Artefactual Structure From Topographic Mappings

NICHOLAS P. HUGHES

Master Of Science
(By Research)

ASTON UNIVERSITY, BIRMINGHAM

September 1999

ASTON UNIVERSITY, BIRMINGHAM

Artefactual Structure From Topographic Mappings

NICHOLAS P. HUGHES

Master Of Science, 1999

Thesis Summary

This thesis is a study of the problem of artefactual structure from topographic mappings, in particular Sammon's Mapping and its close relative Metric Multidimensional Scaling. Such structure is termed artefactual because it is not representative of true underlying structure in the data and is a side-effect of the mapping algorithm. The problem is investigated from both an experimental and a theoretical standpoint, and it is found that the choice of distance metric in the mapping algorithm is fundamental to the degree of artefactual structure observed.

The results of this work are then used to gain insight into a recent and controversial use of techniques from Multidimensional Scaling in the analysis of the connectivity of regions in the macaque monkey visual cortex. In particular it has been debated in the academic literature the extent to which the resulting mappings are corrupted by artefactual structure. This premise is investigated experimentally and the support of the mappings for the "two streams" hypothesis of visual processing is discussed in detail.

Keywords: Data Visualisation, Artefacts, Dimensionality Mismatch, Sammon's Mapping, Multidimensional Scaling, Primate Visual Cortex

Acknowledgements

Firstly I would like to thank my supervisor, Prof. David Lowe, for his help and guidance, and for interesting and stimulating discussions throughout the course of this project. I would also like to thank Dr. Mike Tipping from Microsoft Research for his patience in answering my queries on Multidimensional Scaling and computing matters in general. From the Neural Computing Research Group I wish to thank Dr. Ian Nabney for his help with miscellaneous maths problems and Dr. David Saad for his useful comments and ideas.

Finally I would like to thank Prof. Lionel Tarassenko for his encouragement and advice over the last few years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> The single biggest problem we face is that of visualisation.
>
> *Richard Feynmann*, 1945.

The volume of information produced throughout the world continues to grow at an ever increasing rate. In the finance industry vast quantities of complex time series data are generated on a daily basis, and in the area of molecular biology new DNA and protein sequences are regularly mapped out. This rapid growth of information, facilitated by the continual exponential increase in computer power and storage capacity, has created a need and an opportunity for techniques that are capable of discovering meaningful patterns and relationships from large amounts of data. Without these techniques, the derivation of useful knowledge from such data is often impossible.

This subject, generically known as *information processing*, is of increasing importance to many of the world's leading companies and research labs. Indeed, the past decade has seen an explosion of interest in the areas of Data Mining and Knowledge Discovery in Databases (KDD), where the aim is to extract useful knowledge from the vast quantities of information that are stored in company databases worldwide. An important problem therefore is how to extract this knowledge, and furthermore, the reliability with which decisions can then be based on such knowledge.

In this thesis, the type of information or data that will be used for information processing, is that in numeric form. Typically the data, often termed multivariate, will be characterised by a number of measurements which describe in detail a group of objects. Thus a given dataset can be regarded as a table of values, where each row represents a set of measurements for a specific object and each column the values of a particular measurement for each one of the objects. Importantly, the total number of columns in the table represents the *dimensionality*

of the data. This value indicates the dimension of the space the data naturally resides in.

In practice, most real-world datasets will have a dimensionality greater than three, thus precluding the possibility of direct visualisation of the raw data alone. Hence one goal of information processing is to provide a mechanism for the *visualisation* and *exploration* of such high-dimensional multivariate data. One method for achieving this is to find a projection of the data from the original (high-dimensional) data space to a two- or three-dimensional visualisation space. Such a projection would naturally allow for both a better understanding of the underlying structure in the data and an examination of any clustering which might be present in the data.

An important problem then is how to define, and indeed compute, this projection in practice. In addition to mapping the data to a lower dimensional space, the projection must also seek to retain the "interestingness" in the data. In this context then, the visualisation space should exhibit as much of the original structure in the data as possible, with minimum loss of detail. Conversely, any structure which is retained under the projection should be representative of true underlying structure in the data. The different approaches to producing such projections (and hence the various ways in which the term "interestingness" is defined) are discussed in Chapter Two.

This thesis examines two closely related techniques for data visualisation, those of Sammon's Mapping and Metric Multidimensional Scaling. In particular the validity of the mappings obtained after the projection of high-dimensional unstructured data is investigated. The results of this work are then applied to neuroanatomical connection data to provide new evidence in the study of the macaque monkey visual cortex. The remainder of this chapter gives an introduction to topographic mappings, an overview of the artefactual structure problem, the aims of this thesis and finally a plan of the work contained herein.

1.1 What is a Topographic Mapping?

Before considering what features characterise a topographic mapping, it is perhaps more useful to first consider what is meant by the phrase "a mapping". In the most general sense, a mapping is simply a transformation from one space to another. This transformation may come in the form of a projection from the data space to the map space, in which case it is referred to as a *projection mapping*. Alternatively points lying in the data space can be considered to have been generated by points lying in the map space, in which case the transformation is referred to as a *generative mapping*.

It is also useful to group the various mapping techniques according to their linear-nonlinear and supervised-unsupervised nature [Mao and Jain 1995]. A *linear* mapping defines a linear

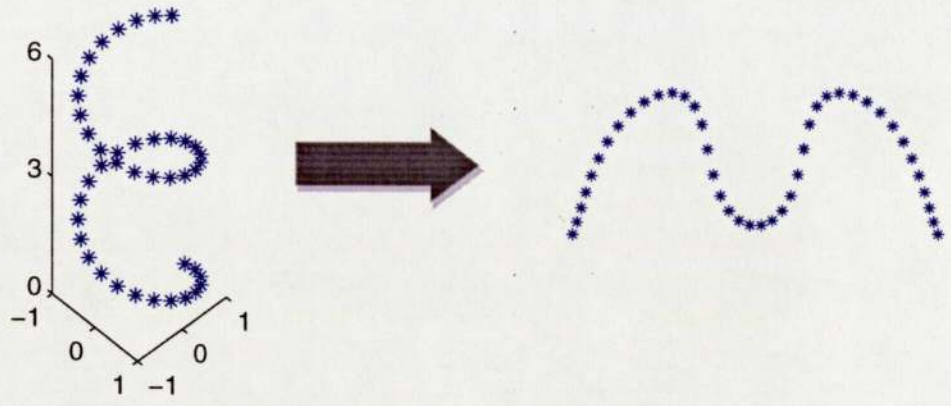


Figure 1.1: A two-dimensional representation of a three-dimensional helix through the use of a topographic mapping. A Sammon mapping (see Section 2.4.1) of the original helix data was performed, resulting in the two-dimensional curve shown above. Clearly the preservation of inter-point distances in the map space accurately captures the sinusoidal nature of the helix.

transformation of the data space to produce the map space. Although such mappings have the advantage that they can be derived analytically, they are constrained by their global linearity and are therefore less flexible than *nonlinear* mappings. Mappings which rely only on the input data itself, and do not make use of any additional class information or target data, are considered to be *unsupervised*. Alternatively, in a *supervised* technique, the mapping utilises additional a priori information about the data (usually in the form of class labels) to produce a map space with improved inter-class separation. Recently a third class of models has been proposed which can be considered to be *relatively supervised* [Lowe 1993]. With these models, a relative measure of the target separation between pairs of vectors in the map space is provided, rather than a set of explicit target map vectors (see Section 2.4.3).

As previously discussed, for the purposes of visualisation and exploratory data analysis, the transformation should map the data from the data space to a two- or three-dimensional visualisation space. Furthermore, as much of the original structure in the data as possible should be preserved under this transformation. One possible way of implementing this constraint is to require that points which lie close together in the data space should lie close together in the map space. This then gives rise to a *topological ordering* in the map space. Such transformations, known as *neighbourhood preserving mappings*, are discussed in more detail in Section 2.3.

An alternative, and perhaps more natural definition, is to require that *all* distance relationships between data points be preserved under the transformation. In this way, points which lie close together in the data space will lie close together in the map space, and similarly points which lie far apart in the data space will lie far apart in the map space. Such a transformation is known as a *topographic mapping* and has the property that the geometric structure

of the original data is optimally preserved in the lower dimensional space.

Figure 1.1 shows a topographic mapping of a three-dimensional helix. By attempting to preserve all the inter-point distances in the transformation, the resulting map[1] retains much of the structure of the original data. Upon closer inspection however, it is evident that although the original data points are uniformly distributed along the length of the helix, the same is not true of the distribution of the points in the map space. In particular the map points appear more closely grouped around the bends of the map curve. This serves to illustrate an important point, that when data undergoes a dimension reducing transformation, some of the original structure is inevitably lost.

1.2 The Artefactual Structure Problem

The artefactual structure problem, also referred to as *dimensionality mismatch*, concerns the topographic mapping of unstructured data. In particular it has been observed that a topographic mapping of high-dimensional randomly distributed data gives rise to an artefactual or illusory structure in the map space [Klock and Buhmann 1997]. In addition, the degree of the resulting artefactual structure is known to increase with the dimensionality of the data (for randomly distributed input data). This structure is termed *artefactual* because it is not representative of the underlying data generator.

Such artefactual structure is a serious problem because for most real-world datasets it is not known in advance whether any structure actually exists in the data (and what form this structure might take). Indeed the aim of performing a topographic mapping is to reveal any hidden structure which may be present in the data through a lower dimensional map. Conversely if the original data is unstructured then the spatial organisation of the resulting map should reflect this. Thus the tendency of topographic mappings to produce maps which exhibit artefactual structure poses a serious concern for the data analyst who makes use of these techniques. In particular the level of confidence which can be attached to any structure derived from a topographic mapping is brought into question.

A recent example of this problem concerns the study of the connectivity of different regions in the visual cortex of the macaque monkey [Young 1992]. A topographic mapping was used to produce a two-dimensional map of the various regions in the macaque visual cortex, for the purposes of visualisation and analysis. However the resulting structure in the map generated significant controversy as to its validity and potentially artefactual nature [Simmen, Goodhill, and Willshaw 1994]. This is discussed in more detail in Section 3.5.

[1] The word "map" is used in this thesis to refer to the resulting two- or three-dimensional image of the mapping process.

1.3 Thesis Aim

The aim of this thesis is to investigate the artefactual structure problem, from both an experimental and a theoretical standpoint. The types of topographic mappings considered are those of Sammon's Mapping and its close relative Metric Multidimensional Scaling. In particular the importance of the choice of distance metric used within the mapping algorithm and the dimensionality of the input data are both investigated. Finally the results of this work are then used to analyse a real-world problem of artefactual structure from the area of neuroanatomy.

1.4 Plan of This Thesis

Chapter 1 is this introduction.

Chapter 2 describes the established techniques for multivariate data projection. These techniques are grouped according to the way in which they attempt to preserve the underlying geometric structure in the data, and also by their supervised-unsupervised and linear-nonlinear nature.

Chapter 3 presents a detailed study of the problem of artefacts in topographic mappings. In particular how the degree of artefactual structure varies with the distance metric used is investigated. The results obtained are then used to gain insight into the organisation of the primate visual cortex through the mapping of neuroanatomical connection data.

Chapter 4 concludes the thesis with a summary of the key results and suggests directions for future research.

Chapter 2

Established Techniques for Multivariate Data Projection

2.1 Introduction

This chapter considers a number of different techniques for the projection of multivariate data. These techniques are divided into three groups, determined by the manner in which they attempt to preseve the underlying geometric structure in the data. At the end of the chapter a unifying taxonomy is presented displaying the relationships between the various methods.

2.2 Classical Techniques

The approaches to multivariate data projection described in this section are related by the fact that they place no explicit criterion on the preservation of geometric structure in the data under the projection. Instead they seek to maximise alternative criteria in order to find a lower dimensional representation of the data that can subsequently be used for data analysis or for further modelling.

