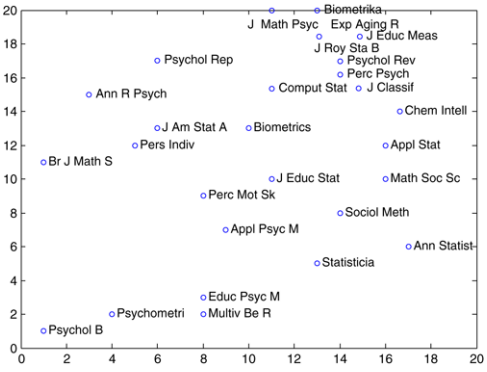
The advantages of the gViSOM are evident as it produces a much more faithful metric scaling and is able to extract a highly nonlinear manifold function. In addition, it can cope with discontinuities in the manifold (for example, holes in



(a) Metric MDS (Sammon mapping).



(b) SOM scaling.



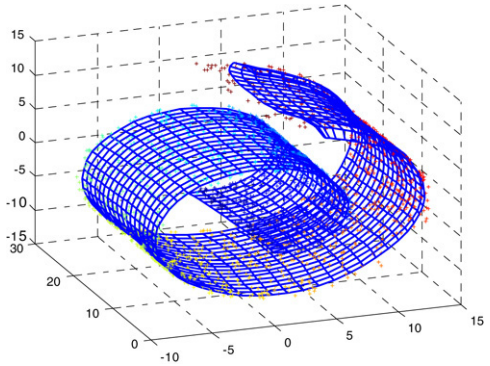
(c) ViSOM ($\lambda = 5$).

Fig. 3. Metric MDS, SOM and ViSOM scaling of Citation data set.

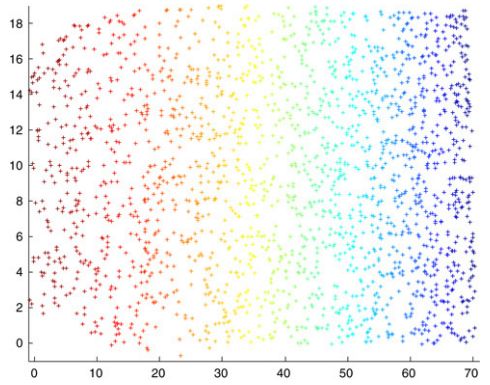
manifold or disjoint manifold), and adapt to the dynamics (slow changes) of the manifold, for which both Isomap and LLE (and other scaling methods) would have to re-capture with the entire data set once any part or whole were updated. SOM-based methods also have better abilities for handling noise.

6. Conclusions

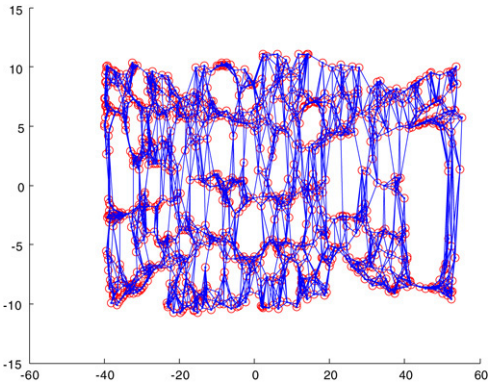
This paper provides a review on the self-organising map (SOM) and the issues surrounding its cost function and topology measures. It then reveals the connection between the SOM (or its variant ViSOM) and multidimensional scaling (MDS) through analysing their cost functions, though such an analogy has been reported in the literature before, mainly based on experimental results. The analysis shows that SOMs



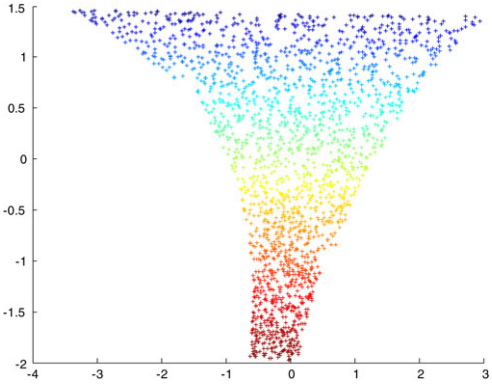
(a) Swissroll data and gViSOM embedding.



(b) Extracted manifold by gViSOM ($\lambda = 1.5$).



(c) Isomap embedding.



(d) LLE embedding.

Fig. 4. Embedding of Swissroll data set by gViSOM, Isomap and LLE.

and MDS are similar mappings for the principle of data visualisation. Furthermore, the ViSOM is closer to MDS than SOM is in terms of metric scaling and distance preservation. However, SOMs can be regarded as a generalised scaling that associate or order two possibly different metric spaces. It also reveals that the metric MDS and the principal manifold essentially produce comparable topographic embeddings for visualisation. The metric MDS is a point-to-point mapping on dissimilarities; while the principal manifold approach can establish an explicit mapping function of the data set. The core of the ViSOM, namely local distance-preserving mapping, has been used to provide an adaptive technique for dealing with highly nonlinear, complex data sets.

Acknowledgements

The author wishes to thank the reviewers for their valuable comments and suggestions. The work was partially supported by the UK EPSRC grant EP/E057101/1.

