# Nonlinear Multidimensional Data Projection and Visualisation

Hujun Yin

Dept. of Electrical Engineering and Electronics
University of Manchester Institute of Science and Technology
Manchester, M60 1QD, UK
h.yin@umist.ac.uk

**Abstract.** Multidimensional data projection and visualisation are becoming increasingly important and have found wide applications in many fields such as decision support, bioinformatics and web/document organisation. Various methods and algorithms have been proposed as either nonparametric or semi-parametric approaches. This paper provides an overview of the subject and reviews some recent developments. Relationships among various key methods such as Sammon mapping, Neuroscale, principal curve/surface, SOM, GTM and ViSOM are analysed and their advantages and limitations are highlighted in the context of nonlinear principal component analysis and independent component analysis.

## 1 Introduction

Data projection and visualisation methods are becoming widely used tools in many fields such as decision support [8], financial analysis [1], information/knowledge management [11] and bioinformatics [51]. Searching for a suitable data mapping method has always been an integral objective of multivariate data analysis and pattern recognition. Projecting data on to its underlying subspace can detect its real structures, facilitate functional analysis, and help making a judgment. A great deal of research has been devoted to this subject and a number of methods have been proposed.

Classic projection methods include the linear principal component analysis (PCA) and the multidimensional scaling (MDS). The PCA projects the data onto its principal directions, which are represented by the orthogonal eigenvectors of the covariance matrix of the data. It is the optimal linear projection in the sense of minimum mean-square-error between the original data points and the projected ones on the principal subspace. But the PCA's linearity has limited its power for practical data, as it cannot capture nonlinear relationships defined by higher than second order statistics. If the input dimensionality is much higher than two, the projection onto a linear plane will provide limited visualisation power. An extension to nonlinear PCA, in principle, could tackle practical problems better. However there is no single and unique solution to nonlinear PCA [34]. Various methods have been proposed, for example the auto-associative networks [27], generalised PCA [20], kernel PCA [46], and the principal

J. Liu et al. (Eds.): IDEAL 2003, LNCS 2690, pp. 377–388, 2003.

curves and surfaces [14, 28]. Other mapping methods include the recently proposed local, geometric based grouping and averaging [49] and local linear embedding [43].

MDS tries to project data points onto a two-dimensional plane by preserving as close as possible the inter-point metrics [9, 42]. The mapping generally is nonlinear and can reveal the overall structure of the data. Sammon [45] mapping is a widely known example of MDS. However, like other MDS methods, the Sammon algorithm is a point-to-point mapping, which does not provide the explicit mapping function [45, 35].

Neural networks present another approach to nonlinear data analysis or projection. A feedforward neural network has been proposed to parametrise the Sammon mapping function and a back-propagation algorithm has been derived for training of the network and minimising the Sammon stress [35]. Neuroscale [33] is another realisation of the MDS using the radial basis function. The self-organising map (SOM) is an abstract mathematical model of the mapping between nerve sensory and cerebral cortex [24, 25]. The SOM has been widely used as a visualisation tool for dimensionality reduction (e.g. [52, 26]). The generative topographic mapping (GTM) [3] parametrises the SOM using mixture of Gaussians. The SOM's topology preserving property can be utilised to visualise the relative mutual relationships among the input. However, the SOM does not directly apply to scaling, which aims to reproduce proximity in (Euclidean) distance on a low visualisation space, as it has to rely on a colouring scheme to imprint the distances –that is very crude and often the distributions of the data points are distorted on the map. The recently proposed visualisation induced SOM (ViSOM) [54, 55] constrains the lateral contraction force between the neurons in the SOM and hence regularises the inter-neuron distances with respect to a scaleable parameter that defines and controls the resolution of the map. It preserves the data structure as well as the topology as faithfully as possible. The ViSOM provides a direct visualisation of both the structure and distribution of the data.

The remaining of the paper provides a review on various methods. The relationships among these methods are analysed and drawn. The advantages and drawbacks of these different approaches are also elucidated.

## 2 MDS and Sammon Mapping

Multidimensional scaling (MDS) is a traditional subject related to dimension reduction and data projection, which includes PCA as one of the projection methods. MDS tries to project data points onto a two-dimensional sheet or plot by preserving as close as possible the inter-point metrics [9, 42]. The projection is generally nonlinear and can reveal the overall structure of the data.