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA), also known as the Karhunen-Loéve Transform, is a linear unsupervised feature extraction technique which makes use of the covariance matrix of the data to find a transformation to new variables that are uncorrelated. For the purposes

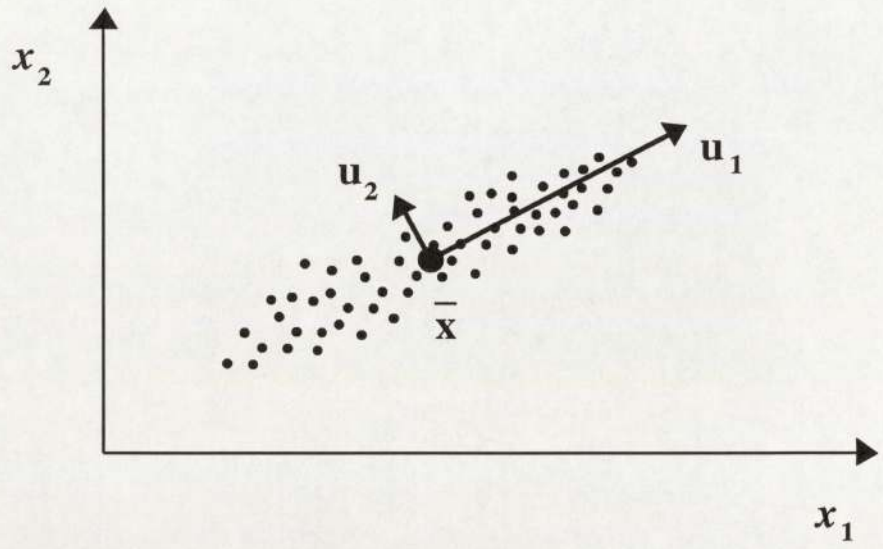


Figure 2.1: An example of dimensionality reduction with Principal Component Analysis.

of dimensionality reduction, the data is projected from its original d dimensional space onto the q (where $q < d$) orthogonal axes that retain the maximum variance.

For a dataset composed of N vectors $\mathbf{x}$, the algorithm proceeds as follows. First the sample covariance matrix of the data is computed, given by:

$$\widehat{\Sigma} = \sum_{n=1}^{N} (\mathbf{x}^n - \hat{\boldsymbol{\mu}})(\mathbf{x}^n - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

where $\hat{\boldsymbol{\mu}}$ is the sample mean, given by:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$$

The eigenvectors and eigenvalues of $\widehat{\Sigma}$ are then found, and the eigenvectors corresponding to the q largest eigenvalues (known as the *principal components*) are retained. Finally the input vectors $\mathbf{x}$ are projected onto the principal components to produce the transformed dataset. Figure 2.1 above shows an example for a two-dimensional dataset.

Since PCA only produces a linear subspace, it will be sub-optimal when the underlying structure in the data is *nonlinear*. This has led to the development of a number of techniques for improving the ability of PCA to capture any nonlinearity which may be present in the data. One technique is to assume that in local regions of the data space, a linear approximation will be sufficient. In this way a globally nonlinear dataset can be modelled by a number of local PCA models [Tipping and Bishop 1997].

An alternative method is to utilise an auto-associate neural network, as shown schematically in Figure 2.2 overleaf. The network is constructed with less hidden units than inputs and

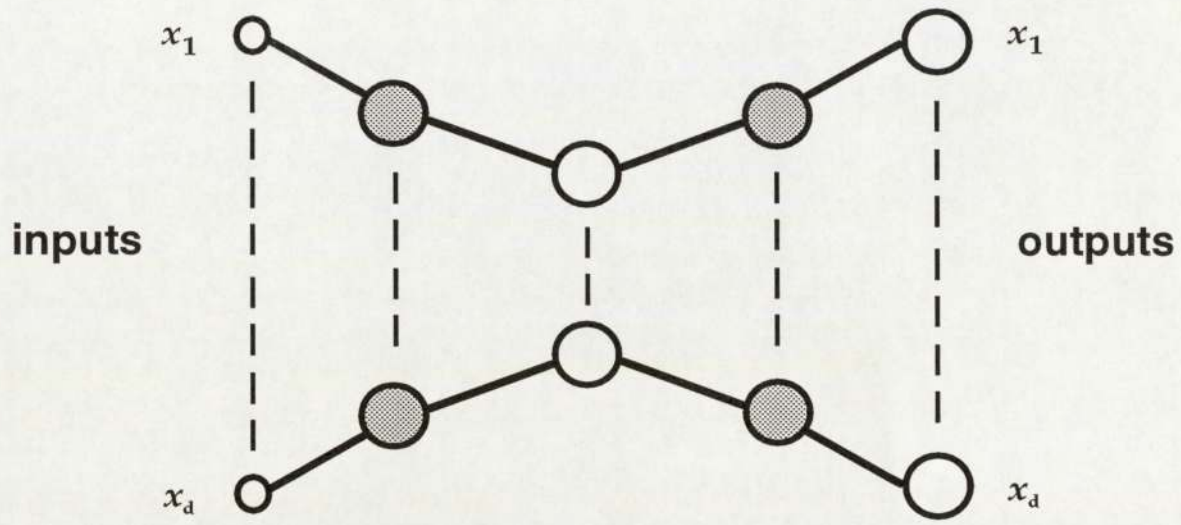


Figure 2.2: A schematic of a four-layer auto-associative neural network. If the shaded units implement the sigmoidal activation function, then the trained network will perform nonlinear PCA.

is then trained to map each input vector onto itself. In this way the network is encouraged to find an effective lower-dimensional representation of the data. If a single hidden layer network is used, the trained network will simply perform PCA, regardless of the choice of hidden unit activation function. However if a network with three hidden layers is used (and sigmoidal activation functions for the first and third hidden layers), then the trained network will effectively perform nonlinear principal component analysis. This is discussed in more depth in Bishop [1995].

2.2.2 Canonical Variates

Canonical Variates is the name given to a linear *supervised* technique which aims to produce an optimal linear dimensionality reduction of the original data. This is achieved by ensuring that the resulting feature space (which is spanned by the canonical variates) optimally distinguishes between the classes present in the data. For the purposes of classification it is then relatively simple to construct a discriminant function which assigns unlabelled data to one of the classes on the basis of its projection.

Before considering Canonical Variates in general, it is useful to first consider *Fisher's Linear Discriminant* - which for a dataset containing two unique classes, finds a projection of the data onto a one-dimensional space that optimally separates the two classes. In this way an input vector $\mathbf{x}$ is projected onto a value y given by:

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

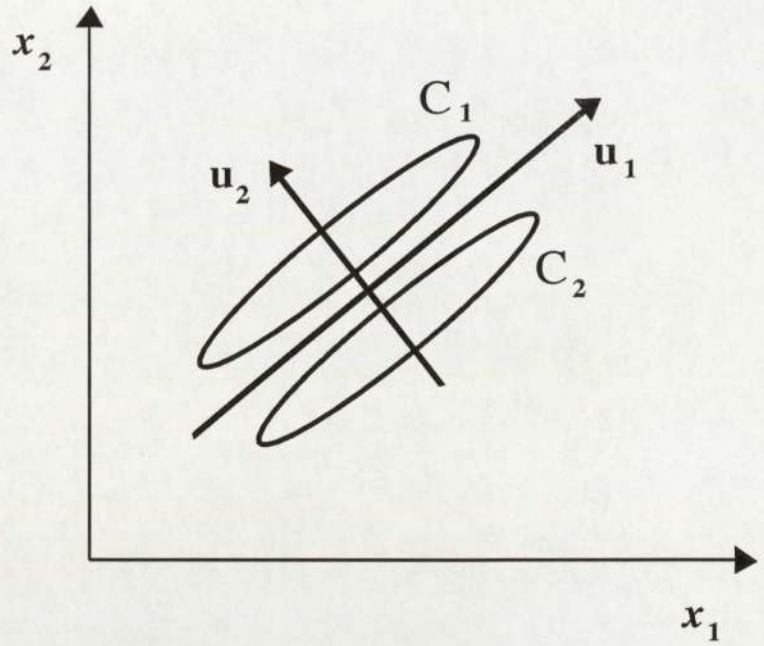


Figure 2.3: An example of a simple classification problem for which dimensionality reduction using Fisher's Linear Discriminant is superior to PCA.

where $\mathbf{w}$ is a vector of adjustable weight parameters. In addition if there are a total of N input vectors in the dataset, with N_1 from class C_1 and N_2 from class C_2, then the mean vector and the individual class mean vectors are given by:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n}\mathbf{x}^n, \qquad \hat{\boldsymbol{\mu}}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}^n, \qquad \hat{\boldsymbol{\mu}}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}^n$$

For optimal separation of the two classes, the criterion is to maximise the separation of the projected class means whilst minimising the projected within class variances. The mathematical embodiment of this is to maximise a function $J(\mathbf{w})$ defined as:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{S}}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{S}}_W\mathbf{w}}$$

where $\hat{\mathbf{S}}_B$ is the *between-class* sum of squares matrix given by:

$$\hat{\mathbf{S}}_B = N_1(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}})^{\mathrm{T}} + N_2(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

and $\hat{\mathbf{S}}_W$ is the *within-class* sum of squares matrix given by:

$$\hat{\mathbf{S}}_W = \sum_{n\in C_1}(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}} + \sum_{n\in C_2}(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_2)^{\mathrm{T}}$$

The function $J(\mathbf{w})$ is then maximised by setting the weight vector $\mathbf{w}$ to the dominant eigenvector of $\hat{\mathbf{S}}_W^{-1}\hat{\mathbf{S}}_B$. Figure 2.3 above shows an example of a problem where PCA fails to retain the discriminatory information under the projection but Fisher's linear discriminant

successfully retains it. Transforming the data to one-dimension using PCA results in a projection onto the vector $\mathbf{u}_1$ which would merge the data from the two classes and render accurate classification impossible. In contrast however, the use of Fisher's Linear Discriminant would result in a projection onto the vector $\mathbf{u}_2$ (or the first canonical variate), which would retain the discriminatory information and hence allow for subsequent classification.

Canonical Variates generalises this technique to a dataset with c classes and for a projection of the input data onto a space of dimension d' (where $1 \leq d' < c$). In addition, let there be N_k input vectors from each class C_k. Then for optimal separation of the c classes in the d' dimensional projection space, a suitable criterion is to maximise $J(\mathbf{W})$ defined as:

$$J(\mathbf{W}) = \mathrm{Tr}\,[\hat{\mathbf{S}}_W^{-1}\hat{\mathbf{S}}_B]$$

where $\hat{\mathbf{S}}_W$ is now the within-class sum of squares matrix in the *projected d' dimensional space* and $\hat{\mathbf{S}}_B$ is a mesaure of the between-class sum of squares matrix also in the *projected d' dimensional space*. These are then defined as:

$$\hat{\mathbf{S}}_W = \sum_{k=1}^{c} \sum_{n \in C_k} (\mathbf{y}^n - \hat{\boldsymbol{\mu}}_k)(\mathbf{y}^n - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}$$

$$\hat{\mathbf{S}}_B = \sum_{k=1}^{c} N_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

where

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{y}^n, \qquad \hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^{c} N_k \hat{\boldsymbol{\mu}}_k,$$

The projection vectors $\mathbf{w}_i$ (which are the rows of the projection matrix $\mathbf{W}$), can then be obtained as the solutions of the generalised eigenvector equation:

$$\hat{\mathbf{S}}_B \mathbf{w}_i = \lambda_i \hat{\mathbf{S}}_W \mathbf{w}_i$$

The resulting discriminant axes which span the projected feature space are termed the *canonical variates*. *Linear Discriminant Analysis* (LDA) takes this procedure one step further by determining the linear discriminant boundaries which optimally separate the different classes.

One limitation of this technique is that for a c class problem there is a maximum of $c - 1$ indepedent projection vectors available. For the purposes of data visualisation and exploration this is rarely a problem since a dataset with three or more classes can be projected onto a two-dimensional space. However if this feature space is subsequently to be used for classification purposes (as with LDA) and the original dataset is of a high dimensionality with only a small number of classes, then the dimensionality reduction is likely to result in the loss of useful discriminatory information. In this case pre-processing with PCA, which

is able to extract any number of independent projection vectors (up to the dimensionality of the original data space), is more appropriate.

As noted previously, Canonical Variates is a linear technique and is therefore unable to capture any nonlinearity which may be present in the data. This problem can be overcome by allowing for a nonlinear transformation of the input data in order to maximise an appropriate discriminant criterion. In particular it has been shown that the hidden units of a multi-layer perceptron (with linear output units) trained with a sum-of-squares error function, perform *nonlinear discriminant analysis* on the input data [Webb and Lowe 1990].

2.3 Neighbourhood Preserving Mappings

The techniques considered in this section all attempt to preserve some notion of the geometric structure in the data. This is achieved by requiring that points which lie close together in the data space also lie close together in the map space. The resulting map is then said to display a *topological ordering* of the data. Such mappings effectively preserve local neighbourhoods or clusters of points under the projection.

2.3.1 Self-Organising Feature Map

The Self-Organising Feature Map (SOFM), also known as the 'Kohonen Map', takes its inspiration from the topologically ordered maps found in the brains of the more developed animal species [Kohonen 1990]. As an example, nerve cells and fibres in the auditory pathway are arranged anatomically in relation to the frequency which causes the greatest response in each neuron. Thus the neurons transform input signals into a *place-coded probability distribution* of the data by sites of maximum relative activity within the map. Kohonen's SOFM is a model of how such self-organisation can take place inside the brain.

