Chapter 2

Established Techniques for Topographic Mapping

2.1 Introduction

This chapter considers three particular established schemes for the generation of mappings that preserve some notion of topography or geometric structure — the *Kohonen Self-Organising Feature Map* (Section 2.2), the *Sammon Mapping* (2.3) and the statistical field of *Multidimensional Scaling* (2.5). Each of these approaches is individually described, comparisons between them are drawn and respective advantages and disadvantages outlined.

2.2 The Kohonen Self-Organising Feature Map

The archetypal topographic neural network is Kohonen's self-organising feature map (often simply referred to as the 'Kohonen Map' or abbreviated to 'SOFM') [Kohonen 1982; Kohonen 1990; Kohonen 1995]. The motivation for Kohonen's model is neuro-biological and was developed as an abstraction of earlier work in the field of ordered neural connections by Willshaw and von der Malsburg [1976].

The SOFM can be viewed as a neural network comprising a set of input neurons and a set of output neurons, each of which is connected by a weight vector $\mathbf{w}$, in the standard manner, to the input. However, in contrast to the standard single-layer model — the simple perceptron — there is an inherent additional structure within the output layer. These neurons may be considered to form a fixed *lattice*, usually one- or two-dimensional, with associated lateral connectivity in addition to the connection to the input layer. In the most common two-dimensional case, the output of the SOFM network is a 'sheet' of interconnected neurons in a rectangular or hexagonal configuration. Such an architecture is illustrated in figure 2.1.

When successfully trained, such a network will exhibit the property that adjacent neurons in this lattice structure respond to similar (nearby) input vectors, or features, and the map is then said to be *topologically ordered*. This ordered mode of neural response has been observed on the neocortex in the brains of higher animals, notably the auditory [Suga and O'Neill 1979], the visual [Blasdel and Salama 1986] and somatosensory [Kaas et al. 1979] cortices. For example, on the human auditory cortex, there is a near-logarithmic frequency ordering of responsive cells — this is the so-called *tonotopic map* — such that nearby neurons on the cortex respond to sounds of a similar pitch. This, and

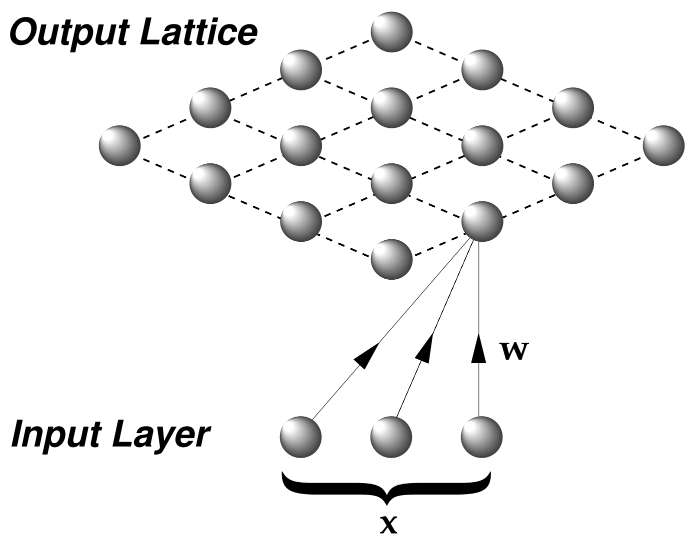

Figure 2.1: A schematic of the architecture of a two-dimensional output layer Kohonen network. For clarity, only the weight connections from the input to a single neuron are shown.

other mappings within the brain, involve a vast number of similar cortical connections (estimated at around 10^{13}), and this almost certainly precludes the possibility that this topological ordering is genetically determined, thus suggesting that these properties evolve during brain development according to some alternative systematic process. The learning procedure for the Kohonen SOFM is a generalised abstraction of such a potential mechanism, and the model has been successfully applied across a considerable variety of distinctly non-biological domains. Good examples include speech recognition, image processing, interpretation of EEG traces and robot arm control, and these, and other applications, are comprehensively reviewed (with references) in [Kohonen 1995]. In addition, there is a large on-line biography ($>$ 1630 references) available concerning theory and applications of the self-organising map [Anonymous 1996].

To construct the SOFM, consider a network as outlined above with inputs from some p-dimensional space connected to a q-dimensional output lattice of K neurons with associated weight vectors $\mathbf{w}_i$, each of which effectively defines a point in the input space. The Kohonen algorithm is thus:

❶ Choose the dimension, size and topology of the map according to the prior knowledge of the problem. Some preprocessing of the data may also be necessary as the map is sensitive to scaling of the input features.

❷ At time step $t = 0$ initialise all the weight vectors $\mathbf{w}_i$ to random values.

❸ Present an input pattern $\mathbf{x}_t$ to the network, drawn according to the input distribution defined by the probability density function $f(\mathbf{x})$.

❹ Determine the "winning" neuron, $v(\mathbf{x}_t)$, whose weight vector $\mathbf{w}_i$ is closest to the input point $\mathbf{x}_t$. That is

$$v(\mathbf{x}_t) = \underset{i}{\mathbf{argmin}} \parallel \mathbf{w}_i - \mathbf{x}_t \parallel$$

❺ Adjust the weight vector of the winning neuron, and those of its neighbours, in a direction towards the input vector. That is

$$\mathbf{w}_i = \mathbf{w}_i + \eta(t)\Lambda[i, v(\mathbf{x}_t)](\mathbf{x}_t - \mathbf{w}_i),$$

where $\eta(t)$ is a learning-rate parameter. The function $\Lambda[i, v(\mathbf{x}_t)]$ is the *neighbourhood* function, and is described in detail below.

❻ Repeat from step ❸ until the map has stabilised.

The key to the topographic nature of the mapping is the neighbourhood function, $\Lambda(i, v(\mathbf{x}))$. This is some function defined over the output lattice space which is generally non-negative and decreases with the distance (in the lattice space) between the winning neuron $v(\mathbf{x})$ and any other neuron i. This implies that the weight vector of the winning neuron receives maximum perturbation, while the corresponding vectors of more distant neurons are adjusted to a lesser extent. Popular choices for this function are the Gaussian, $\Lambda(r) = \exp(-r^2/2\sigma^2)$ — where r is the distance from the winning neuron to another neuron — and the 'bubble', which is simply a constant value over a fixed neighbourhood width σ. The parameter σ therefore controls the degree of weight adjustment with distance, a property that is sometimes referred to as the "stiffness" of the lattice.