A general fitness function or so-called *stress* can be described as,

$$S = \frac{\sum_{i,j}[d_{ij} - D_{ij}]^2}{\sum_{i,j} D_{ij}^2} \tag{1}$$

where $d_{ij}$ represents the proximity of points $i$ and $j$ in the original space, $D_{ij}$ represents the distance (usually Euclidean) between mapped points $i$ and $j$ in the new space,.

There is no exact and unique procedure to find the projection. Instead, the MDS relies on an optimisation algorithm to search for a configuration that gives as low stress as possible. A gradient method is commonly used for this purpose. Inevitably, various computational problems such as local minima and divergence may occur to the optimisation process. The methods are also often computationally intensive. The final solution depends on the starting configuration and parameters used in the algorithm [5].

### 2.1 Sammon Mapping

Sammon mapping is a well-known example of MDS. The objective of Sammon mapping is to minimise the differences between inter-point distances in the original space and those in the projected plane. The Sammon mapping has been shown to be useful for data structure analysis (e.g. [45], [41]). However, like other MDS methods, the Sammon algorithm is a point-to-point mapping, which does not provide the explicit mapping function and cannot naturally accommodate new data points [45, 35]. It also requires to compute and store all the inter-point distances. This proves difficult or even impossible for many practical applications where data arrives sequentially, the quantity of data is large, and/or memory space for the data is limited.

In Sammon mapping intermediate normalisation (of original space) is used to preserve good local distributions and at the same time maintain a global structure. The transformation function is simply the Euclidean distance between points $i$ and $j$ in the original space. The Sammon stress is expressed as,

$$S_{Sammon} = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{[d_{ij} - D_{ij}]^2}{d_{ij}} \tag{2}$$

Sammon proposed a recursive learning algorithm using the Newton optimisation method for the optimal configuration [45]. It converges faster than the simple gradient method, but the computational complexity is even higher. It still has the local minima and inconsistency problems.

In addition to being computationally costly, especially for large data sets, and not adaptive, another major drawback of MDS methods including Sammon mapping is lack of an explicit projection function. Thus for any new input data, the mapping has to be recalculated based on all available data. Although some methods have been proposed to accommodate the new arrivals using triangulation [29, 5, 10], the methods are generally not adaptive.

### 2.2 Neural Approaches to MDS

Mao and Jain [35] have proposed to use a feedforward neural network, termed SAMANN, to parametrise the Sammon mapping function and a unsupervised training methods has been derived for training of the network. The derivation is similar to the back-propagation algorithm, by minimising the Sammon stress instead of the total errors between desired and actual output. The network takes a pair of input points at

each time in the training. An evaluation has to be carried out, using all the data points, after a fixed number of iterations.

In the SAMANN, all the inter-point distances have to be normalised before being input to the network. This will result clamping of any new data points whose distances to previous data points are larger the initial normalising scale. The algorithm is a gradient descent method and relies on a good initialisation.

Neuroscale uses radial basis function (RBF) to generalise the MDS [33]. It minimises a simple unweighted stress function, i.e. Eq (1) without the denominator. A subjective element has also been incorporated into the objective function and produces a projection that takes into account of subjective perception of data attributes, if such a prior knowledge exists such as in grading systems.

## 3 Nonlinear Extensions of PCA and ICA

PCA is a classic linear data analysis method aiming at finding orthogonal principal directions from a set of data, along which the data exhibit the largest variances. By discarding the minor components, the PCA can effectively reduce data variables and display the dominant ones in a linear, low dimensional subspace. It is the optimal linear projection in the sense of the mean-square error between original points and projected ones, i.e.,

$$\min \sum_{\mathbf{x}} [\mathbf{x} - \sum_{j=1}^{m} (\mathbf{q}_j^T \mathbf{x}) \mathbf{q}_j]^2 \tag{3}$$