The architectural layout of the most common form[1] of SOFM is shown in figure 2.4 overleaf. The output units are arranged on a regular lattice, with each unit connected to the input layer through a specific weight vector. Learning in the SOFM then consists of adapting the weight vectors such that the presentation of an input pattern to the trained network gives rise to a localized region of activity in the resulting feature map.

For a dataset of dimensionality d, and for a SOFM with K output units, the Kohonen Self-Organising algorithm is as follows:

❶ At time step $t = 0$ select an inital set of K weight vectors $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K\}$ for each of the K units in the feature map. The weight vectors are of dimension d and their initial values may be chosen randomly.

❷ Choose a pattern $\mathbf{x}$ from the training dataset and identify the "winning" unit i whose weight vector $\mathbf{w}_k$ is nearest to $\mathbf{x}$. That is, the unit for which:

$$\| \mathbf{x} - \mathbf{w}_k \| < \| \mathbf{x} - \mathbf{w}_{k'} \| \qquad \forall\, k' \neq k$$

❸ Move $\mathbf{w}_k$ closer to $\mathbf{x}$ by an amount determined by the learning rate $\eta(t)$

$$\mathbf{w}_k^* = \mathbf{w}_k + \eta(t)(\mathbf{x} - \mathbf{w}_k)$$

[1] Hexagonal grid configurations of the units in the output layer are also sometimes used.

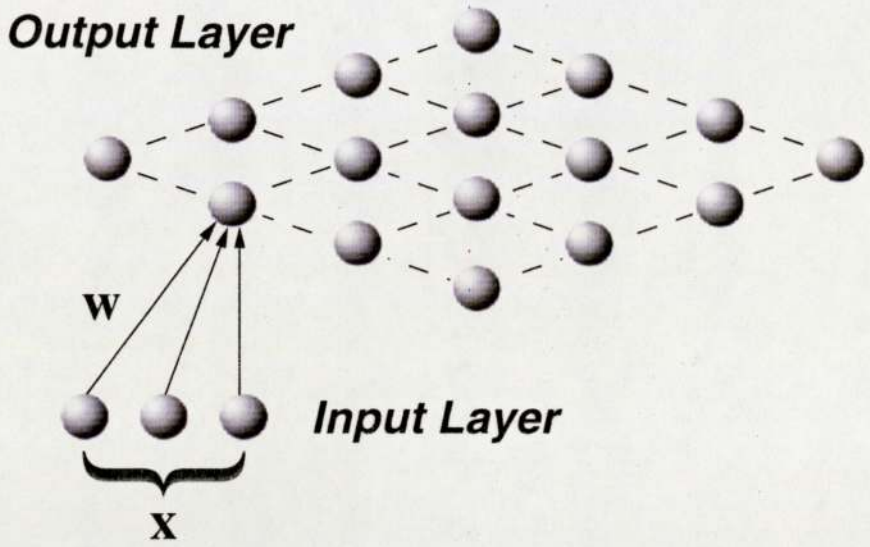


Figure 2.4: A schematic of Kohonen's Self-Organising Feature Map

❹ In addition move units which are in the *neighbourhood* of the winning unit (indicated here by the subscript $k \pm p$) by an amount proportional to $\beta(t)$

$$\mathbf{w}^*_{k\pm p} = \mathbf{w}_{k\pm p} + \eta(t)\beta(t)(\mathbf{x} - \mathbf{w}_{k\pm p})$$

❺ Increment the time step and repeat from step ❷ until the map has stabilised.

For the purposes of visualisation each input pattern $\mathbf{x}$ is then projected onto the unit in the feature map whose weight vector $\mathbf{w}$ is closest to $\mathbf{x}$. The key to the *self-organising* nature of the map is in step ❹ and in particular the neighbourhood function β. This function is usually chosen to have the shape of a 'Mexican Hat' or a Gaussian so that its value decreases with the distance (in the output layer) between the winning unit and any other unit. Hence the weights of units close to the winning unit will move closer to the input pattern $\mathbf{x}$, and the weights of units further away by smaller amounts. For a neighbourhood function of zero width the learning algorithm reverts to the adaptive K-means clustering algorithm [Tarassenko 1998]. The magnitude of the learning rate η and the width of the neighbourhood function β are usually decreased with time during the training process. In this way the coarse global topology of the map is formed during the early stages of training while the local detail is fine-tuned later.

The SOFM has achieved many successes in practical applications such as speech recognition, image processing, robotics, process control and telecommunications. The learning algorithm scales well with the size of the dataset involved and it is this tractability that makes it such a popular tool for data visualisation. However, the SOFM does suffer from a number of significant drawbacks. Firstly there is no explicit error measure defined by the algorithm and therefore it is impossible to gauge a measure of the quality of a map once training is

completed. In addition there in no proof of convergence and the learning rate as well as the neighbourhood size reduction rate must be chosen by trial and error. Finally the fixed lattice of units in the output layer inevitably leads to some distortion in the representation of the true structure in the data.

2.3.2 Generative Topographic Mapping

The Generative Topographic Mapping (GTM) is a probabilistic alternative to Kohonen's SOFM [Bishop, Svensén, and Williams 1998]. Although the name of the technique indicates the mapping is topographic in nature, it is in fact strictly a neighbourhood-preserving or topological mapping when considered with the definitions of Section 1.1 in mind.

GTM is a *latent variable model* which seeks to represent the distribution of data in a space of several dimensions in terms of a smaller number of latent, or hidden, variables. For the purposes of data visualisation the number of such latent variables, L, is usually chosen to be two or three. In addition the GTM algorithm is a *generative model* since it defines a mapping from the latent space to the data space. This mapping can then be inverted through the use of Bayes' theorem to produce a mapping from the data space to the latent (or visualisation) space.

The technique begins by first defining a function $\mathbf{x}(\mathbf{y}; \mathbf{W})$ which maps each point $\mathbf{y}$ in the (two- or three-dimensional) latent space to a point $\mathbf{x}(\mathbf{y}; \mathbf{W})$ in the data space. The unconditional probability distribution in latent space, $p(\mathbf{y})$, is chosen to be a sum of delta functions centred on the nodes of a regular lattice in latent space. The function $\mathbf{x}(\mathbf{y}; \mathbf{W})$ is usually given by a generalised linear regression model of the form:

$$\mathbf{x}(\mathbf{y}; \mathbf{W}) = \mathbf{W}\phi(\mathbf{y})$$

where the elements of $\phi(\mathbf{y})$ consist of fixed basis functions. As a result of the choices for $\mathbf{x}(\mathbf{y}; \mathbf{W})$ and $p(\mathbf{y})$, each node in latent space becomes the centre in a constrained Gaussian mixture model in data space - the free parameters of which are determined through use of the EM algorithm [Bishop 1995]. In this way each data point $\mathbf{x}$ induces a posterior distribution in $\mathbf{y}$-space. For visualisation purposes the data points are then projected to the mean or mode of their posterior distribution.

Although the GTM algorithm is a principled technique which can be used to model any n-dimensional distribution of data, it is best suited to modelling specific types of distributions. In particular optimal performance will be achieved when modelling moderately curved L-dimensional distributions (embedded in the original data space) of roughly rectangular shape [Svensén 1998]. Prior knowledge concerning the underlying structure in the input data is therefore useful in determining the applicability of GTM for the purposes of data visualisation.

2.4 Topographic Mappings

This section introduces a class of techniques for multivariate data projection known as topographic mappings. The aim of such mappings is the optimal preservation of the underlying geometric structure in the data under the transformation.

2.4.1 Sammon's Mapping

Sammon's Mapping, also known as the Nonlinear Mapping, is an algorithm for finding a transformation of a dataset of dimensionality p onto a nonlinear map space of dimensionality q (where $q < p$), whilst preserving as well as possible the inter-point distances in the data [Sammon 1969]. This is achieved through the minimisation of an error or stress function, defined as:

$$E_{sammon} = \frac{1}{\sum_i \sum_{j \neq i} d^*_{ij}} \sum_i \sum_{j>i} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}}$$

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 points i and j in the map space $\mathbb{R}^q$. The distance measure $\| \ldots \|$ may be any valid distance metric, although usually the Euclidean distance is chosen.

The term $(d^*_{ij} - d_{ij})^2$, in the numerator of the error function, is a measure of the deviation between the inter-point distances in the input space and those in the map space. Hence minimisation of this error function involves adjusting the map points $\mathbf{y}_i$ in order to improve the representation of the geometric structure of the input data in the map space. In Sammon's original paper a simple gradient descent technique was proposed for this minimisation. However in practice more advanced nonlinear optimisation techniques such as *conjugate-gradient* descent or *quasi-Newton* methods are to be preferred [Bishop 1995]. The training of a Sammon Mapping is illustrated schematically in figure 2.5 overleaf.

The purpose of the constant term, $1/\sum_i \sum_{j \neq i} d^*_{ij}$, is simply to reduce the sensitivity of the error value to the number of points used and their scaling. However its presence does not normalise the error value (ie. restrict it to the range $[0, 1]$) since the values of d_{ij} may be much greater than those of d^*_{ij} (especially at the start of the procedure).

The purpose of the d^*_{ij} term in the denominator of the sum is to weight the errors relative to the magnitude of the distances involved. In this way a pair of points which are a distance 9 apart in the map (d_{ij}) but 10 apart in the data space (d^*_{ij}), and another pair of points which are a distance 90 apart in the map but 100 apart in the data space, will both contribute equally to the overall error value. However since the overall geometric structure of the input data is generally unknown at the outset of the mapping procedure, there is no reason why we should weight the preservation of local structure over global structure.

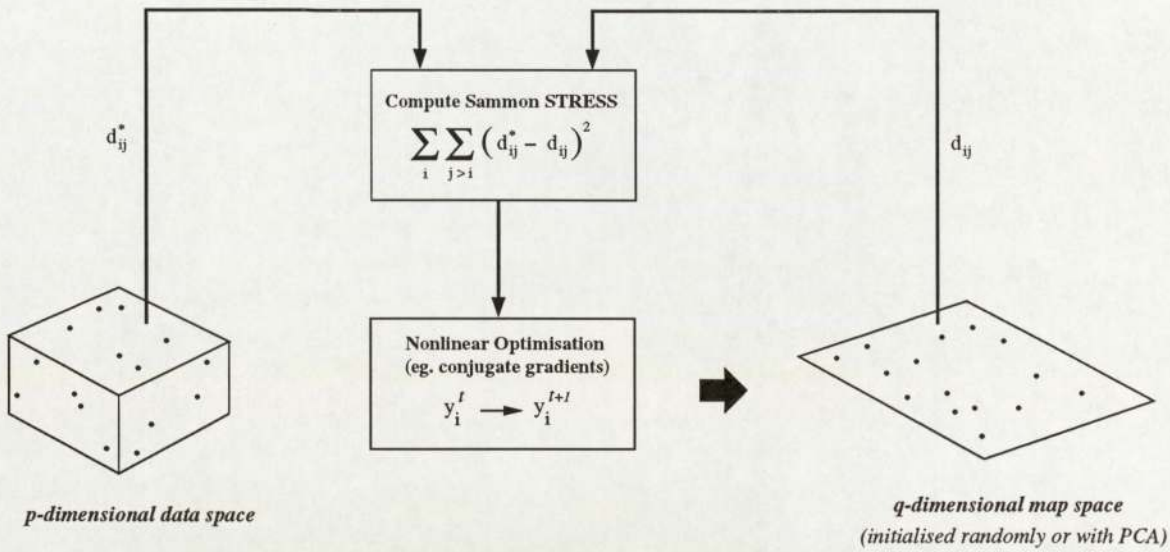


Figure 2.5: A process diagram to illustrate the training of a Sammon Mapping.