Thus, when an input pattern is presented to the network, the nearest, winning, neuron will be moved in the direction of that input vector along with its neighbours by an amount that decreases with their distance (within the *lattice*) from the winning neuron. In this way, the weights for nearby neurons will converge to the same region of input space, thus exhibiting the characteristic topological ordering. The 'width' of this neighbourhood function, the parameter σ, is t-dependent. It is usually set to a relatively high value (as much as half the lattice width or greater) at initialisation, and decreases with time. This allows the coarse global structure of the map to be formed in the early stages of training, while the local structure is fine-tuned later. The learning-rate $\eta(t)$ also varies with time, decreasing monotonically to some arbitrarily small value when the map is "frozen".

Three example plots taken during the training of a SOFM are illustrated in Figure 2.2 below. These show the evolution of the output lattice for input data sampled uniformly at random from within a 2-dimensional unit square. The weight vectors $\mathbf{w}_i$ are plotted in the input space, and those corresponding to adjacent nodes in the neuron lattice are shown connected. Note that these connections are not explicit within the SOFM architecture; it is the action of the neighbourhood function that implicitly inter-connects all the output layer neurons to some degree.

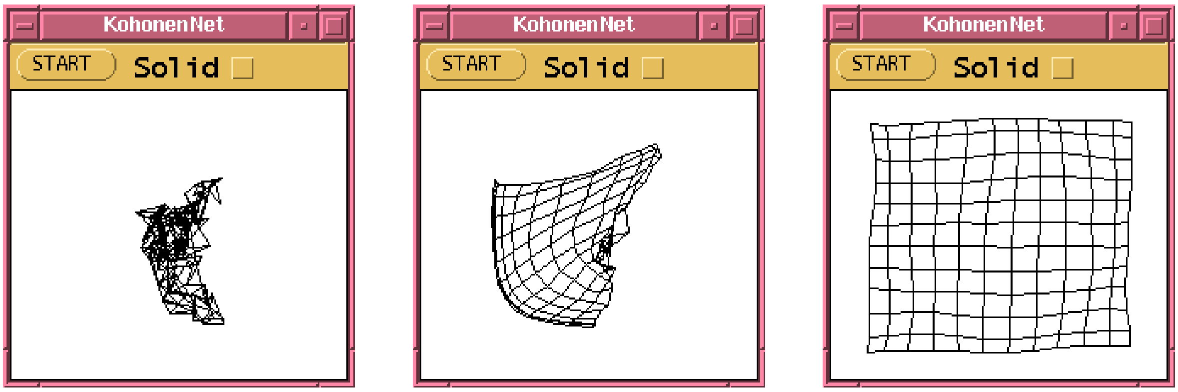

Figure 2.2: A 2-D Kohonen map of data sampled uniformly at random from the unit square. The final map is after 5,000 time steps.

For a neighbourhood function of zero width, there is no topological ordering in the map and the algorithm becomes equivalent to a *vector quantisation* (VQ) scheme (specifically, the LBG algorithm of Linde, Buzo, and Gray [1980]), where the weights $\mathbf{w}_i$ are analogous to the codebook vectors of VQ. It is also, therefore, closely related to the *k-means* technique for data clustering [MacQueen 1967].

Ideally, it would be convenient if the *density* of the distribution of the $\mathbf{w}_i$ in the input space were directly representative of the input probability density $f(\mathbf{x})$. This, however, is not the case. An exact result has been derived only for the single-dimensional feature map which reveals that the density of the $\mathbf{w}_i$, $m(\mathbf{w})$, is in fact given by [Ritter and Schulten 1986; Ritter 1991]:

$$m(\mathbf{w}) \propto f(\mathbf{x})^{\beta}, \quad \text{with} \tag{2.1}$$

$$\beta = \frac{2}{3} - \frac{1}{3[\sigma^2 + (\sigma+1)^2]}, \tag{2.2}$$

where σ is the number of neurons to each side of the winning neuron that are adjusted at each training step. For $\sigma = 0$, this is equivalent to VQ, and $m(\mathbf{w}) \propto f(\mathbf{x})^{1/3}$.

In this case, and in general for higher dimensions, the Kohonen SOFM over-emphasises regions of low input density at the cost of under-emphasising those of high density. An illustration of this effect may be seen later in Section 2.4.

One important feature of the SOFM is that it exists only at the algorithmic level. It has been shown [Erwin, Obermayer, and Schulten 1992] that the above procedural description of the Kohonen Map cannot be interpreted as minimising a single energy (or error) function. This implies that there is no direct measure of "quality" of a map, although several indirect alternatives have been proposed [Bauer and Pawelzik 1992; Bezdek and Pal 1995; Goodhill, Finch, and Sejnowski 1995].

There have been several extensions made to the basic SOFM model since its introduction. For application to classification problems, there are the *Learning Vector Quantisation* (LVQ) schemes [Kohonen 1990], where sets of weight vectors are allocated exclusively to a single class and the learning algorithm adjusted such that inter-class decision boundaries are emphasised. There have also been variants of the map proposed which permit arbitrary and dynamic output layer topology, such as the *neural gas* of Martinetz and Schulten [1991] and the *growing cell structures* of Fritzke [1994]. Such schemes permit a better match between the network topology and that of the data distribution, but considerably complicate the generation of convenient visualisations, such as that illustrated for the standard SOFM in Section 2.4. This restriction makes these approaches less suitable for data analysis, and they will not be considered further in this thesis.

2.3 The Sammon Mapping

If the definition of a topographic mapping is to be understood as implying a retention of global metric relationships, then the *Sammon Mapping* [Sammon 1969], sometimes referred to as the *Non-Linear Mapping* or NLM, is the most intuitive basis for its definition. In contrast to the Kohonen mapping, the Sammon mapping may be determined by the optimisation of an error, or 'STRESS', measure which attempts to preserve all inter-point distances under the projection. The Sammon STRESS is defined as

$$E_{ss} = \frac{1}{\sum_i \sum_{j<i} d^*_{ij}} \sum_i \sum_{j<i} \frac{[d^*_{ij} - d_{ij}]^2}{d^*_{ij}}, \tag{2.3}$$

where d^*_{ij} is the distance $\| \mathbf{x}_i - \mathbf{x}_j \|$ between points i and j in the input space $\mathbb{R}^p$, and d_{ij} is the distance $\| \mathbf{y}_i - \mathbf{y}_j \|$ between their images in the map, or feature, space $\mathbb{R}^q$. These distance measures are generally Euclidean but need not strictly be so.