where $\mathbf{x}=[x_1, x_2, \ldots x_n]^T$ is the $n$-dimensional input vector, $\{\mathbf{q}_j, j=1,2, \ldots m, m \le n\}$ are orthogonal vectors representing principal directions. They are the first $m$ principal eigenvectors of the covariance matrix of the input. The second term in the above bracket is the reconstruction or projection of $\mathbf{x}$ on these eigenvectors. The term $\mathbf{q}_j^T \mathbf{x}$ represents the projection of $\mathbf{x}$ onto the $j$-th principal dimension. Traditional methods for solving eigenvector problem involve numerical methods. Though fairly efficient and robust, they are not usually adaptive and often require the presentation of the entire data set. Several Hebbian-based learning algorithms and neural networks have been proposed for performing PCA such as, the subspace network [37] and the generalised Hebbian algorithm [44]. The limitation of linear PCA is obvious, as it cannot capture nonlinear relationships defined by higher than the second order statistics. If the input dimension is much higher than two, the projection onto linear principal plane will provide limited visualisation power.

The extension to nonlinear PCA (NLPCA) is not unique, due to the lack of a unified mathematical structure and an efficient and reliable algorithm, in some cases due to excessive freedom in selection of representative basis functions [34, 20]. Principal curves and principal surfaces [14, 41] were primary nonlinear extension of PCA, but a valid algorithm is required for a good implementation. Several networks have been proposed for nonlinear PCA such as, the five-layer feedforward associative network [27] and the kernel PCA [46]. The first three layers of the associative network project the original data on to a curve or surface, providing an activation value for the bottle-

neck node. The last three layers define the curve and surface. The weights of the associative NLPCA network are determined by minimising the following objective function,

$$\min \sum_{\mathbf{x}} \| \mathbf{x} - \mathbf{f}\{s_f(\mathbf{x})\} \|^2 \tag{4}$$

where $\mathbf{f}$: $\mathbb{R}^1 \rightarrow \mathbb{R}^n$ (or $\mathbb{R}^2 \rightarrow \mathbb{R}^n$), the function modelled by the last three layers, defines a curve (or a surface), $s_f$: $\mathbb{R}^n \rightarrow \mathbb{R}^1$ (or $\mathbb{R}^n \rightarrow \mathbb{R}^2$), the function modelled by the first three layers, defines the projection index.

The kernel-based PCA [46] uses nonlinear mapping and kernel functions to generalise PCA to NLPCA and has been used for character recognition. The nonlinear function $\Phi(\mathbf{x})$ maps data onto high-dimensional feature space, where the standard linear PCA can be performed via kernel functions: $k(\mathbf{x}, \mathbf{y})=(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$. The projected covariance matrix is then

$$\overline{C} = \frac{1}{N} \sum_{i=1}^{N} \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T \tag{5}$$

The standard linear eigenvalue problem can now be written as $\lambda\mathbf{V}=\mathbf{K}\mathbf{V}$, where the columns of $\mathbf{V}$ are the eigenvectors and $\mathbf{K}$ is a $N$x$N$ matrix with elements as kernels $K_{ij}:=k(\mathbf{x}_i, \mathbf{x}_j)=(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j))$.

(Linear) ICA is another way to extend and generalise the PCA. With the assumption of linear mixtures, ICA can decompose and map data set into statistically independent components. It has been studied intensively recently. Many theories regarding various aspects of the linear ICA have been established and confirmed experimentally [30, 17, 12]. However, in general and for many practical problems, mixtures are more likely to be nonlinear or subject to some kind of nonlinear distortions due to sensory or environmental limitations. Extension of existing theories and methods to nonlinear ICA (NLICA) is not straightforward. There have been few initial attempts (e.g. [39]). ICA and NLICA can be approached through non-linear PCA (NLPCA), SOM, and mixture models, and some work has demonstrated tentative links [38, 15, 40, 53].

The generic NLICA problem can be formulated as $\mathbf{X}(t)=F[\mathbf{S}(t)]$, where $\mathbf{S}(t)$ are a set of unknown source signals and $\mathbf{X}(t)$ the observation or measurements, $F$ is unknown and generally a nonlinear transformation. In real problems there is a noise process $\mathbf{V}$(t) associated with either source or observed signals. This noise term can be additive and multiplicative, and with various distributions (either correlated or not with the source signals). The complexity of the noisy NLICA model suggests the use of a flexible method that may need to be tailored to the experimental context. Some research has addressed a compromise between standard linear and purely NLICA methods, such as the ICA mixture models [32] and the local linear ICA using $k$-means clustering [21]. The first tries to relax the independence assumption of the generic ICA model. While the second is closely related to the batch version of the SOM, but with standard $k$-means clustering used, because of the fairly small number of clusters involved.