Thus an alternative, and simplified definition of Sammon's Mapping, is given by the Sammon STRESS:

$$\text{Sammon STRESS} = \sum_i \sum_{j>i} \left(d^*_{ij} - d_{ij}\right)^2$$

Although Sammon's Mapping is traditionally considered to be an unsupervised technique (in the sense that only objective distance measurements are used), it can be extended to incorporate additional subjective class information into the mapping process. This is achieved by modifying the original STRESS function, to give:

$$\text{Generalised Sammon STRESS} = \sum_i \sum_{j>i} \left(\delta_{ij} - d_{ij}\right)^2$$

where now:

$$\delta_{ij} = (1-\alpha)d^*_{ij} + \alpha s_{ij} \qquad \text{and} \qquad 0 \leq \alpha \leq 1$$

Once again, d^*_{ij} and d_{ij} are the inter-point distances in the input space and the map space respectively. The term s_{ij} represents a subjective dissimilarity value between points i and j. In a simple case this value could be 0 if the two points are from the same class and 1 otherwise. However if more detailed information is available it is possible to construct more advanced subjective dissimilarities, which can be thought of as measurements from a *subjective metric* over the input space [Tipping 1996, Chapter 3]. Given this definition then, the purpose of the parameter α is to govern the degree to which the subjective metric influences the final output configuration. If $\alpha = 1$, the map is purely supervised since only the subjective metric information is utilised in the training process. Alternatively with $\alpha = 0$, the map is purely unsupervised since only the objective distance information d^* is used. Intermediate values

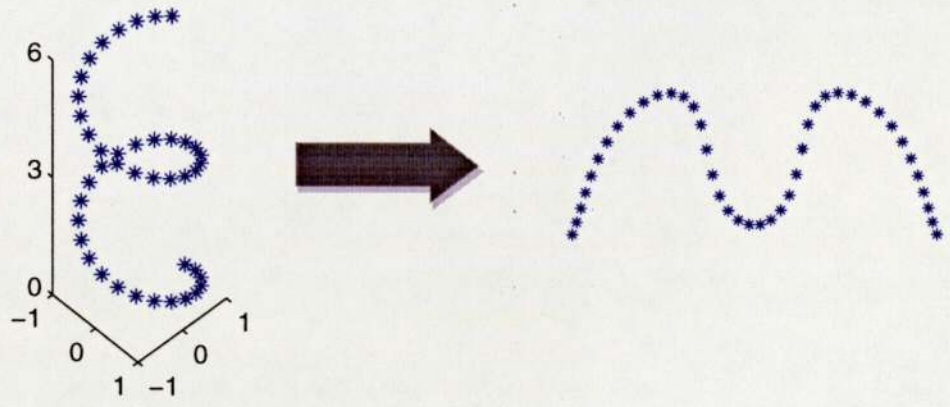


Figure 2.6: The result of applying Sammon's mapping to a three-dimensional helix.

of α therefore give rise to a *semi-supervised* mapping in which the objective spatial distances and the subjective dissimilarities are linearly combined to give a hybrid map space. In practice, for a given application of Sammon's Mapping to a particular dataset, the optimal value of α cannot be computed explicitly and must be determined through trial and error.

Figure 2.6 shows a Sammon Mapping of a three-dimensional helix. The geometric structure of the resulting map clearly indicates the topographic nature of the technique; points close together on the helix are close together on the map and points far apart on the helix and are similarly far apart on the map. In addition the mapping captures the *sinusoidal* nature of the helix in the reduced two-dimensional map space. However some information is inevitably lost in the mapping process. For example, although the individual data points are equally spaced along the length of the helix, the map points appear to cluster slightly around the regions of the map curve with the greatest curvature. This is a consequence of the fact that the mapping is attempting to match inter-point distances derived from a three-dimensional space with those from a two-dimensional space. Thus unless the input data lies on a two-dimensional manifold in the data space, the resulting mapping will always exhibit some distortion of the original data.

Although Sammon's Mapping optimally preserves the topographic nature of the input data under the transformation, it does suffer from a number of drawbacks. Firstly the time complexity of the algorithm is $O(N^2)$, ie. proportional to the square of the number of data points, since each evaluation of the Sammon STRESS function requires a double loop over i and j. In practice then Sammon's Mapping is usually limited to datasets composed of no more than 1000 data points. To overcome this restriction, Sammon himself suggested the application of an initial clustering phase to the dataset (using for example the *K-means algorithm*) to generate a set of K prototype vectors in the input space (where $K \ll N$). Sammon's Mapping can then be applied to these prototype vectors at a much reduced computational

cost. Another drawback with Sammon's Mapping is that it is generated iteratively through the use of a nonlinear optimisation algorithm and is therefore prone to sub-optimal local minima. In practice the procedure is usually performed a number of times with different initial configurations and the mapping with the smallest final STRESS value is chosen.

2.4.2 Multidimensional Scaling

The techniques considered so far in this chapter have all made use of a set of input vectors (where each vector corresponds to a number of measurements or features) to produce a set of lower dimensional map vectors. Multidimensional Scaling (MDS) however, takes as its input a set of *proximity* values which provide a measure of the *similarity* or *dissimilarity* between the individual objects. Such proximity data is usually in the form of an $(N \times N)$ symmetric matrix, where the rows and columns represent the objects under consideration, and the values of the matrix represent the relative proximity between pairs of objects.

Given this data, the purpose of MDS is to find a configuration of N points in a two- or three-dimensional space, where each point represents an object and the geometric layout of the points reflects the relationships between the objects defined by the proximity matrix. In this way the information contained within the proximity matrix can be captured by a more succint *spatial model* which aids visualisation of the data and improves understanding of the process that generated it.

The set of values forming the proximity matrix may have been derived in a number of different ways, for example the data might be the results of a psychological experiment, or connectivity values from an analysis of different regions of the brain. Regardless of how the data has been derived, it can be considered to belong to one of four distinct levels of measurement [Schiffman, Reynolds, and Young 1981]. These are defined as:

- *Ratio*: Objects are placed on a scale such that the position along the scale represents the absolute magnitude of the attribute. Both the intervals and the zero point are relevant (eg. mass, velocity, etc).
- *Interval*: Objects are placed on a scale such that the magnitude of the differences between objects is shown by the scale. Thus intervals are meaningful, but not the zero point (eg. Celsius or Fahrenheit temperature scale).
- *Ordinal*: Objects are arranged in rank order of magnitude. The individual values and hence the intervals have no meaning (eg. a subject's rating of taste stimuli in a psychology experiment).
- *Nominal*: Objects are sorted into distinct groups (eg. males and females). This is the weakest or lowest level of measurement.

Since the input data has no explicit "data space", it is impossible to formulate a mapping from a data space to the desired map space (as with other mapping techniques). Instead MDS works by attempting to match the inter-point distances in the configuration space with the inter-object dissimilarity values given by the dissimilarity matrix (if a "similarity" matrix is given, it can be converted to a "dissimilarity" matrix simply by subtracting the values from an appropriate constant). The varying degrees of importance placed on matching these values to inter-point distances gives rise to two closely related techniques, known as *Metric* and *Nonmetric* Multidimensional Scaling.

The aim of Metric MDS is to match as closely as possible the dissimilarity values between pairs of objects to the corresponding spatial distances between pairs of points. In this way Metric MDS assumes the data measurements to be at the *ratio* or *interval* level. With Nonmetric MDS only the rank ordering of the data is deemed important and the aim of the technique is to match the ordering of the dissimilarities with the ordering of the distances. Hence Nonmetric MDS assumes the data to be at the weaker *ordinal* level of measurement.

Although Nonmetric MDS is generally considered a more powerful technique than its Metric counterpart (due mainly to its ability to handle data at the ordinal level), it cannot strictly be termed "topographic" since there is no real notion of "geometric structure preservation" with ordinal input data. Thus when the dissimilarity data is representative of spatial distance measurements between the objects, Metric MDS (which can be viewed as a form of Sammon's Mapping) is a more suitable technique.

A good example of the effectiveness of Metric MDS is given by the application of the technique to proximity data based on the distances between various cities. Table 2.1 overleaf shows the road distances between 18 cities located in various regions of mainland UK, and the resulting two-dimensional configuration[2] is shown in figure 2.7. The layout of the configuration points is clearly intuitive and the model captures the salient information in the proximity data in a visual and compact form. Although this example is artificial in the sense that the resulting configuration is already known *a priori*, it serves to illustrate the use of MDS as a powerful tool for visualisation when the inherent structure in the data *is* unknown.

The optimisation criterion for Metric MDS can be formulated in a very similar fashion to that of Sammon's Mapping. In particular, if the dissimilarity between objects i and j is denoted by δ_{ij} and the corresponding inter-point distance in the configuration space by d_{ij}, then we wish to minimise a STRESS function given by:

$$\text{STRESS} = \sum_i \sum_{j>i} (\delta_{ij} - d_{ij})^2$$

[2] After a suitable rotation about its centre of mass.

	Abd	Aby	Brm	Btn	Brs	Cam	Crl	Gla	Inv	Lds	Man	New	Nor	Oxf	Pez	Sot	Yor	Ldn
Aberdeen (Abd)	0	466	431	606	514	463	232	147	106	331	353	234	488	503	697	569	322	545
Aberystwyth (Aby)	466	0	124	285	128	216	233	331	492	171	132	272	278	156	305	221	198	239
Birmingham (Brm)	431	124	0	171	88	98	199	296	457	116	90	203	159	68	272	134	129	120
Brighton (Btn)	606	285	171	0	169	120	375	472	633	263	265	350	169	109	287	66	276	59
Bristol (Brs)	514	128	88	169	0	171	282	379	540	213	172	299	233	74	194	75	226	120
Cambridge (Cam)	463	216	98	120	171	0	257	354	515	147	160	229	63	81	361	131	156	60
Carlisle (Crl)	232	233	199	375	282	257	0	97	258	126	121	60	282	271	465	337	117	313
Glasgow (Gla)	147	331	296	472	379	354	97	0	173	222	217	154	379	368	562	434	213	410
Inverness (Inv)	106	492	457	633	540	515	258	173	0	384	379	266	540	529	723	595	375	571
Leeds (Lds)	331	171	116	263	213	147	126	222	384	0	43	94	172	171	396	237	24	199
Manchester (Man)	353	132	90	265	172	160	121	217	379	43	0	144	185	161	355	227	71	203
Newcastle (New)	234	272	203	350	299	229	60	154	266	94	144	0	254	257	482	324	88	285
Norwich (Nor)	488	278	159	169	233	63	282	379	540	172	185	254	0	144	423	193	181	115
Oxford (Oxf)	503	156	68	109	74	81	271	368	529	171	161	257	144	0	264	66	184	56
Penzance (Pez)	697	305	272	287	194	361	465	562	723	396	355	482	423	264	0	221	409	310
Southampton (Sot)	569	221	134	66	75	131	337	434	595	237	227	324	193	66	221	0	250	80
York (Yor)	322	198	129	276	226	156	117	213	375	24	71	88	181	184	409	250	0	211
London (Ldn)	545	239	120	59	120	60	313	410	571	199	203	285	115	56	310	80	211	0

Table 2.1: Road distances (in miles) between various mainland UK towns and cities.

Figure 2.7: The resulting 2-D configuration after applying Metric MDS to the dissimilarity data of Table 2.1

This function provides a measure of the deviation of the dissimilarities from the corresponding distances, and can be minisimed through the use of a nonlinear optimisation algorithm. An alternative definition[3] of Metric MDS is the SSTRESS measure, given by:

$$\text{SSTRESS} = \sum_i \sum_{j>i} \left(\delta_{ij}^2 - d_{ij}^2\right)^2$$

This function provides a measure of the deviation of the *squared* dissimilarities from the corresponding *squared* distances. The advantage of SSTRESS is that it can be minimised through the use of an *alternating least squares* procedure, which is based on the technique of *iterative majorisation* [Webb 1995]. Indeed this measure forms the basis of the standard implementation of MDS known as ALSCAL, which is included as part of the popular SPSS software package for statistical data analysis [Young and Harris 1990].

In practice, when comparing the results of STRESS and SSTRESS configurations, it should be noted that SSTRESS emphasises the fitting of large dissimilarities over small ones. As an example consider a particular dissimilarity δ_{ij} and its initial corresponding inter-point distance in the configuration space d_{ij}. If $\delta_{ij} = 10$ and $d_{ij} = 8$, then the resulting contributions to STRESS and SSTRESS will be 2^2 and 36^2 respectively. If however $\delta_{ij} = 20$ and $d_{ij} = 18$, the resulting contributions will be 2^2 and 76^2 respectively. Furthermore, as will be shown in the next chapter, the use of SSTRESS produces very different configurations from those of STRESS when the dissimilarity matrix is derived from unstructured data in a high-dimensional space.

2.4.3 NEUROSCALE

Although the techniques considered so far in this section are grouped under the heading of "topographic mappings", they are not actually mappings in the strict mathematical sense of the word. This is because there is no *transformation* defined from the input space to the map space. The result of Sammon's Mapping for example, is a set of configuration points which forms the low-dimensional representation of the high-dimensional input data. However there is no mechanism for the projection of new or *unseen* data without expensively regenerating the entire configuration with the new data points included in the original dataset. Thus Sammon's Mapping is best viewed as a "look-up" table in which only the original training data points are mapped to lower-dimensional configuration points.

The NEUROSCALE technique is a novel neural network implementation of Sammon's Mapping which provides a transformation from the data space to the map space, and thus the ability to project new data. The technique is effected by a feed-forward radial basis function

[3] Traditionally STRESS and SSTRESS are defined with a square root over the double summation. However it is dropped here for convenience since minimisation of either form (ie. with or without the square root) leads to the same solution.