The $[d^*_{ij} - d_{ij}]^2$ term is clearly a measure of the deviation between corresponding distances, and the Sammon STRESS thus represents an optimal, in the least-squares sense, matching of inter-point distances in the input and map spaces. The first fractional term in the expression is a normalising constant which reduces the sensitivity of the measure to the number of input points and their scaling. The inclusion of the d^*_{ij} term in the denominator of the sum serves to moderate the domination of errors in large distances over those in smaller distances, and renders the overall measure dimension-less. The inclusion of this term is not justified in the original paper, and has the effect of making the mapping more sensitive to absolute (though not proportional) errors in local distances.

Given this STRESS measure E_{ss}, it is straightforward to differentiate with respect to the mapped coordinates $\mathbf{y}_i$ and optimise the map using standard error-minimisation methods. Setting the constant $c = \sum_i \sum_{j<i} d^*_{ij}$ for simplification gives

$$\frac{\partial E_{ss}}{\partial \mathbf{y}_i} = \frac{-2}{c} \sum_j \frac{(d^*_{ij} - d_{ij})}{d^*_{ij} d_{ij}} (\mathbf{y}_i - \mathbf{y}_j). \tag{2.4}$$

All the points $\mathbf{y}_i$ in the configuration can thus be simultaneously iteratively adjusted to minimise the error. It should be noted that each partial derivative requires N cycles of computation and therefore to calculate the entire set of derivatives will require a double sum over the data. (In fact, $N(N-1)/2$ loops.) Sammon used a simple gradient-descent technique in his original paper, but less naive methods may be employed, and a *conjugate-gradient* routine [Press, Teukolsky, Vetterling, and Flannery 1992] was found to be considerably more effective.

The Sammon mapping originated in the engineering field and was designed as a computational tool for data structure analysis and for visualisation, and indeed, its use is still popular in many domains — Domine et al. [1993] provide a good review of applications in the field of chemometrics. Feature space dimension, q, is thus naturally chosen as either 2 or 3. Because of the metric nature of the map, clusters of data points tend to be retained under the projection and are manifest in the feature space. In addition to this *local* clustering structure, the inter-cluster *global* relationships are also preserved to some extent. Sammon emphasised this latter feature in the paper, giving several illustrative examples where a linear projection onto the first two principal axes (the orthogonal axes that maximise the variance under projection) confused multiple distinct clusters in contrast to the Sammon mapping which maintained their separation.

While minimisation of E_{ss} implies preservation of the input geometry, the extent to which the integrity of the structure of the input space can be retained is dependent both upon the intrinsic dimensionality of the data, and also upon its topology. In the process of dimension reduction, some information, in all but the most degenerate cases, will be lost, and furthermore, apparent structure may be elucidated which is truly artefactual in nature. A minor example of such structure will be seen for spherical data in the next chapter, and reference to a more controversial case will be made shortly in discussion of multidimensional scaling. Some investigation of artefactual structure was undertaken by Dzwinel [1994]. One particular illustration was given for data generated uniformly at random from within a 100-dimensional hypercube, which resulted in a circular configuration when mapped down to two dimensions. The cause of this particular configuration was actually explained by the author, with reference to the "curse-of-dimensionality", but this and other such projections may often appear inconsistent to the human observer because "our intuitive notions of low dimensions don't carry over well to high dimensions" [Friedman 1995].

Despite the simple, intuitive appeal of the Sammon Mapping, there are, however, some significant disadvantages and limitations to its application.

① The mapping is generated iteratively and has been observed to be particularly prone to sub-optimal local minima.

② The computational requirements scale with the square of the number of data points, making its application intractable for large data sets.

③ There is no method to determine the dimensionality of the feature space *a priori*.

④ The map is generated as a 'look-up table' — that is, there is no way to project new data without re-generating the entire map with the new data points included.

Sammon himself appreciated the restriction posed by item ② above, conceding that with the computing facilities available at that time, a practical upper limit of 200 data points was imposed. To partially overcome this, he proposed applying some *a priori* clustering process to extract prototypes, and then mapping these with the algorithm. This, and other approaches to the computational problem, will be considered in Chapter 7, with an investigation of problem ①, local minima, in Chapter 6. A considerable part of this thesis will be concerned with approaches to the problem posed by item ④, and this will be considered in more detail in the following chapters.

2.4 Comparison of the Kohonen SOFM and the Sammon Mapping

It has already been stated that, unlike the Sammon mapping, the generation of a Kohonen SOFM cannot strictly be interpreted as the minimisation of a single energy or cost function [Erwin, Obermayer, and Schulten 1992]. Aside from this, there is a more fundamental underlying difference between the two methods. It is the mechanism of the local neighbourhood function in the Kohonen map that affords the topographic nature of the scheme. However, there is no explicit retention of global structure, and indeed, the emphasis of the algorithm is to model the *density* of the underlying input data distribution. This contrast between the two techniques may be illustrated by the following simplistic example.

Both the SOFM and Sammon's algorithm are applied to the mapping of a synthetic dataset comprising three clusters in three dimensions. The clusters, C_1, C_2 and C_3, are centred at $(0,0,0)$, $(1,-1,0)$ and $(4,5,0)$, and contain 50, 100 and 50 points respectively. Each cluster is dispersed uniformly at random inside a cube centred at each point, with the size of each edge of the cube for C_3 being double that of C_1 and C_2. This distribution of data is illustrated via two orthogonal projections in figure 2.3. A (12×10) Kohonen Map of this data and the corresponding Sammon mapping are shown in figure 2.4. For comparison, the first two principal components of the data are plotted in figure 2.5.