Various neural networks have also been applied to the NLICA problem. The use of multilayer perceptrons (MLP) in biomedical applications has been studied [31], while earlier a two-layer perceptron was employed [6]. An RBF network [48] has been used

to recover the unknown sources from their nonlinear mixtures in presence of cross-nonlinearities. This method appears robust against additive, uniformly distributed white noise, but a further noise suppression technique is necessary to denoise the separated source signals. SOMs have been used [15, 40] to extract independent components from nonlinearly mixed discrete or continuous sources. The network complexity increases with the number of neurons while the quantization error (interpolation error, in the continuous case) cannot be disregarded. A SOM-based NLICA method has been used to denoise multiplicative noise [13]. A post-nonlinear mixing has been proposed by [32] and [47]. In this case, the sources are assumed to be linearly mixed and then transformed by a nonlinear transfer channel. This parametric approach uses sigmoidal functions and MLP networks to approximate the inverse nonlinearity. However, the approach is limited to a certain class of nonlinear mixtures. A generalization to a rather larger class of functions has been given by [18] using the notion of conformal mapping into the complex domain. Generally existing NLICA methods can be classified into two categories. The first models the nonlinear mixing as a linear process followed by a nonlinear transfer channel. These methods are of limited flexibility as they are often parametrized. The second category employs parameter-free methods, which are more useful in representing more generic nonlinearities. A common neural technique in this second category is the SOM, which can be used to model and extract the underlying nonlinear data structures.

## 4 Principal Curves and Surfaces

The principal curves and principal surfaces (14, 28] are primary nonlinear extension of PCA, but a valid algorithm is required for a good implementation. The principal curve was first defined as a smooth and self-consistency curve, which does not intersect itself. Denote $\mathbf{x}$ as a random vector in $\mathbb{R}^n$ with density $p$ and finite second moment. Let $f(\cdot)$ be a smooth unit-speed curve in $\mathbb{R}^n$, parametrised by the arc length $\rho$ (from one end of the curve) over $\Lambda \in \mathbb{R}$, a closed interval.

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

$$\rho_f(\mathbf{x}) = \sup_{\rho \in \Lambda} \{\rho : \|\mathbf{x} - f(\rho)\| = \inf_{\vartheta} \|\mathbf{x} - f(\vartheta)\|\} \tag{6}$$

The curve is called self-consistent or a principal curve of $p$ if

$$f(\rho) = E[\mathbf{X} \mid \rho_f(\mathbf{X}) = \rho] \tag{7}$$

The principal component is a special case of the principal curves if the distribution is ellipsoidal. Although 1-D principal curves have been mainly studied, extension to higher dimension, e.g. principal surfaces is feasible in principle. However, in practice, a good implementation of principal curves/surfaces relies on an effective and efficient algorithm.

The principal curves/surfaces are more of a concept that invites practical implementations. The HS algorithm is a nonparametric method [14], which directly iterates the two steps of the above definition. It is similar to the LGB VQ algorithm, combined with some smoothing techniques.

*HS algorithm:*

*Initialisation*: Choose the first linear principal component as the initial curve, $f^{(0)}(\mathbf{x})$.

*Projection*: Project the data points onto the current curve and calculate the projections index, i.e. $\rho^{(t)}(\mathbf{x})=\rho_{f(t)}(\mathbf{x})$.

*Expectation*: For each index, take the mean of data points projected onto it as the new curve point, i.e., $f^{(t+1)}(\rho)=E[\mathbf{X}|\rho_{f(t)}(\mathbf{X})=\rho]$.

The projection and expectation steps are repeated until a convergence criterion is met, e.g. when the change of the curve between iterations is below a threshold.