(RBF) network which transforms the p-dimensional input space into a q-dimensional map or feauture space (where $q < p$). In common with Sammon's Mapping, NEUROSCALE allows for the incorporation of additional subjective (or supervisory) information into the training process in order to produce an enhanced feature space. The general NEUROSCALE error or stress measure is thus defined by:

$$E_{ns} = \sum_i \sum_{j>i} (\delta_{ij} - d_{ij})^2$$

where:

$$\delta_{ij} = (1 - \alpha) d^*_{ij} + \alpha s_{ij} \qquad \text{and} \qquad 0 \leq \alpha \leq 1$$

As before, d^*_{ij} and d_{ij} are the inter-point distances in the input space and the feature space respectively, and s_{ij} is the subjective metric whose influence is controlled by the parameter α. For any particular application of NEUROSCALE (eg. data visualisation, feature extraction, etc), the optimal value of this parameter must be determined through trial and error.

Training of a NEUROSCALE model involves adjusting the RBF parameters in order to minimise the error measure E_{ns}. Once the parameters of the hidden units have been determined (for example - as the parameters of the basis functions in a Gaussian Mixture Model trained on the data), the output layer weights can be adjusted using a *relative supervision* training algorithm [Lowe 1993]. To train the network, pairs of input points are presented, and while there is no corresponding pair of explicit target vectors, there is a measure of *relative* error available and this can be used to minimise the error E_{ns}.

One particularly efficient implementation of this form of training algorithm is known as *Shadow Targets*. The algorithm works by first computing a set of estimated or shadow targets based on the current map points $\mathbf{y}_i$ and the error derivatives $\frac{\partial E}{\partial \mathbf{y}_i}$. These targets are used as the desired network outputs, as with a supervised problem, and the network weights are then found. This procedure is repeated iteratively, each time a new set of shadow targets is computed and the corresponding network weights are found; until a minimum of E_{ns} has been reached. The algorithm effectively decomposes the training process into two steps, one of which is linear and can be computed efficiently. Indeed it has been shown that Shadow Targets is an order of magnitude more efficient at reaching a minimum of E_{ns} than other non-linear optimisation algorithms [Tipping 1996, Chapter 7].

One important property of NEUROSCALE is that it has excellent *generalisation* properties and furthermore it is relatively insensitive to model complexity. Both these properties result from the use of the relative supervision training algorithm. In particular the NEUROSCALE model naturally tends to adopt a solution with low curvature, and hence the trained network produces a smooth transformation which gives rise to improved generalisation. In addition this training algorithm automatically incorporates a regularising component which reduces the

sensitivity of the transformation to the complexity of the model (ie. the number of hidden units).

An alternative neural network implementation of Sammon's Mapping, and one which is particularly popular in the pattern recognition literature, is SAMANN [Mao and Jain 1995]. This model uses a multi-layer perceptron (MLP) to effect the transformation from the data space to the feature space. However it suffers from a number of drawbacks compared with the NEUROSCALE model. Firstly the training of SAMANN involves a back-propagation step in order to compute the error derivatives with respect to the weights. When on-line learning is used then $N(N-1)$ back-propagations are required for $N(N-1)/2$ pairs of input points. Alternatively, for a model linear in the weights such as an RBF network, these derivatives can be derived directly in a straightforward fashion. Another disadvantage of the SAMANN model is that the use of sigmoidal output units in the MLP requires that the training set of inter-point distances be scaled such that all the values are in the range [0,1]. This can be achieved by dividing through by the largest inter-point distance value. However if a new unseen data point is presented to the trained SAMANN network it could be projected incorrectly if its distance to a point in the original dataset is larger than any of the original inter-point distances. One solution to this problem is to use linear output units in the network. However it was shown by de Ridder and Duin that the training of such networks does not converge well, even if the weights are initialised such that the network performs a PCA projection of the original dataset [de Ridder and Duin 1997]. As an alternative the authors recommend training a network on the result of a Sammon Mapping of the original data. However such *a posteriori* training has been shown to lead to networks with high curvature and poor generalisation performance [Tipping 1996, Chapter 6].

NEUROSCALE is a highly effective and flexible neural network implementation of Sammon's Mapping which defines a transformation from the input space to the map space. Recent studies of neural network based feature extraction methods [Lerner et al. 1999] have tended to focus on the SAMANN model, suggesting that NEUROSCALE has yet to be explored fully by the pattern recognition community at large.

2.5 Conclusions

This chapter has described the established techniques for multivariate data projection, incorporating methods from both the fields of statistical pattern analysis and neural computing. These techniques have been grouped according to the emphasis they place on the preservation of the underlying geometric structure in the data. In addition the techniques may also be categorised according to their *linear-nonlinear* and *supervised-unsupervised* nature.

Figure 2.8 shows a taxonomy of the techniques covered in this chapter, based on the aforementioned categories. Metric and Nonmetric MDS are categorised as unsupervised and supervised respectively; the former considers the dissimilarity data to be representative of spatial distance measurements, whereas the latter treats the data as being of a more subjective nature. Sammon's Mapping is classed in its traditional sense as an unsupervised technique (although as previously noted it is possible to include supervisory information into the mapping process). NEUROSCALE which can incorporate varying degrees of *supervisory information* depending on the value of the parameter α, spans both the supervised and unsupervised domains.

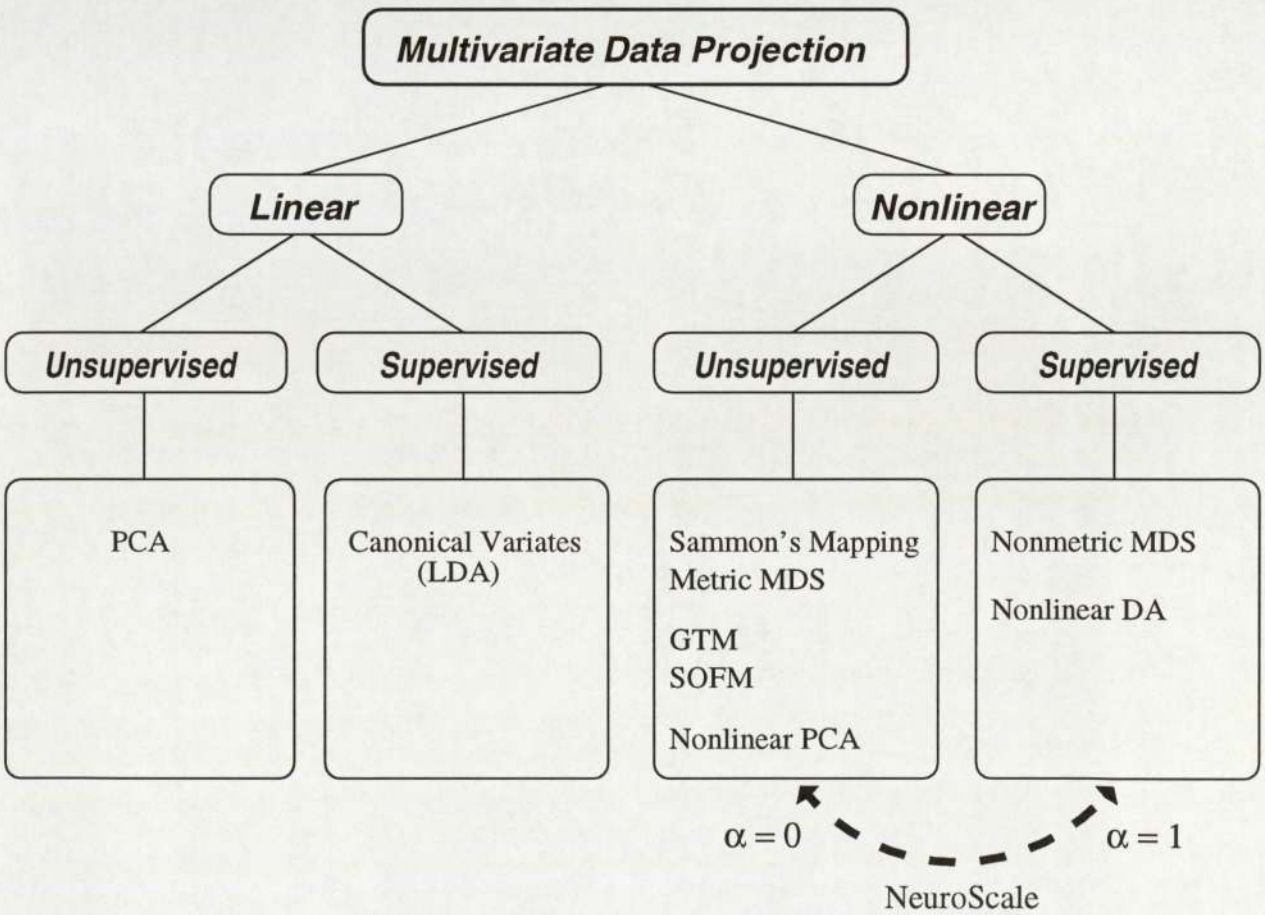


Figure 2.8: A taxonomy of the different approaches to multivariate data projection.

Chapter 3

Investigation into Artefactual Structure

3.1 Introduction

The main emphasis of the techniques considered up to this point has been on the degree to which the underlying geometric structure in the data is preserved. The previous chapter detailed the established methods for the projection of high-dimensional data and the extent to which they can be considered *topographic*. In particular it was noted that mappings which minimise simple STRESS measures of the form $\sum_{ij}(d^*_{ij} - d_{ij})^2$, optimally preserve the inherent structure in the data through the retention of spatial distance relationships, on both a local and a global scale.

However, if a mapping technique is to be used in practice as a tool for the exploration and visualisation of high-dimensional datasets, then it is important that a level of confidence can be placed on the knowledge that is derived from the resulting low-dimensional configurations. This can be achieved if the following two conditions are met. Firstly, as previously noted, the technique should produce a visualisation space whose geometric structure reflects that of the original data space. Secondly, any structure which is observed in the visualisation space should be truly *representative* of similar structure in the data space. This latter point can also be expressed as the requirement that the mapping should be free from *artefacts*, ie. the occurence of structure in the map data which is not present in the original input data.

This chapter presents a series of investigations into the problem of artefactual structure in topographic mappings. The problem is investigated both experimentally and theoretically, and a variety of different topographic constraints are considered. The insights gained from

this work are then used to examine the use of topographic mappings in the analysis of neuroanatomical connection data, and in particular the spatial organisation of the primate visual system.

3.2 Topographic Mappings of Unstructured Data

This section presents the experimental work which was performed in order to investigate the artefactual structure problem. Initially four different datasets were produced each of which characterised *unstructured* data of a particular dimension. For each dataset, 1000 points were drawn independenly from a uniform random distribution on the interval $[0, 1]^p$, where p indicates the dimension of the input space - for which values of 5, 10, 30, and 100 were chosen. Thus each of the four datasets represented 1000 points randomly[1] scattered inside a unit hypercube of dimension p.

A Sammon mapping of each dataset was then performed as follows. Firstly an initial set of map points was generated by randomly sampling 1000 points from within a unit square. Then, for a given error (or stress) measure, the set of map points was iteratively adjusted through the use of the conjugate gradients optimisation algorithm. When the solution had stabilised (or the maximum number of iterations had been reached), the final error value was recorded. This procedure was repeated fifty times (each time with a different initial configuration of the map points) and the configuration with the lowest final error value was chosen. In this way it was hoped that the resulting configuration would not represent a solution corresponding to a sub-optimal local minimum.

The remainder of this section shows the resulting Sammon maps for the various error measures used and discusses the nature of the spatial configurations derived from unstructured data of different dimensions. Appendix A details the error gradient $\frac{\partial E}{\partial \mathbf{y}_i}$ for each of the error measures used.

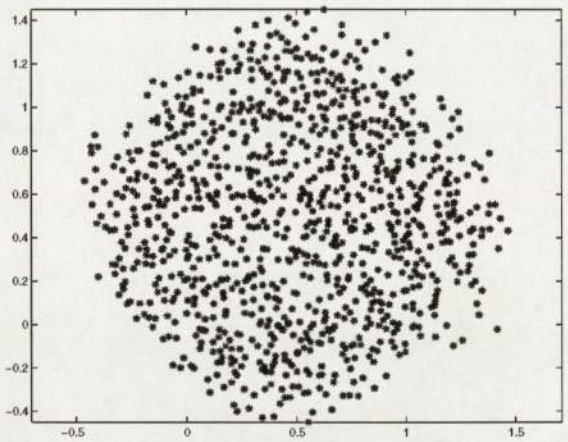
3.2.1 STRESS Based Mappings

Recall that the standard STRESS measure (or Sammon STRESS) is defined as:

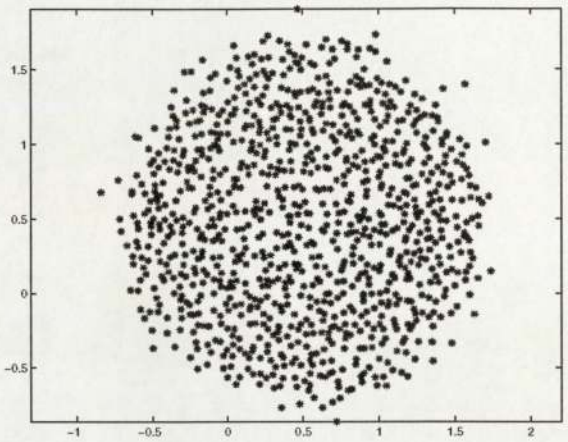
$$\text{STRESS} = \sum_i \sum_{j>i} \left(d^*_{ij} - d_{ij}\right)^2$$

where d^*_{ij} and d_{ij} are the Euclidean distances between points i and j in the input space and the map space respectively.