This particular distribution of data was chosen deliberately to emphasise the differences in the methods. As illustrated in figure 2.4, the Sammon mapping offers a good representation of the original topography. The variation of inter-cluster separation is still clear, and the increased dispersion of C_3 is also evident. The Kohonen SOFM has retained the local topology, but because it is a density-driven approach, fails to capture both the global relationships between the clusters and the local dispersion of C_3. Underlining this behaviour, the concentration of neurons in the region of class C_2, which contained twice the point density of the other classes, is also evident. The number of nodes activated for each of the three classes is 23, 40 and 27 respectively, which indicates that the map has over-represented the lower density clusters. That the cluster C_3 is significantly larger than C_1 and C_2 is not evident, and neither is its greater distance from those clusters.

In this simple case, a principal component projection is apparently adequate for retaining the topography (although close inspection will reveal better dispersion within the three clusters in the case of the Sammon Mapping). For real, higher-dimensional, datasets, this linear technique is generally limited in its application, as will be illustrated in Chapter 4.

An additional phenomenon inherent in the SOFM is the introduction of some topographic distortion due to the fixed topology of the lattice of output neurons. As asserted by Li, Gasteiger, and Zupan [1993], "global topology distortions are ... inevitable" in all but the most trivial situations. This effect is a result of mismatch between the topology of the lattice and that of the input data. This conclusion is also confirmed by Bezdek and Pal [1995] who claim that "the Sammon method preserves metric relationships much better than [the SOFM]." This assertion is a result of assessing the alternative mappings according to a measure of *metric topology preservation*, derived from Spearman's rank coefficient. With respect to this criterion, the Sammon mapping scored higher for all datasets tested.

The distortive aspect may be demonstrated by the example in figures 2.6 and 2.7. This illustrates a (12×12) 2D-sheet mapping of data points lying on three concentric 3-dimensional spheres, with radii 0,1 and 2 units respectively. Fifty points were distributed at random over each of the spheres and a small amount of Gaussian random noise was added, making the centre sphere effectively a cluster. The diagram in figure 2.6 shows the map, with its inevitable discontinuities, and below, in figure 2.7, is an illustration of the form of the sheet embedded in the input space — the 'frustration' in the lattice is clearly visible in this latter diagram. When such mismatch occurs, it may also induce poor performance from a clustering point of view. Such degradation, in comparison with the standard 'k-means' procedure, has been observed by Balakrishnan, Cooper, Jacob, and Lewis [1994].

Regarding these criticisms it should be noted that Kohonen's SOFM was developed as an analogue of observed neuro-biological behaviour, rather than being explicitly motivated by the criterion of faithful preservation of universal topography. In addition, in stark contrast to Sammon's technique, it has the attractive feature of good computational behaviour. It is this tractability for sizable datasets which makes the Kohonen SOFM a popular topographic mapping tool. However, on the basis of the discussion in this section, for applications in data analysis and visualisation, the Sammon mapping should be preferred for smaller datasets.

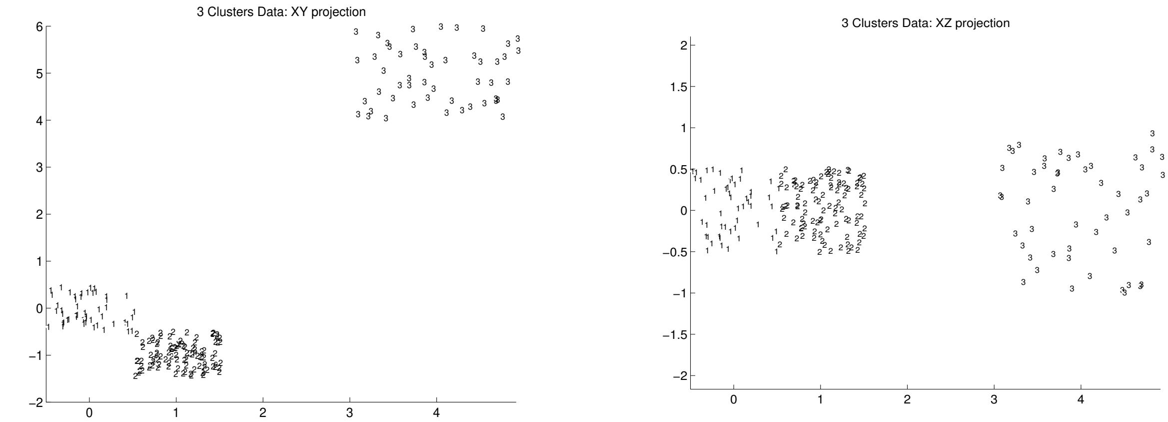

Figure 2.3: Synthetic data distributed in 3 clusters in 3-dimensional space.

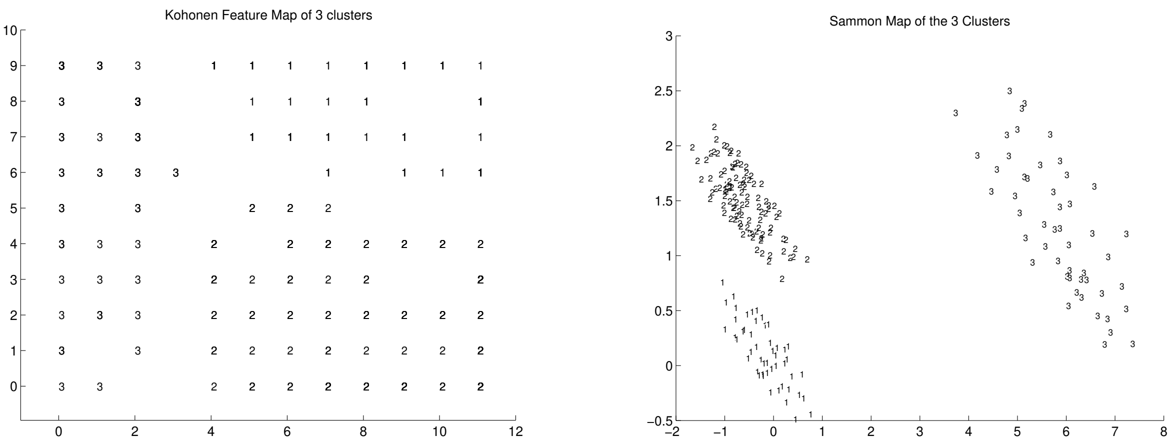

Figure 2.4: Kohonen and Sammon Mappings of the 3 clusters.