For a finite data set, the density $p$ is often unknown, the above expectation is replaced by a smoothing method such as the locally weighted running-line smoother or smoothing splines. For kernel regression, the smoother is,

$$f(\rho)=\frac{\sum_{i=1}^{N}\mathbf{x}_i\kappa(\rho,\rho_i)}{\sum_{i=1}^{N}\kappa(\rho,\rho_i)} \tag{8}$$

The arc length is simply computed from the line segments. There are no proofs of convergence of the algorithm, but no convergence problems have been reported, though the algorithm is biased in some cases [14]. Banfield and Raftery [2] have modified the HS algorithm by taking the expectation of the residual of the projections in order to reduce the bias. Kegl et al [23] have proposed an incremental, e.g. segment by segment, and arc length constrained method for practical construction of principal curves.

Tibshirani [50] has introduced a semi-parametric model for the principal curve. A mixture model was used to estimate the noise along the curve; and the expectation and maximisation (EM) method was employed to estimate the parameters. Other options for finding the nonlinear manifold include the GTM [3] and probabilistic principal surfaces (PPS) [7]. These methods model the data by a means of a latent space. They belong to the semi-parametrised mixture model, although types and orientations of the local distributions vary from method to method.

## 5 SOM, GTM, and ViSOM

The SOM is an unsupervised learning algorithm that uses a finite grid of neurons to frame the input space. As the map is often arranged in a low dimensional, e.g. 2-D, grid, it can be used for visualisation of potentially high dimensional data on a visible dimension. In the SOM, neighbourhood learning is adopted to form topological ordering among the neurons in the map. The mapping is generally nonlinear. The close data points are likely projected to nearby nodes. Thus the map can be used to show the relative relationships among the data points. However, the SOM does not directly

show the inter-neuron distances on the map. For visualisation, the SOM requires assistance from a colouring scheme to imprint the inter-neuron distances and therefore the clusters and boundaries can be marked. The colour or grey tone of a node or a region between nodes is proportional to the mean or median of the distances between that node and its nearest neighbours. Such a colouring method has been used in many data visualisation applications, e.g. WEBSOM [16] and World Welfare Map [22]. The colouring methods indeed enhance the visualisation ability of the SOM. However, the cluster structures and distribution of the data shown on the map often are not apparent and appear in distorted and unnatural forms. Other techniques to mark the inter neuron distances include calculating the magnification factors or the Jacobians [4] and interpolation [57]. The SOM can serve as a visualisation map only in showing the relative closeness and relationships among data points and clusters. This is also the case for the GTM, which is a parametrised approach to the SOM [3]. It uses a set of RBFs to map a latent 2-D grid into the high dimensional data space,

$$Y(\mathbf{x}, \mathbf{W})=\mathbf{W}\phi(x) \tag{9}$$

In the GTM, the data is modelled by a mixture of homoscedastic Gaussians. Then the EM algorithm is used to learn the parameters: the mapping $\mathbf{W}$ and the common variance $\sigma$ of the Gaussians. The PPS [7] has adopted a general Gaussian mixture and oriented covariance noise model in the GTM and results in a better approximation to principla curves and surfaces. The recently proposed latent trait model (LTM) [19] generalises the GTM for both discrete and categorical data distributions. The SOM can also be directly parametrised using mixture of Gaussians such as the Bayesian SOM [58] for modelling the data density.

In many cases, however, a direct and faithful display of structure shapes and distributions of the data is highly desirable in visualisation applications. ViSOM has been proposed to directly preserve distances on the map [54, 55]. For the map to capture the data structure naturally and directly, the distance quantity must be preserved on the map, along with the topology. Ideally the nodes should be uniformly and smoothly placed in the nonlinear manifold of the data space. 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 the ViSOM, lateral contraction force is constrained in the learning rule,

$$\mathbf{w}_k(t+1)=\mathbf{w}_k(t)+ \alpha(t)\eta(v,k,t)\{[\mathbf{x}(t)-\mathbf{w}_v(t)]+ \beta[\mathbf{w}_v(t)-\mathbf{w}_k(t)]\} \tag{10}$$

where the simplest constraint can be $\beta:=d_{vk}/(\Delta_{vk}\lambda)-1$, with $d_{vk}$ the distance in the input space, $\Delta_{vk}$ the distance on the map, and $\lambda$ a resolution constant.