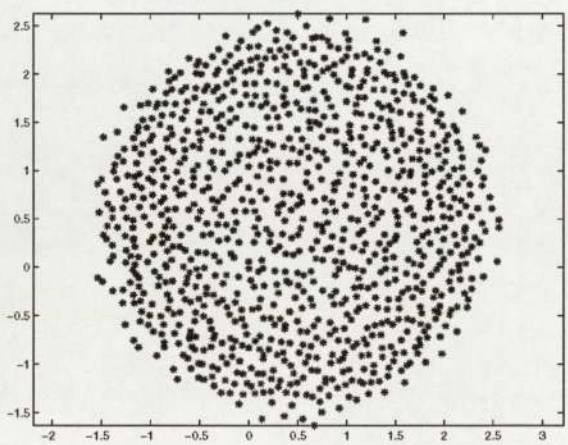
[1] Strictly, the distribution of the points was pseudo-random.



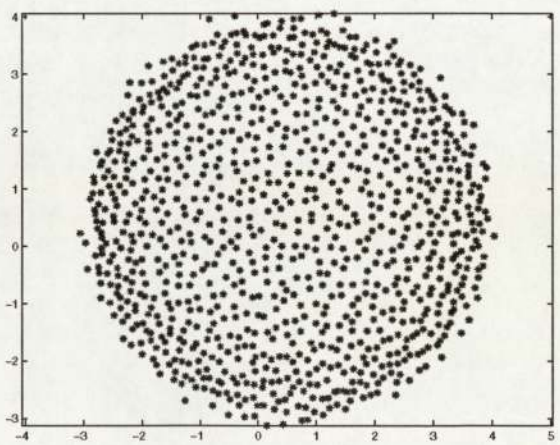
(a) 5-dimensional input data



(b) 10-dimensional input data

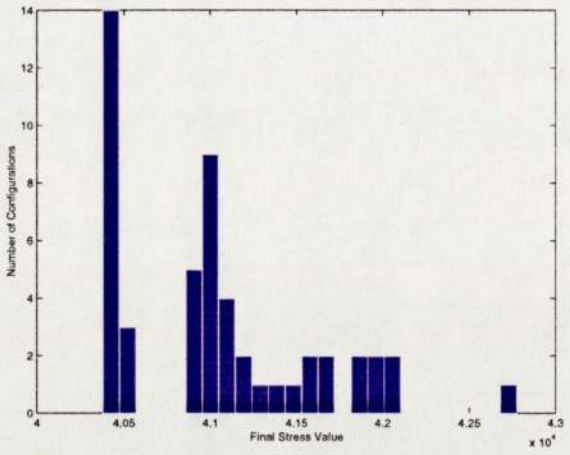


(c) 30-dimensional input data

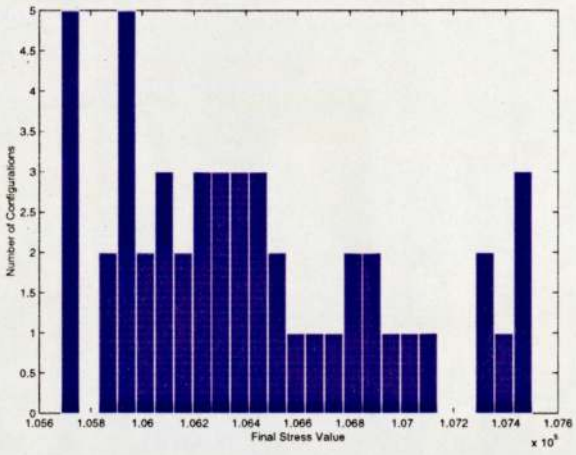


(d) 100-dimensional input data

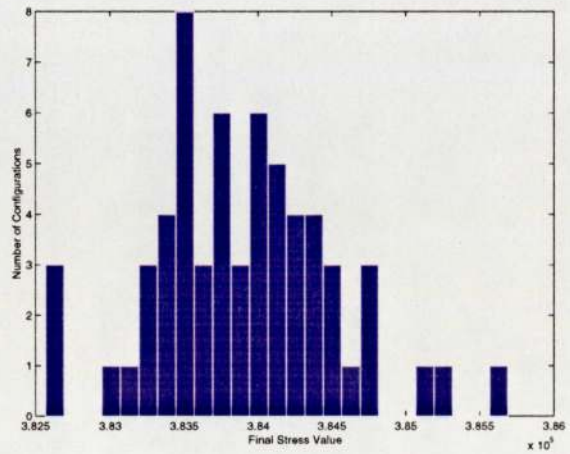
Figure 3.1: Maps produced with the STRESS measure from uniformly randomly distributed data of different dimensions.



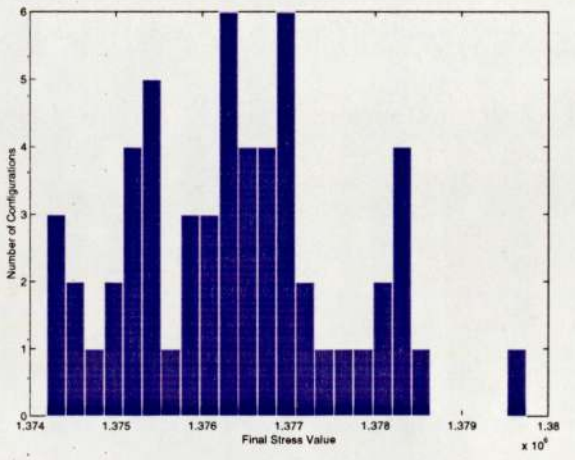
(a) 5-dimensional input data



(b) 10-dimensional input data



(c) 30-dimensional input data



(d) 100-dimensional input data

Figure 3.2: Histograms showing the number of configurations with a given final STRESS value for each of the datasets used.

Figure 3.1 overleaf shows the resulting minimum STRESS maps for each of the datasets and figure 3.2 shows the histograms of the final STRESS values over the fifty runs.

The resulting map plots indicate that the map points are approximately located inside a ball or a disc, and as the dimension of the input data increases the shape of the resulting configuration becomes more "disc like". This is interesting since it might naively be expected that a mapping of data randomly distributed within a hypercube would result in a configuration of map points randomly distributed within a square. Indeed this would appear to be the most intuitive and perhaps more importantly, the most informative configuration possible given the nature of the input data.

An insight into this effect can be gained by considering the spatial arrangement of the input data space. Firstly, as the dimension p of the input space increases, the number of corners of the resulting hypercube increases exponentially or, to be precise, as 2^p. If the dimension of the map space is q (where $q \ll p$) then there will be an inevitable distortion of the data during the mapping procedure due to the lack of dimensions to capture the spatial arrangement of the input data. In particular it was found experimentally that a mapping of only the corners of a high-dimensional hypercube produced a circular configuration in the map space. This is not surprising however, since any hypercube of side d can be positioned perfectly within a hypersphere of radius $\frac{d}{2}\sqrt{p}$ so that each corner of the hypercube just touches the surface of the hypersphere. Hence points lying on the corners of a hypercube can also be considered to lie on the surface of the matching hypersphere. A mapping of these points to a two-dimensional space therefore gives rise to a circular configuration. Thus the optimal solution (in terms of the minimisation of STRESS) is to position the map points corresponding to input points near to the corners of the hypercube, close to the circumference of a disc in the map space.

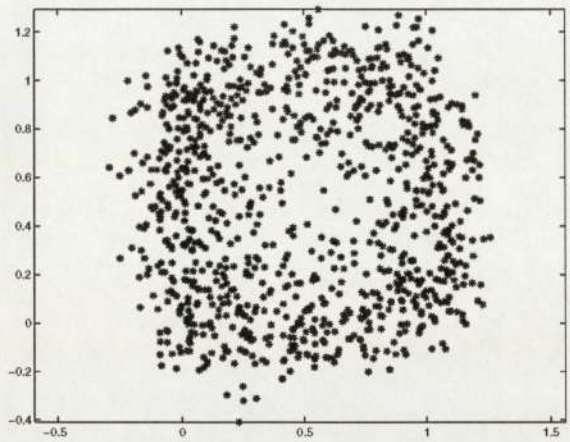
3.2.2 SSTRESS Based Mappings

As described in Section 2.4.2, the SSTRESS measure is defined as:

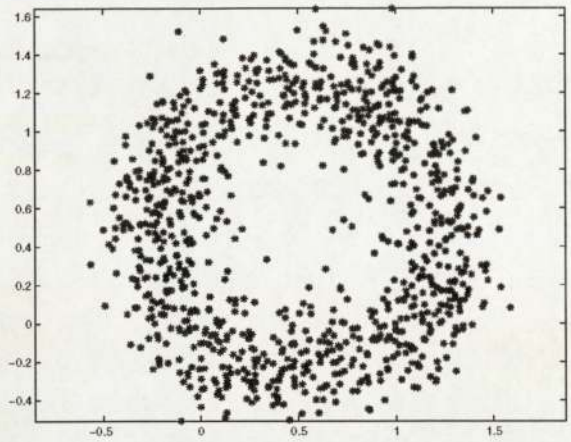
$$\text{SSTRESS} = \sum_i \sum_{j>i} \left(d_{ij}^{*2} - d_{ij}^2\right)^2$$

where d_{ij}^* and d_{ij} are the *squared* Euclidean distances between points i and j in the input space and the map space respectively.

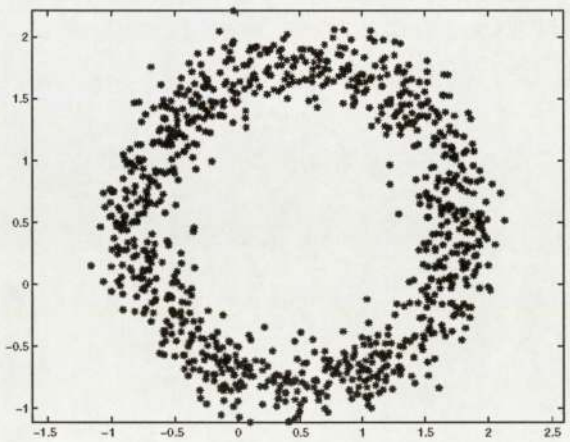
Figure 3.3 overleaf shows the resulting minimum SSTRESS maps for each of the datasets and figure 3.4 shows the histograms of the final SSTRESS values over the fifty runs. It is clear from the resulting maps that significant artefactual structure is produced by the SSTRESS mappings of uniform random data. As the dimension of this data increases, a noticeable ring structure is observed which becomes more well-defined with the increasing dimension.