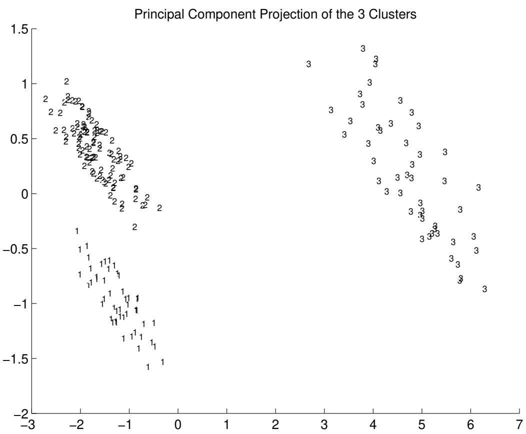

Figure 2.5: Projection onto the first two principal axes of the 3 clusters.

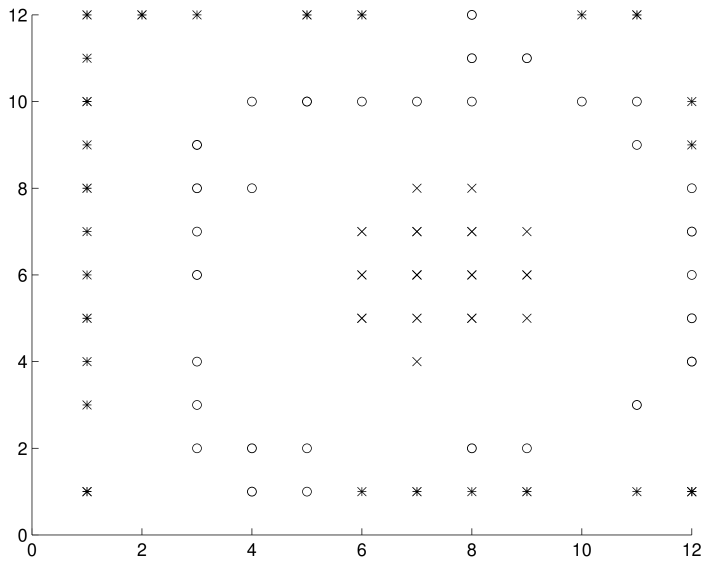

Figure 2.6: A Kohonen Mapping of data on 3 concentric spheres.

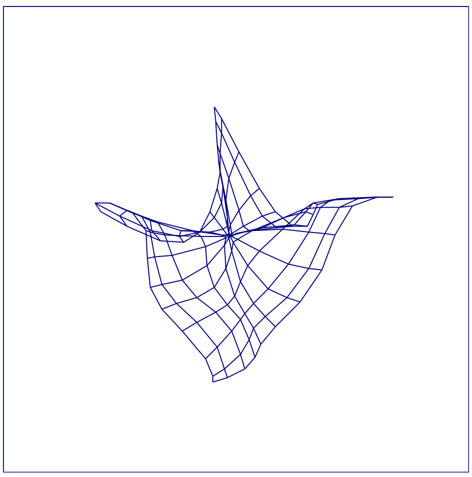

Figure 2.7: The Kohonen lattice embedded in the original space.

2.5 Multidimensional Scaling

2.5.1 The Underlying Principle

Multidimensional Scaling (MDS) is described by Davison [1983] as

> ...a set of multivariate statistical methods for estimating the parameters in, and assessing the fit of, various spatial distance models for proximity data.

This definition is a relatively narrow one, and some authors (e.g. Carroll and Arabie [1980]) accept a broader view and include other methods for modelling multivariate proximity data (such as factor analysis or cluster analysis) within the scope of MDS. However, it is the topographic properties of MDS that are relevant in this context, so the spatial distance definition is the most appropriate here.

The raw data to which MDS is applied is *proximity data*. This is generally in the form of a square symmetric ($N \times N$) matrix, where each row and column enumerates a set of objects and the elements of the matrix are measures of the relative proximity of those respective objects. In this context, proximity may refer to either *similarity* or *dissimilarity* of objects. It is then the purpose of MDS techniques to represent the structure of the proximity matrix in a more simple and perspicuous geometrical model. The classic example is that, given a matrix of road-distances (which can be considered analogous to dissimilarities) between cities, the data can be modelled by a two-dimensional map (e.g. see Krzanowski and Marriott [1994], pp113). In this instance, the scaling procedure greatly facilitates visualisation of the data and eliminates the redundancy in the description.

For the case of the road-distances, the geometric interpretation is intuitive and clearly valid. Typically, however, the proximity data processed by MDS models will have been gathered in a more subjective manner, often by means of psychological experiment where human subjects are asked to assess the likeness, or *similarity*, of each pair of objects, or *stimuli*. The fundamental assumption underlying the application of MDS in these contexts is that these empirical observations can be meaningfully fitted to a set of points in some metric space, where the distance between the points representing each pair of stimuli corresponds to their perceived *dissimilarity*. (The measure of 'dissimilarity' may be simply derived from that of 'similarity', for example, by subtracting from a constant.) This basic principle was originally proposed by Richardson [1938]. Given this assumption, it is then hoped that such a fitted configuration will aid visualisation of the data and also provide insight to the processes that generated it. These techniques have been successfully applied in a variety of fields — the behavioural and social sciences, psychology, acoustics, olfactory analysis, education and industrial relations are examples. A comprehensive list of many such applications is given by Davison [1983]. MDS remains a very popular tool, with a search of citation indices revealing relevant annual publications in the hundreds. A prominent, recent, and controversial example of the application of MDS techniques is in the study of connectivity of regions in the visual cortex of the macaque monkey [Young 1992]. This has provoked some significant debate over the validity of the structure inferred from such a model [Goodhill, Simmen, and Willshaw 1995], as to whether it is artefactual or truly representative of the underlying relationships in the data.