The ViSOM regularises the contraction force so that the distances between the nodes on the map are analogous to the distances of their weights in the data space. The aim is to adjust inter-neuron distances on the map in proportion to those in the data space, i.e. $\Delta_{vk} \propto d_{vk}$. When the data points are eventually projected on a trained map, the distance between point $i$ and $j$ on the map is proportional to that of the original space, subject to the quantisation error (the distance between a data point and its neural representive). This has a similar effect to Sammon mapping, which also aims at achieving this proportionality, $D_{ij} \propto d_{ij}$, though here $D_{ij}$ represents the distance of two mapped data

points. When the number of nodes increases, the quantisation errors reduces. The key feature of the ViSOM is that the distances between the neurons on the map (in a neighbourhood) reflect the corresponding distances in the data space. When the map is trained and data points mapped, the distances between mapped data points on the map will resemble approximately those in the original space (subject to the resolution of the map). This makes visualisation more direct, quantitatively measurable, and visually appealing. The size or covering range of the neighbourhood function can also be decreased from an initially large value to a final smaller 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 of this size the flatter the final map is in the data space.

The SOM has also been related to the discrete principal curve/surface algorithm [42, 36]. However the differences remain in both the projection and smoothing processes. In the SOM the data are projected onto the nodes rather than onto the curve. The principal curves perform the smoothing entirely in the data space –see Eq. (8). The smoothing process in the SOM and ViSOM, as a convergence criterion, is [56],

$$\mathbf{w}_k = \frac{\sum_{i=1}^{L} \mathbf{x}_i \eta(v,k,i)}{\sum_{i=1}^{L} \eta(v,k,i)} \tag{11}$$

The smoothing is governed by the indexes of the neurons in the map space. The kernel regression uses the arc length parameters ($\rho$, $\rho_i$) or $\|\rho-\rho_i\|$ exactly, while the neighbourhood function uses the node indexes ($k$, $i$) or $\|k-i\|$. Arc lengths reflect the curve distances between the data points. However, node indexes are integer numbers denoting the nodes or the positions on the map grid, not the positions in the input space. So $\|k-i\|$ does not resemble $\|\mathbf{w}_k-\mathbf{w}_i\|$ in the common SOM. In the ViSOM, however, as the inter-neuron distances on the map represent those in the data space (subject to the resolution of the map), the distances of nodes on the map are in proportion to the difference of their positions in the data space, i.e. $\|k-i\| \sim \|\mathbf{w}_k-\mathbf{w}_i\|$. The smoothing process in the ViSOM resembles that of the principal curves as shown below,

$$\mathbf{w}_k = \frac{\sum_{i=1}^{L} \mathbf{x}_i \eta(v,k,i)}{\sum_{i=1}^{L} \eta(v,k,i)} \approx \frac{\sum_{i=1}^{L} \mathbf{x}_i \eta(\mathbf{w}_v,\mathbf{w}_k,i)}{\sum_{i=1}^{L} \eta(\mathbf{w}_v,\mathbf{w}_k,i)} \tag{12}$$

The ViSOM is a better approximation to the principal curves/surfaces 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 curves/surfaces.

## 6 Conclusions

Nonlinear projections of multidimensional data have been approached from various aspects such as MDS, nonlinear PCA, nonlinear ICA, and principal manifold such as principal curves and surfaces. Various realisation methods have been proposed and

proved to be useful in various data modelling and visualisation applications. Among these methods, the SOM and ViSOM seem to be the most versatile and nonparametric methods not only in data visualisation, but also in capturing the nonlinear manifold of the data. The ViSOM is a natural algorithm for extracting discrete principal curves and surfaces.

## References


1 Arciniegas, I., Daniel, B., Embrechts, M. J.: Exploring Financial Crises Data with self-organising maps. *Advances in Self-Organising Maps*. Allinson, N, Yin, H., Allinson, L. and Slack J. (Eds). (2001) 39–46

2 Banfield, J. D. and Raftery, A. E.: Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87 (1992) 7–16.

3 Bishop, C. M., Svensén, M., and Williams, C. K. I.: GTM: The generative topographic mapping. *Neural Computation*, 10 (1998) 215–235

4 Bishop, C. M., Svensén, M., and Williams, C. K. I.: Magnification factors for the SOM and GTM algorithms. *Proceedings of Workshop on Self-Organizing Maps (WSOM'97),* 333–338.