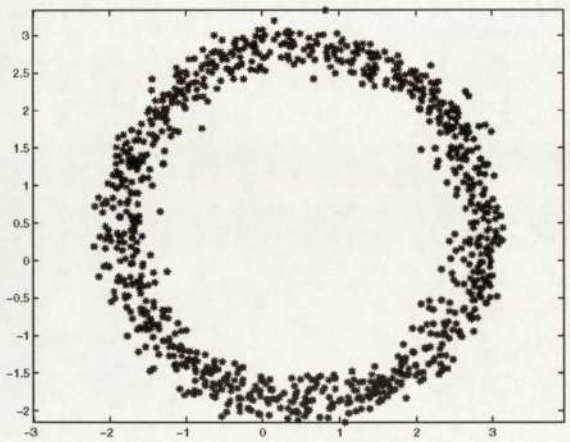
(a) 5-dimensional input data



(b) 10-dimensional input data

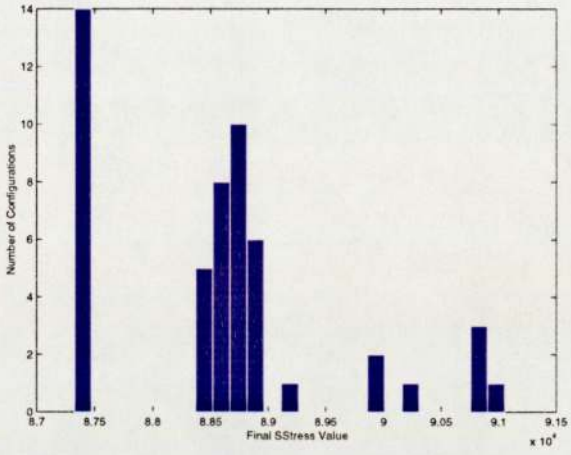


(c) 30-dimensional input data

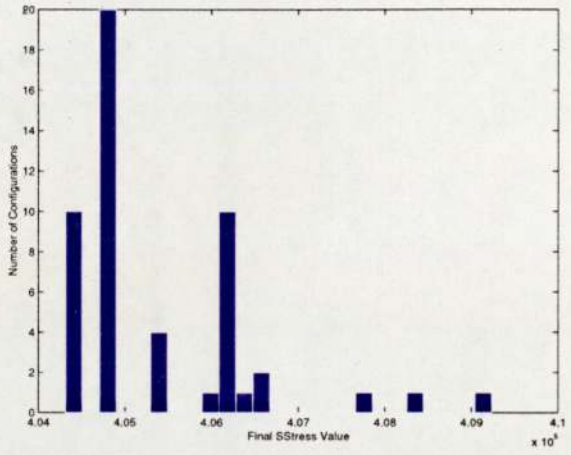


(d) 100-dimensional input data

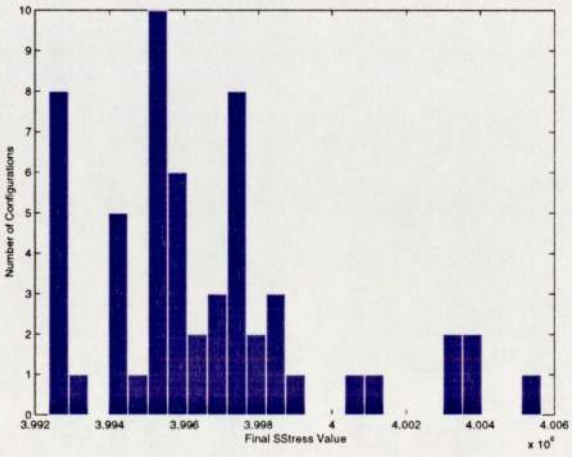
Figure 3.3: Maps produced with the SSTRESS measure from uniformly randomly distributed data of different dimensions.



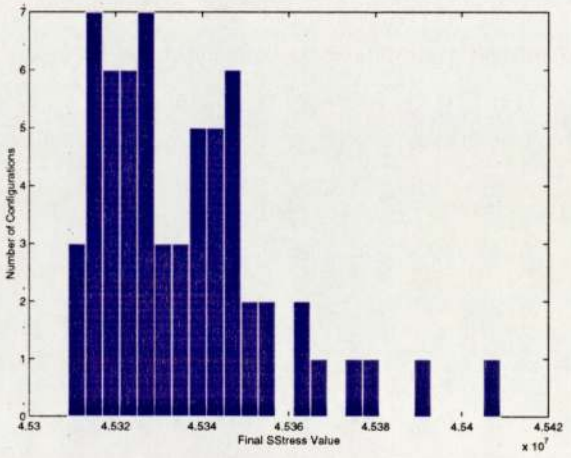
(a) 5-dimensional input data



(b) 10-dimensional input data



(c) 30-dimensional input data



(d) 100-dimensional input data

Figure 3.4: Histograms showing the number of configurations with a given final SSTRESS value for each of the datasets used.

For 5-dimensional input data the configuration looks almost random, with just some slight clustering of the map points away from the centre of the configuration. However for 10-dimensional input data the map points are clearly clustered in a ring shape, although there is an overall fuzziness to the structure. For 100-dimensional input data however this fuzziness has disappeared and the resulting configuration is sharp and annular in nature.

Clearly then, the use of the SSTRESS measure gives rise to a significant artefactual structure when used to map the high-dimensional random datasets considered in this chapter. In addition the greater the dimensionality of the dataset the greater the degree of artefactual structure observed. This effect is known as *dimensionality mismatch* and refers to the general observation that the distortion of a map increases with the difference or mismatch between the dimension of the data space and the dimension of the map space.

The histograms displaying the number of configurations with a given final SSTRESS value indicate that the majority of the solutions correspond to low SSTRESS configurations. Thus the training of SSTRESS maps on unstructured data does not appear particularly prone to sub-optimal local minima and therefore any single SSTRESS mapping of high-dimensional random data is likely to suffer from artefactual structure.

3.2.3 Mappings with Different Powers of the Euclidean Distance

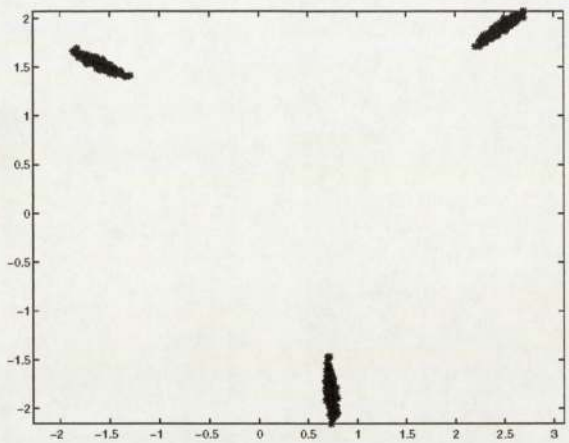
Since the SSTRESS measure utilises a *squared* Euclidean distance metric, it is worthwhile investigating if other powers of the Euclidean distance metric give rise to similar artefactual structure.

The SSTRESS function can be generalised to any arbitrary power n of the distance metric, to give:

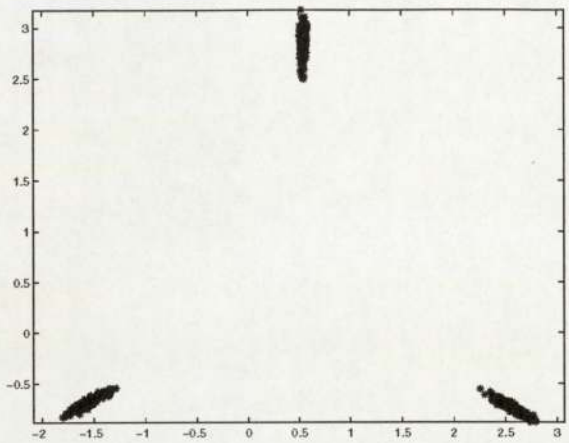
$$\text{SSTRESS}_n = \sum_i \sum_{j>i} \left(d_{ij}^{*n} - d_{ij}^{n}\right)^2$$

Figure 3.5 shows the resulting maps for $n = 3$, 4, 5, and 6 trained on the 100-dimensional uniform random dataset. For powers 3, 4 and 5 of the Euclidean distance metric, the configurations produced are very similar and each indicates that approximately equal numbers of the 1000 points are located on the corners of an equilateral triangle in the map space. Since any configuration is invariant (in terms of the resulting error) under rotation, reflection and translation, the final orientation of the configurations is not relevant to this analysis.

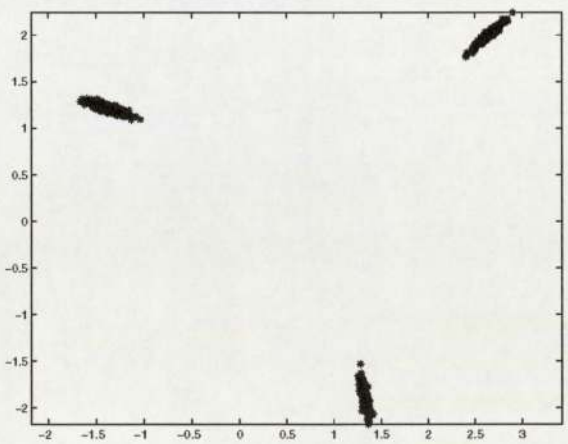
One possible explanation for this tendency to adopt an equilateral triangle configuration, is that such a grouping of the map points results in only two unique values for the inter-point spatial distances in the map space. The points are either zero apart (assuming they are located at one corner exactly), or they are a distance l apart - where l represents the



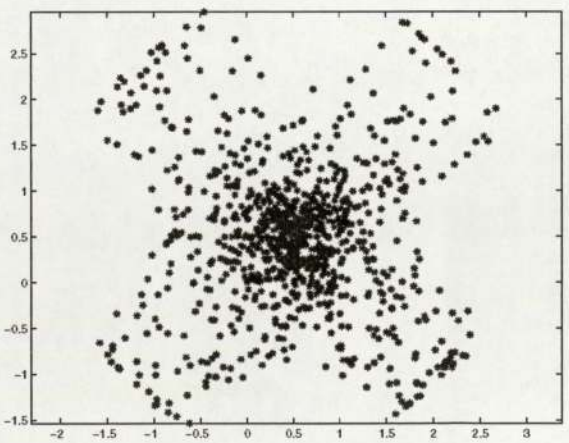
(a) Euclidean Distance Power 3



(b) Euclidean Distance Power 4



(c) Euclidean Distance Power 5



(d) Euclidean Distance Power 6

Figure 3.5: Maps produced with different powers of the Euclidean distance metric for the 100-dimensional uniform random dataset.

length of a side of the equilateral triangle (measured with the given metric). Although the configuration will result in a large error for points whose inter-point distance in the data space lies away from these two extremes, it is possible that the reduction in the error for points which are very close together ($d_{ij}^{*n} \approx 0$) or very far apart ($d_{ij}^{*n} \approx l$) outweighs this error; and hence this configuration provides an effective minimum error solution.

For the power 6 Euclidean distance metric, the configuration changes and adopts a cross shape with some significant clustering in the centre. As the power of the Euclidean metric increases then, at some point the equilateral triangle configuration becomes unfavourable in terms of the minimisation of the error measure.

3.2.4 Mappings with Minkowski Metrics

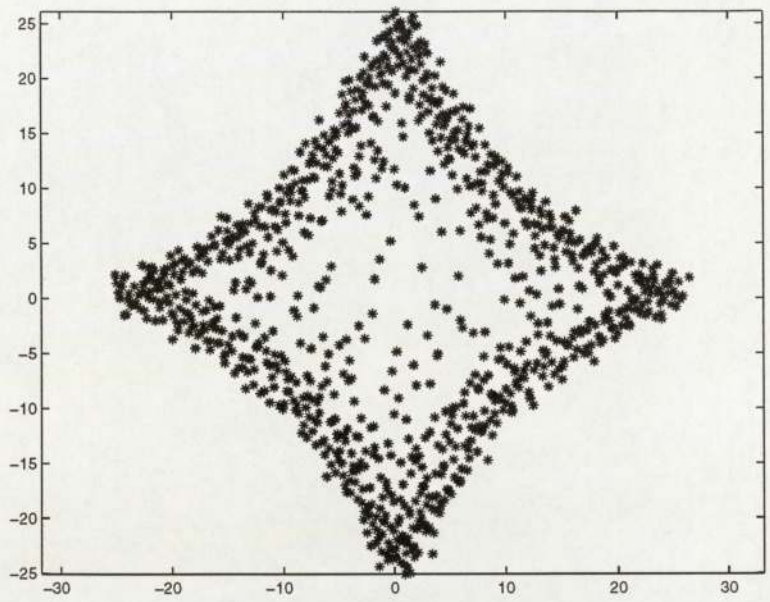
For a d-dimensional feature vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, the Minkowski metric giving the distance between $\mathbf{x}_i$ and $\mathbf{x}_k$ is defined as:

$$d(i,k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{\frac{1}{r}} \qquad \text{where } r \geq 1$$

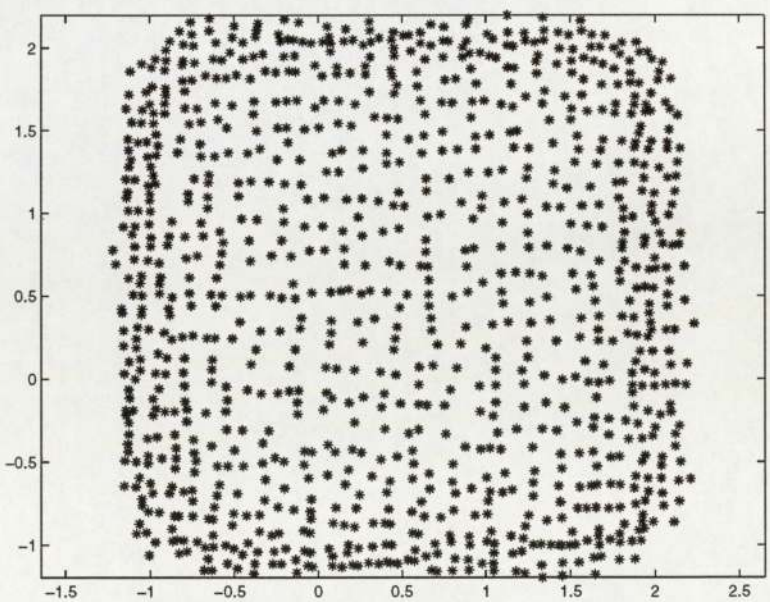
For $r = 1$ the metric is simply the sum of the distances (between the two points) along each axis and is known as the Manhattan or city-block distance. For $r = 2$ the metric is equal to the standard Euclidean distance.