One particularly good illustration of MDS applied to psychological data concerns a study of colour vision. In this experiment, performed by Ekman [1954], participants estimated the similarity of all combinations of pairs of 14 different sample colours presented to them. An MDS technique was used to convert these similarity measurements into a configuration of points in two dimensions, where they were found to lie in a spectrally ordered manner on an annular, horseshoe, structure. This 'bending' of the colour line is a result of the phenomenon that many subjects (reasonably) perceive similarity between the two extreme ends of the spectrum — red and violet. The similarity data and the resulting mapping, which is clearly informative in this case, are given in figures 2.8 and 2.9 respectively.

$$
\begin{bmatrix}
1 & .86 & .42 & .42 & .18 & .06 & .07 & .04 & .02 & .07 & .09 & .12 & .13 & .16 \\
.86 & 1 & .50 & .44 & .22 & .09 & .07 & .07 & .02 & .04 & .07 & .11 & .13 & .14 \\
.42 & .50 & 1 & .81 & .47 & .17 & .10 & .08 & .02 & .01 & .02 & .01 & .05 & .03 \\
.42 & .44 & .81 & 1 & .54 & .25 & .10 & .09 & .02 & .01 & .00 & .01 & .02 & .04 \\
.18 & .22 & .47 & .54 & 1 & .61 & .31 & .26 & .07 & .02 & .02 & .01 & .02 & .00 \\
.06 & .09 & .17 & .25 & .61 & 1 & .62 & .45 & .14 & .08 & .02 & .02 & .02 & .01 \\
.07 & .07 & .10 & .10 & .31 & .62 & 1 & .73 & .22 & .14 & .05 & .02 & .02 & .00 \\
.04 & .07 & .08 & .09 & .26 & .45 & .73 & 1 & .33 & .19 & .04 & .03 & .02 & .02 \\
.02 & .02 & .02 & .02 & .07 & .14 & .22 & .33 & 1 & .58 & .37 & .27 & .20 & .23 \\
.07 & .04 & .01 & .01 & .02 & .08 & .14 & .19 & .58 & 1 & .74 & .50 & .41 & .28 \\
.09 & .07 & .02 & .00 & .02 & .02 & .05 & .04 & .37 & .74 & 1 & .76 & .62 & .55 \\
.12 & .11 & .01 & .01 & .01 & .02 & .02 & .03 & .27 & .50 & .76 & 1 & .85 & .68 \\
.13 & .13 & .05 & .02 & .02 & .02 & .02 & .02 & .20 & .41 & .62 & .85 & 1 & .76 \\
.16 & .14 & .03 & .04 & .00 & .01 & .00 & .02 & .23 & .28 & .55 & .68 & .76 & 1
\end{bmatrix}
$$

Figure 2.8: The proximity matrix for Ekman's colour data. Each value is a normalised, averaged, measure of observed similarity between 14 distinct sample colours.

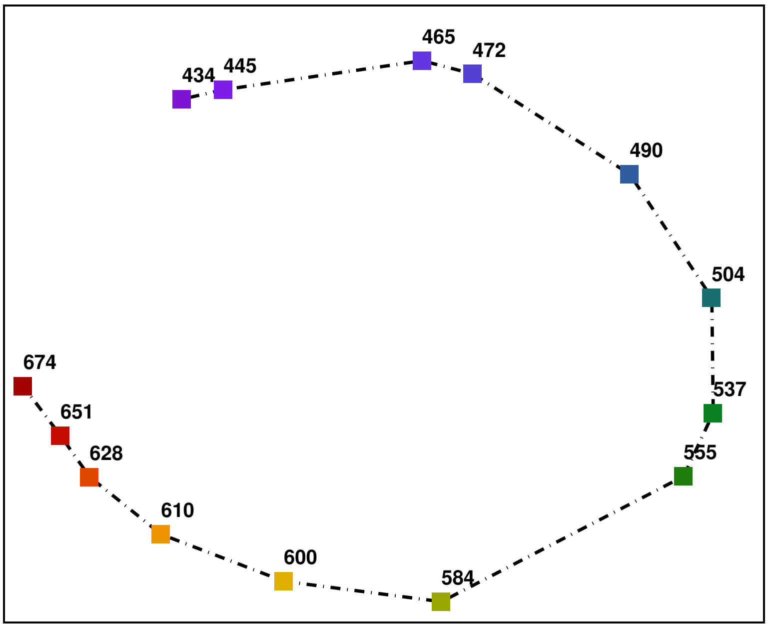

Figure 2.9: The resultant map, with wavelength shown for each sample, for Ekman's colour data.

2.5.2 Scaling Algorithms

The measured dissimilarity between a pair of objects (i, j), known as a *stimulus pair*, can be formalised as the variable δ_{ij}, which is an element of the $(N \times N)$ dissimilarity matrix $\mathbf{\Delta}$. It is the purpose of MDS to turn this data into a $(N \times q)$ configuration matrix $\mathbf{Y}$. In general, and in common with the Sammon Mapping, the dimension of the feature space q is unknown *a priori*.

The configuration of points $\mathbf{y}_i : i \in \{1 \ldots N\}$ must be determined such that the values δ_{ij} match some *distance function*, $d(i, j)$, defined over all possible pairs of points $(\mathbf{y}_i, \mathbf{y}_j)$. For $d(i, j)$ to be a distance function, the following four axioms must hold:

$$d(a, b) \geq 0, \tag{2.5}$$
$$d(a, a) = 0, \tag{2.6}$$
$$d(a, b) = d(b, a), \tag{2.7}$$
$$d(a, b) + d(b, c) \geq d(a, c). \tag{2.8}$$

In a psychological context (e.g. consider the colour data), these first three axioms, (2.5)-(2.7) appear intuitively reasonable, although there is no apparent support or contradiction for (2.8), known as the triangular inequality axiom. Whilst much experimental work corroborates the distance model for psychological data, the results of some tests appear to violate some of the axioms (e.g. Rothkopf 1957). However, this is just one aspect of the application of MDS — in other contexts these contradictions are not manifest.

Usually, the metric employed in the configuration space is the standard Euclidean, although general Minkowski distances have been used in particular applications.

There are two main branches of MDS models — the *metric* and the *nonmetric* methods. In the former, the dissimilarities should correspond as closely as possible to the inter-point distances in the generated configuration. In the latter scheme this constraint is relaxed, with psychological justification, such that the *ordering* of the dissimilarities should correspond to the *ordering* of the distances. The metric techniques, originally developed by Torgerson [1952] as *classical* MDS, have been superseded by the more flexible and effective nonmetric models. The following two subsections cover these methods in more detail.

2.5.3 Classical Multidimensional Scaling (CMDS)

One of the first MDS algorithms was proposed by Torgerson [1952, 1958]. By definition as a metric method, it assumes the identity relationship between distance in the feature space and corresponding object dissimilarity:

$$\delta_{ij} = d(i, j). \tag{2.9}$$

As such, it requires somewhat restrictive assumptions, and is seldom used in its original form, although many more developed algorithms build on it. It does, however, have the advantage of an analytical derivation.