5 Biswas, G., Jain, A. K., & Dubes, R. C.: Evaluation of project algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-3 (1981) 701–708

6 Burel, G.: Blind separation of sources: A nonlinear neural algorithm. *Neural Networks*, 5 (1992) 937–947.

7 Chang, K.-Y. and Ghosh, J.: A unified model for probabilistic principal surfaces, *IEEE Trans. on Pattern Analysis and Machine Intelligence,* PAMI-23 (2001) 22–41.

8 Condon, E., Golden, B, Lele, S., Raghavan, S, Wasil, E.: A visualization model based on adjacency data. *Decision Support systems*, 33 (2002) 349–362.

9 Cox, T. F., & Cox, M. A. A.: *Multidimensional Scaling*, Chapman & Hall (1994).

10 De Ridder, D. and Duin R.P.W.: Sammon mapping using neural networks: a comparison. *Pattern Recognition Letters* 18 (1997) 1307–1316

11 Freeman, R. and Yin, H.: Self-organising maps for hierarchical tree view document clustering using contextual information. *LNCS-2412*. Yin, et al (Eds). (2002) 123–128

12 Girolami, M.: *Self-Organising Neural Networks: Independent Component Analysis and Blind Source Separation*. Springer (1999).

13 Haritopoulos, M., Yin, H., and Allinson, N.M.: Image denoising using self-organising map –based nonlinear independent component analysis. *Neural Networks* 15 (2002) 1085–1098.

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

15 Herrmann, M. and Yang, H. H.: Perspectives and limitations of self-organising maps in blind separation of source signals. *Proc. ICONIP'96* (1996) 1211–1216.

16 Honkela, T., Kaski, S., Lagus, K., and Kohonen, T.: WEBSOM-self-organizing maps of document collections. *Proceedings of Workshop on Self-Organizing Maps (WSOM'97),* 310–315.

17 Hyvärinen, A, Karhunen, J. and Oja, E.: *Independent Component Analysis*. John Wiley & Sons, Inc. (2001).

18 Hyvärinen, A. and Pajunen, P.: Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12 (1999) 429–439.

19 Kaban, A. and Girolami, M.: A combined latent trait class and trait model for the analysis and visualisation of discrete data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-23 (2001) 859–872.
20 Karhunen, J., and Joutsensalo, J.: Generalisation of principal component analysis, optimisation problems, and neural networks. *Neural Networks*, 8 (1995) 549–562.
21 Karhunen, J. and Malaroiu, S.: Local independent component analysis using clusternig. *Proc. 1st Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)* (1999) 43–48.
22 Kaski, S., and Kohonen, T.: Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. *Neural Networks in Financial Engineering*, Apostolos-Paul N. Refenes, Yaser Abu-Mostafa, John Moody, and Andreas Weigend (Eds.) World Scientific, (1996) 498–507.
23 Kegl, B., Krzyzak, A., Linder, T., and Zeger, K.: A polygonal line algorithm for constructing principal curves. *Neural Information Processing Systems (NIPS'98)*, 11 (1998) 501–507.
24 Kohonen, T.: Self-organised formation of topologically correct feature map. *Biological Cybernetics*, 43 (1982) 56–69.
25 Kohonen, T.: *Self-Organising Maps,* Springer: Berlin (Second edition). (1995).
26 Kraaijveld, M.A., Mao, J., and Jain, A.K.: A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Trans. Neural Networks*, 6 (1995) 548–559.
27 Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. *AICHE Journal*, 37 (1991) 233–243.
28 LeBlanc, M., and Tibshirani, R. J.: Adaptive principal surfaces. *J. Amer. Statist. Assoc.* 89 (1994) 53–64.
29 Lee, R.C.T., Slagle, J.R., and Blum, H.: A triangulation method for the sequential mapping of points from n-space to two-space. *IEEE Trans. on Computers*, 27 (1977) 288–292.
30 Lee, T.-W.: *Independent Component Analysis: Theory and Applications*. Kluwer Academic (1998).
31 Lee, T.-W., Koehler, B.-U. and Orglmeister, R.: Blind source separation of nonlinear mixing models. *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP'97)* (1997) 406–415.
32 Lee, T.-W., Lewicki, M.-S. and Sejnowski, T.-J.: Unsupervised Classification with Non-Gaussian Mixture Models using ICA. *Advances in Neural Information Processing Systems*, 11 (1999) 508–514.
33 Lowe, D and Tipping, M.E.: Feed-forward neural networks and topographic mappings for exploratory data analysis. *Neural Computing and Applications*. 4 (1996) 83–95.
34 Malthouse, E. C.: Limitations of nonlinear PCA as performed with generic neural networks. *IEEE Trans. Neural Networks*, 9 (1998) 165–173.
35 Mao, J., and Jain, A. K.: Artificial Neural Networks for Feature Extraction and Multivariate Data Projection. *IEEE Trans. on Neural Networks*, 6 (1995). 296–317.
36 Mulier, F., and Cherkassky, V.: Self-organisation as an iterative kernel smoothing process. *Neural Computation*, 7 (1995) 1165–1177.
37 Oja, E.: Neural networks, principal components, and subspaces. *Int. Journal of Neural Systems*, 1 (1989), 61–68
38 Oja, E.: PCA, ICA, and nonlinear Hebbian learning. *Proc. Int. Conf. on Artificial Neural Networks (ICANN'95)* 89–94.
39 Pajunen, P. and Karhunen, J.: A maximum likelihood approach to nonlinear blind source separation. *Proc. Int. Conf. on Artificial Neural Networks (ICANN'97)* 541–546.