References

Bauer, H. -U., & Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, *3*, 570–579.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). Springer.

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2006). Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *PNAS*, *103*, 1168–1172.

Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. Chapman & Hall.

de Bolt, E., Cottrell, M., & Verleysen, M. (2002). Statistical tools to assess the reliability of self-organizing maps. *Neural Networks*, *15*, 967–978.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, *77*, 84–116.

Durbin, R., & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, *343*, 644–647.

Erwin, E., Obermayer, K., & Schulten, K. (1992a). Self-organizing maps: Ordering, convergence properties and energy functions. *Biological Cybernetics*, *67*, 47–55.

Erwin, E., Obermayer, K., & Schulten, K. (1992b). Self-organizing maps: Stationary states, metastability and convergence rate. *Biological Cybernetics*, *67*, 35–45.

Estévez, P. A., & Figueroa, C. J. (2006). Online data visualization using the neural gas network. *Neural Networks*, *19*, 923–934.

Flexer, A. (1997). Limitations of self-organizing maps for vector quantization and multidimensional scaling. In *Proc. NIPS'97* (pp. 445–451).

Goodhill, G. J., & Sejnowski, T. (1997). A unifying objective function for topographic mappings. *Neural Computation*, *9*, 1291–1303.

Groenen, P. J. F., & Heiser, W. J. (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika*, *61*, 529–550.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, *84*, 502–516.

Heskes, T. (1999). Energy functions for self-organizing maps. In E. Oja, & S. Kaski (Eds.), *Kohonen maps* (pp. 303–315).

Karhunen, J., & Joutsensalo, J. (1995). Generalization of principal component analysis, optimization problems, and neural networks. *Neural Networks*, *8*, 549–562.

Kohonen, T. (1982). Self-organized formation of topologically correct feature map. *Biological Cybernetics*, *43*, 56–69.

Kohonen, T. (1984). *Self-organization and associative memory*. Springer-Verlag.

Kohonen, T. (1986). Representation of sensory information in self-organizing feature maps, and relation of these maps to distributed memory networks. In *Proc. SPIE*: *Vol. 634* (pp. 248–259).

Kohonen, T. (1991). Self-organizing maps: Optimization approaches. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kanguas (Eds.), *Artificial neural networks*: *Vol. 2* (pp. 981–990). Amsterdam: North-Holland.

Kohonen, T. (1997). *Self-organizing maps* (2nd ed.). Springer.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AICHE Journal*, *37*, 233–243.

Lampinen, J., & Oja, E. (1992). Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, *2*, 261–272.

Lin, S., & Si, J. (1998). Weight-value convergence of the SOM algorithm for discrete input. *Neural Computation*, *10*, 807–814.

Lo, Z. P., & Bavarian, B. (1991). On the rate of convergence in topology preserving neural networks. *Biological Cybernetics*, *65*, 55–63.

Lowe, D., & Tipping, M. E. (1996). Neuroscale: Novel topographic feature extraction using RBF networks. *Neural Computing and Applications*, *4*, 83–95.

Luttrell, S. P. (1990). Derivation of a class of training algorithms. *IEEE Transactions on Neural Networks*, *1*, 229–232.

Luttrell, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation*, *6*, 767–794.

Malthouse, E. C. (1998). Limitations of nonlinear PCA as performed with generic neural networks. *IEEE Transactions on Neural Networks*, *9*, 165–173.

Mao, J., & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, *6*, 296–317.

Mitchison, G. (1995). A type of duality between self-organizing maps and minimal wiring. *Neural Computation*, *7*, 25–35.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Ritter, H., Martinetz, T., & Schulten, K. (1992). *Neural computation and self-organizing maps: An introduction*. Addison-Wesley Publishing Company.

Ritter, H., & Schulten, K. (1988). Convergence properties of Kohonen's topology conserving maps: Fluctuations, stability, and dimension selection. *Biological Cybernetics*, *60*, 59–71.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computer*, *18*, 401–409.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.

Villmann, T., Der, R., Herrmann, M., & Martinetz, T. M. (1997). Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, *8*, 256–266.

von der Malsburg, C., & Willshaw, D. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, *4*, 85–100.

Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London, Series B*, *194*, 431–445.

Wu, S., & Chow, T. W. S. (2005). PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks*, *16*, 1362–1380.

Yin, H. (1996). Self-organizing maps: Statistical analysis, treatment and applications. *Ph.D. thesis*. University of York.

Yin, H. (2002a). ViSOM-A novel method for multivariate data projection and structure visualization. *IEEE Transactions on Neural Networks*, *13*, 237–243.

Yin, H. (2002b). Data visualisation and manifold mapping using the ViSOM. *Neural Networks*, *15*, 1005–1016.

Yin, H. (2003). Resolution enhancement for the ViSOM. In *Proc. workshop on self-organizing maps* (pp. 208–212).

Yin, H. (2007). Connection between self-organizing maps and metric multidimensional scaling. In *Proc. IJCNN 2007* (pp. 1025–1030).

Yin, H., & Allinson, N. M. (1995). On the distribution and convergence of the feature space in self-organizing maps. *Neural Computation*, *7*, 1178–1187.