The CMDS procedure is as follows:

❶ From the dissimilarity matrix $\mathbf{\Delta}$, generate the *double-centred inner product matrix* $\mathbf{B}^*$, given by:

$$\mathbf{B}^* = -\frac{1}{2}\mathbf{H}\mathbf{\Delta}_2\mathbf{H}, \tag{2.10}$$

with $\mathbf{\Delta}_2$ the matrix whose elements are the square of those of $\mathbf{\Delta}$. That is, $\mathbf{\Delta}_2 = \{\delta_{ij}^2\}$. The matrix $\mathbf{H}$ is the *centring* matrix, given by $\mathbf{I} - \mathbf{1}/N$, where $\mathbf{1}$ is the square matrix whose elements are all 1.

❷ Factorise $\mathbf{B}^*$ into:

$$\mathbf{B}^* = \mathbf{U\Lambda U}^{\mathsf{T}}, \tag{2.11}$$

$$= \mathbf{YY}^{\mathsf{T}}. \tag{2.12}$$

The matrix $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{B}^*$, with $\mathbf{U}$ the corresponding matrix of eigenvectors.

❸ The matrix $\mathbf{Y} = \mathbf{U\Lambda}^{1/2}$ is the configuration of points in $p = N$ dimensions that satisfies exactly the dissimilarity measures specified in $\mathbf{\Delta}$. To reduce to q dimensions, select the q columns of $\mathbf{U}$ corresponding to the q largest eigenvalues, giving an $(N \times q)$ data matrix $\mathbf{Y}_q$.

There are several points to be noted about the CMDS procedure:

- The points $\mathbf{Y}$ are centred at the origin, so $\sum_i^N \mathbf{y}_i = \mathbf{0}$.
- Calculation of $\mathbf{\Lambda}^{1/2}$ requires that $\mathbf{B}^*$ is positive semi-definite. Techniques adopted for dealing with the problem of negative eigenvalues are typically heuristic. One is to ignore the small, negative eigenvalues. Another is the *trace criterion*, where the sum of the discarded negative eigenvalues should equal the sum of the positive discards, so the sum of the remaining eigenvalues is still equal to the trace of the matrix. A large negative eigenvalue is nevertheless a major problem. Mardia [1978] proposed "goodness of fit" measures for such non-Euclidean data.
- If $\mathbf{B}^*$ is positive semi-definite, then $\mathbf{\Delta}$ is a Euclidean distance matrix. That is, the dissimilarities δ_{ij} correspond exactly to the Euclidean distances between a set of points embedded in at most $(N-1)$ dimensions.
- If $\mathbf{B}^*$ is positive semi-definite, then the points $\mathbf{y}_i$ are referenced to their principal axes. (That is, $\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$ is diagonal). Furthermore, if the elements of $\mathbf{\Delta}$ are the inter-point Euclidean distances of a given set of data points, then the CMDS solution in q dimensions is identical to a projection onto the first q principal axes of the data. Indeed, CMDS is sometimes known, after Gower [1966], as *principal co-ordinates analysis*.
- For a *similarity* matrix $\mathbf{S}$, where $0 \leq s_{ij} \leq 1$ and $s_{ii} = 1$ (such as that given for the colour data in figure 2.8), then a corresponding dissimilarity matrix can be formed by $\delta_{ij} = \sqrt{(1 - s_{ij})}$. In that case, $\mathbf{\Delta}$ is a Euclidean distance matrix [Gower and Legendre 1986].

2.5.4 Nonmetric Multidimensional Scaling (NMDS)

In Nonmetric, or *ordinal*, Multidimensional Scaling (NMDS) the requirement that distances in the projected space optimally fit the dissimilarities is relaxed so that only the *ordering* of distances is retained. That is, the two most dissimilar stimuli should also be the two most distant points in the configuration and the second most dissimilar pair of stimuli be the second most distant pair of points etc. It is therefore not necessary for all corresponding pairs of distances and dissimilarities to be identical. In fact, the ordinal constraint implies that it is only necessary that the dissimilarities be some arbitrary monotonically increasing function of the distances.

Thus in contrast to equation (2.9), for nonmetric models the relationship between dissimilarity and spatial distance becomes

$$\delta_{ij} = f(d_{ij}) = f\left[\sum_k (x_{ik} - x_{jk})^2\right]^{1/2}, \tag{2.13}$$

where f is a monotone function such that

$$\forall i, j, i', j' : d_{ij} < d_{i'j'} \Rightarrow f(d_{ij}) < f(d_{i'j'}). \tag{2.14}$$

Note that the function f need never be known explicitly, although its form may be recovered after the scaling procedure.

As well as adhering to the first three Euclidean distance axioms, this concept preserves many intuitive psychological properties and in many cases permits the generation of more useful, lower-dimension, lower-STRESS mappings. Furthermore, in some experiments, only ordinal data is available, notably where human subjects are required to rank various stimuli in order of merit or preference.

In contrast to the classical method, these ordinal configurations are generated by minimisation of a particular cost function, or STRESS measure, and must be generated iteratively via some nonlinear optimisation procedure. Because of the ordinal constraint, generating a configuration is particularly computationally expensive due to the additional requirement of a *monotonic regression* step.

The first nonmetric scheme was proposed by Shepard [1962a, 1962b], in response to experimental evidence that in certain applications, observed dissimilarities were related to some nonlinear function of the spatial distances in a putative model (e.g., Shepard 1958). These methods were further and more formally developed by Kruskal and that work remains the basis of modern implementations. Kruskal [1964a] formalised the method by defining a measure of *goodness-of-fit*. The proposed MDS technique is thus to determine a point configuration $\mathbf{Y}$ that optimises this. A practical computer implementation of the algorithm is described in a companion paper [Kruskal 1964b]. To clarify the NMDS technique, consider the following description of Kruskal's procedure.