40 Pajunen, P., Hyvärinen, A. and Karhunen, J.: Nonlinear blind source separation by self-organising maps. *Proc. ICONIP'96,* 1207–1210.
41 Ripley, B. D.: *Pattern Recognition and Neural Networks*, Cambridge University Press: Cambridge, (1996).
42 Ritter, H., Martinetz, T., and Schulten, K.: *Neural Computation and Self-organising Maps: An Introduction*. Addison-Wesley Publishing Company (1992).
43 Roweis, S. T., and Saul, L. K.: Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (2000) 2323–2326.
44 Sanger, T. D.: Optimal unsupervised learning in a single-layer linear feedforward network. *Neural Networks*, 2 (1991) 459–473.
45 Sammon, J. W.: A nonlinear mapping for data structure analysis. *IEEE Trans. on Computer*, 18 (1969) 401–409.
46 Schölkopf, B., Smola, A., & Müller, K. R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 (1998) 1299–1319.
47 Taleb, A. and Jutten, C.: Source separation in postnonlinear mixtures. *IEEE Trans. on Signal Processing*, 47 (1999) 2807–2820.
48 Tan, Y., Wang, J. and Zurada, J. M.: Nonlinear blind source separation using a radial basis function network. *IEEE Trans. on Neural Networks*, 12 (2001) 124–134.
49 Tenenbaum, J. B., de Silva, V., & Langford, J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (2000) 2319–2323.
50 Tibshirani, R.: Principal curves revisited. *Statistics and Computation*, 2 (1992) 183–190.
51 Törönen, P., Kolehmainen, K., Wong, G., Castrén, E.: Analysis of gene expression data using self-organising maps. *FEBS Letters*, 451 (1999) 142–146.
52 Ultsch, A.: Self-organising neural networks for visualisation and classification. *Information and Classification*, O. Opitz, B. Lausen and R. Klar (Eds.) (1993) 864–867.
53 Xu, L, Cheung, C. C. and Amari, S.-I.: Learned parametric mixture based ICA algorithm. *Neurocomputing*, 22 (1998) pp. 69–80.
54 Yin, H.: Visualisation induced SOM (ViSOM). In: *Advances in Self-Organising Maps (Proc. WSOM'01)*, N. Allinson, H. Yin, L. Allinson, & J. Slack (Eds.), Springer, 81–88.
55 Yin, H.: ViSOM-A novel method for multivariate data projection and structure visualisation. *IEEE Trans. on Neural Networks*, 13 (2002) 237–243.
56 Yin, H.: Data visualisation and manifold mapping using the ViSOM. *Neural Networks*, 15 (2002) 1005–1016.
57 Yin, H., and Allinson, N. M.: Interpolating self-organising map (iSOM). *Electronics Letters* 35 (1999) 1649–1650.
58 Yin, H., and Allinson, N. M.: Bayesian self-organising map for Gaussian mixtures. *IEE Proc. –Vis. Image Signal Processing,* 148 (2001) 234–240.