Given a set of experimentally obtained dissimilarity data δ_{ij} and a configuration of N points in q dimensions, the dissimilarities can be ranked according to their magnitude

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \ldots < \delta_{i_M j_M} \tag{2.15}$$

where $M = N(N-1)/2$. It is then possible to determine a set of *disparities* $\hat{d}_{ij}$ that are "nearly equal" to d_{ij}, whilst still retaining monotonicity with respect to the corresponding δ_{ij}. That is:

$$\hat{d}_{i_1 j_1} \leq \hat{d}_{i_2 j_2} \leq \ldots \leq \hat{d}_{i_M j_M} \tag{2.16}$$

The $\hat{d}_{ij}$ are said to be *monotonically related* to the d_{ij}, and the fitting of those values is a monotonic regression of distance upon dissimilarity. The multi-pass procedure for the determining the disparities is as follows.

At each monotonic regression phase, the disparities $\hat{d}_{ij}$ are initialised as the distances d_{ij}, and are listed such that their corresponding dissimiliarities δ_{ij} are in ascending order. In the first pass, each pair of adjacent (list-wise) disparity values is compared and if the two variables are not correctly ordered, they are combined into a 'group' with a common disparity value equal to the arithmetic mean of the combined values. Subsequent passes are then similar, except previously combined groups may be compared together as well as with other adjacent groups and/or single disparity values. The procedure terminates when no further groupings are required and the listed disparities are either equal (within groups) or in the requisite ascending order.

Kruskal then defined an objective measure based on these disparities:

$$\text{STRESS} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}, \tag{2.17}$$

where the denominator again normalises for the number and scaling of the dissimilarities.

The monotonic regression step occurs first, determining the $\hat{d}_{ij}$, following which those calculated values are used in a gradient-descent minimisation step of equation (2.17). These alternate steps are then repeated until a local minimum of STRESS is attained. Inevitably this procedure is computationally demanding and prone to finding sub-optimal local minima. Also, there is again no indication of choice of dimensionality q, so it is usually necessary to generate several configurations in a number of dimensions for comparative purposes.

The modern approach to NMDS is the *Alternating Least Squares* procedure, ALSCAL, and is available in the popular SPSS software package [Young and Harris 1990]. Since their introduction, the nonmetric schemes have become the dominant scaling models, with the classical procedure now rarely used.

2.6 Comparison of MDS and the Sammon Mapping

In Sammon's original paper he briefly mentioned the connection to MDS methods, and this relationship was further clarified by Kruskal [1971].

Sammon's mapping is effectively a metric, but nonlinear, scaling method. As such, its exact analogue does not exist in the MDS domain. Whilst all these latter scaling techniques may be applied to dissimilarities generated directly from a set of points, doing so defeats the primary motivation behind their development which is to produce such a spatial configuration from non-spatial data. Nevertheless, comparison of equations (2.17) and (2.3) indicates that the operation of the Sammon mapping, ignoring normalisation terms, is identical to a nonmetric scaling procedure without the monotonic regression step.

Algorithmically, MDS and the Sammon Mapping are effectively identical; it is only the difference in the source of the input data that differentiates between the two schemes. Conceptually, this is illustrated in figure 2.10 below.

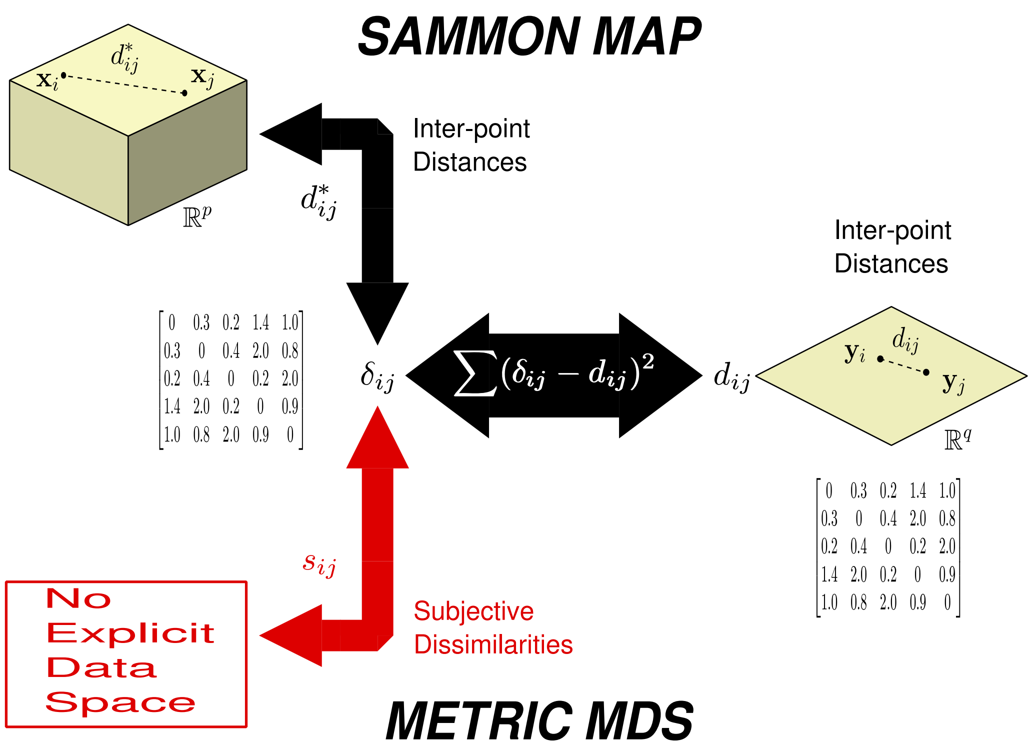

Figure 2.10: A schematic of the operation of the Sammon Mapping and MDS, emphasising the conceptual distinction between the two.

2.7 Conclusions

This chapter has described the standard methods for topographic feature extraction in both the neural network and statistical domains. The next chapter will introduce an alternative, *feed-forward* neural network approach to topographic mapping, which will be based on Sammon's projection. The emphasis of the particular method proposed is for the purposes of visualisation or exploratory data analysis, and this has motivated the choice of the latter technique as the basis for its design. As illustrated in Section 2.4, Sammon's approach to topographic mapping retains significantly more of the salient global data structure than the SOFM paradigm.

The key principle from MDS outlined in Section 2.5 — that informative configurations of points can be generated via topographic constraints from non-spatial data — will be incorporated into the method to enable the exploitation of additional subjective knowledge.

A generalisation of this neural network model to classical MDS in particular will be examined in Chapter 5, in the context of principal components analysis with neural networks.