Figure 3.6 overleaf shows the resulting maps for Minkowski metrics of $r = 1$ and $r = 3$ trained on the 100-dimensional uniform random dataset. The configuration for $r = 1$ indicates that the points are clustered around the edges of a curved diamond shape. However for $r = 3$ the resulting configuration shows that the map points are approximately randomly distributed inside a square (although there is some slight clustering around the edges). This is interesting since it represents the most intuitive and indeed the most informative configuration when the input points are randomly distributed within a high-dimensional hypercube.

However, there is a fundamental objection to the use of non-Euclidean distance metrics in data visualisation algorithms. Since the observer of the map is limited to viewing the map space from a strictly Euclidean standpoint, the topographic structure "seen" by the observer does not reflect that defined by the inter-point distances (as measured by the metric used). This problem stems from the simple fact that human beings are only capable of seeing through "Euclidean eyes". Thus from a purely theoretical stance, the standard Euclidean metric is to be preferred for data visualisation.



(a) Minkowski Metric: r = 1 (Manhattan Distance)



(b) Minkowski Metric: r = 3

Figure 3.6: Maps produced with two different Minkowski metrics for the 100-dimensional uniform random dataset.

3.3 Theoretical Analysis of Artefactual Structure

This section presents a theoretical analysis of the artefactual structure problem. The low-dimensional map configuration is considered to be produced by a mapping trained with uniformly randomly distributed input data on the SSTRESS measure. The set of map points is defined by a matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^{\mathrm{T}}$ and the set of input data points by a matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^{\mathrm{T}}$.

We begin by first defining the SSTRESS error measure:

$$E = \sum_i \sum_{j>i} \left(d_{ij}^{*2} - d_{ij}^2\right)^2$$

Differentiating E with respect to a particular map vector $\mathbf{y}_i$ we obtain:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -4 \sum_{j \neq i} (d_{ij}^{*2} - d_{ij}^2)^2 \, (\mathbf{y}_i - \mathbf{y}_j) \tag{3.1}$$

For the Euclidean distance metric $\| \ldots \|$, we have:

$$d_{ij}^{*2} = \| \mathbf{x}_i - \mathbf{x}_j \|^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j) = \| \mathbf{x}_i \|^2 + \| \mathbf{x}_j \|^2 - 2\, \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \tag{3.2}$$

$$d_{ij}^{2} = \| \mathbf{y}_i - \mathbf{y}_j \|^2 = (\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_j) = \| \mathbf{y}_i \|^2 + \| \mathbf{y}_j \|^2 - 2\, \mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j \tag{3.3}$$

Substituting (3.2) and (3.3) into (3.1) and rearranging, gives:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -4 \sum_{j \neq i} \left[(\| \mathbf{x}_i \|^2 - \| \mathbf{y}_i \|^2) + (\| \mathbf{x}_j \|^2 - \| \mathbf{y}_j \|^2) + 2\, (\mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j - \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j) \right] (\mathbf{y}_i - \mathbf{y}_j)$$

This equation can be expanded into a number of distinct terms, given by:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -4 \sum_{j \neq i} \left(\| \mathbf{x}_i \|^2 - \| \mathbf{y}_i \|^2 \right) \mathbf{y}_i \tag{3.4}$$

$$+ 4 \sum_{j \neq i} \left(\| \mathbf{x}_i \|^2 - \| \mathbf{y}_i \|^2 \right) \mathbf{y}_j \tag{3.5}$$

$$- 4 \sum_{j \neq i} \left(\| \mathbf{x}_j \|^2 - \| \mathbf{y}_j \|^2 \right) \mathbf{y}_i \tag{3.6}$$

$$+ 4 \sum_{j \neq i} \left(\| \mathbf{x}_j \|^2 - \| \mathbf{y}_j \|^2 \right) \mathbf{y}_j \tag{3.7}$$

$$- 8 \sum_{j \neq i} (\mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j) \, \mathbf{y}_i \tag{3.8}$$

$$+ 8 \sum_{j \neq i} (\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j) \, \mathbf{y}_i \tag{3.9}$$

$$+ 8 \sum_{j \neq i} (\mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j) \, \mathbf{y}_j \tag{3.10}$$

$$- 8 \sum_{j \neq i} (\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j) \, \mathbf{y}_j \tag{3.11}$$

We can simplify terms (3.4) and (3.8) - (3.11), as:

$$\sum_{j\neq i}\left(\|\mathbf{x}_i\|^2-\|\mathbf{y}_i\|^2\right)\mathbf{y}_i=(N-1)\left(\|\mathbf{x}_i\|^2-\|\mathbf{y}_i\|^2\right)\mathbf{y}_i$$

$$\sum_{j\neq i}\left(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j\right)\mathbf{y}_i=\sum_{j\neq i}\left(\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right)\mathbf{y}_j=\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\sum_{j\neq i}\left(\mathbf{y}_j\right)$$

$$\sum_{j\neq i}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\right)\mathbf{y}_i=\sum_{j\neq i}\left(\mathbf{y}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{x}_j=\mathbf{y}_i\mathbf{x}_i^{\mathrm{T}}\sum_{j\neq i}\left(\mathbf{x}_j\right)$$

$$\sum_{j\neq i}\left(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j\right)\mathbf{y}_j=\sum_{j\neq i}\left(\mathbf{y}_j\mathbf{y}_j^{\mathrm{T}}\right)\mathbf{y}_i$$

$$\sum_{j\neq i}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\right)\mathbf{y}_j=\sum_{j\neq i}\left(\mathbf{y}_j\mathbf{x}_j^{\mathrm{T}}\right)\mathbf{x}_i$$

Since we are considering the position of the map vector $\mathbf{y}_i$ at the minimum SSTRESS solution (ie. $\frac{\partial E}{\partial \mathbf{y}_i}=\mathbf{0}$), then we can divide through by $-4(N-1)$. This gives:

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{y}_i} \propto\ & \left(\|\mathbf{x}_i\|^2-\|\mathbf{y}_i\|^2\right)\mathbf{y}_i && (3.12)\\
& -\left(\|\mathbf{x}_i\|^2-\|\mathbf{y}_i\|^2\right)\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\right) && (3.13)\\
& +\frac{1}{N-1}\sum_{j\neq i}\left(\|\mathbf{x}_j\|^2-\|\mathbf{y}_j\|^2\right)\mathbf{y}_i && (3.14)\\
& -\frac{1}{N-1}\sum_{j\neq i}\left(\|\mathbf{x}_j\|^2-\|\mathbf{y}_j\|^2\right)\mathbf{y}_j && (3.15)\\
& +2\,\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\right) && (3.16)\\
& -2\,\mathbf{y}_i\mathbf{x}_i^{\mathrm{T}}\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{x}_j\right) && (3.17)\\
& -2\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\mathbf{y}_j^{\mathrm{T}}\right)\mathbf{y}_i && (3.18)\\
& +2\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\mathbf{x}_j^{\mathrm{T}}\right)\mathbf{x}_i && (3.19)
\end{aligned}$$

Now, since the value of the SSTRESS measure is only dependent on the inter-point distances, δ and d, we can centre the mean of the set of points $\mathbf{x}$ and $\mathbf{y}$ on the origin without affecting the overall geometric structure of the resulting map and hence the value of E. Mathematically

then, we have for large N:

$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{x}_j \simeq \mathcal{E}[\mathbf{x}] = \mathbf{0}$$
$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j \simeq \mathcal{E}[\mathbf{y}] = \mathbf{0}$$

where $\mathbf{0}$ represents the zero vector of the appropriate dimension and $\mathcal{E}[\cdot]$ the expectation operator. This leads to terms (3.13), (3.16) and (3.17) vanishing. In addition terms (3.18) and (3.19) simplify to give:

$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\mathbf{y}_j^{\mathrm{T}} \simeq \mathcal{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = cov(\mathbf{y},\mathbf{y})$$
$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\mathbf{x}_j^{\mathrm{T}} \simeq \mathcal{E}[\mathbf{y}\mathbf{x}^{\mathrm{T}}] = cov(\mathbf{y},\mathbf{x})$$

where $cov(\mathbf{a},\mathbf{b})$ represents the covariance matrix of the vectors $\mathbf{a}$ and $\mathbf{b}$. The expression governing the minimum SSTRESS solution is therefore:

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{y}_i} &\propto \left(\|\mathbf{x}_i\|^2 - \|\mathbf{y}_i\|^2\right)\mathbf{y}_i - 2\,cov(\mathbf{y},\mathbf{y})\,\mathbf{y}_i + 2\,cov(\mathbf{y},\mathbf{x})\,\mathbf{x}_i \\ &\quad + \frac{1}{N-1}\sum_{j\neq i}\left(\|\mathbf{x}_j\|^2 - \|\mathbf{y}_j\|^2\right)(\mathbf{y}_i - \mathbf{y}_j) \\ &= \mathbf{0}\end{aligned}$$

This represents a general formula for the value of the map vector $\mathbf{y}_i$ at the minimum SSTRESS solution, *regardless* of the nature of the input data $\mathbf{x}$. However in this analysis we are interested in the case where the input data is uniformly randomly distributed in a high-dimensional space.

First consider the covariance matrix $cov(\mathbf{y},\mathbf{y})$. If the input data $\mathbf{x}$ is uniformly randomly distributed about the origin, then the overall geometric structure of the set of map points $\mathbf{y}$ will be *isotropic*, since the problem is inherently symmetric about the origin. Therefore $cov(\mathbf{y},\mathbf{y})$ is *diagonal*, and equal to:

$$cov(\mathbf{y},\mathbf{y}) = \begin{bmatrix} \sigma_y^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_y^2 \end{bmatrix} \quad , \quad \sigma_y^2 = \frac{1}{N-1}\sum_i y_{i1}^2 = \frac{1}{N-1}\sum_i y_{i2}^2$$

Next consider the covariance of $\mathbf{y}$ and $\mathbf{x}$, given by $cov(\mathbf{y},\mathbf{x})$. If the dimension of the input vectors is p and the dimension of the map vectors is q, then $cov(\mathbf{y},\mathbf{x})$ is a matrix of dimension $q \times p$. Each element in the matrix represents the *degree of association* between particular

features of $\mathbf{y}$ and $\mathbf{x}$. Thus, we have:

$$cov(\mathbf{y},\mathbf{x}) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{q1}^2 & \sigma_{q2}^2 & \cdots & \sigma_{qp}^2 \end{bmatrix} \quad , \quad \sigma_{jk}^2 = \frac{1}{N-1}\sum_i y_{ij}x_{ik}$$

where for clarity σ^2 represents σ_{yx}^2. Since the input data $\mathbf{x}$ is uniformly randomly distributed about the origin and the set of map points is isotropic and centred on the origin, then the associativity between any two features of $\mathbf{y}$ and $\mathbf{x}$ is zero if $p \gg q$. Thus $cov(\mathbf{y},\mathbf{x})$ is equal to the zero matrix for high-dimensional input data.

Finally, consider the term:

$$\frac{1}{N-1}\sum_{j\neq i}\left(\|\mathbf{x}_j\|^2 - \|\mathbf{y}_j\|^2\right)(\mathbf{y}_i - \mathbf{y}_j)$$

This can be expanded to give:

$$\left(\frac{1}{N-1}\sum_{j\neq i}\sum_{k=1}^{p} x_{jk}^2 - \frac{1}{N-1}\sum_{j\neq i}\sum_{k=1}^{q} y_{jk}^2\right)\mathbf{y}_i \tag{3.20}$$

$$- \frac{1}{N-1}\sum_{j\neq i}\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\mathbf{y}_j \tag{3.21}$$

$$- \frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j\mathbf{y}_j \tag{3.22}$$

For large N, term 3.20 simplifies to:

$$\left(p\,\sigma_x^2 - q\,\sigma_y^2\right)\mathbf{y}_i$$

Assuming the input data is uniformly (and independently) randomly distributed along each coordinate axis and of a high dimensionality (ie. p is large), then term 3.21 simplifies to:

$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\mathbf{y}_j = \frac{p}{N-1}\sum_{j\neq i}\frac{1}{p}\left(\sum_{k=1}^{p} x_{jk}^2\right)\mathbf{y}_j \simeq p\,\sigma_x^2\left(\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j\right) = \mathbf{0}$$

Finally term 3.22 is the third order moment, or *skewness*, of the distribution of map points $\mathbf{y}$. This provides a measure of the asymmetry of the distribution about its mean. Thus if the map points are isotropic, then for large N we have:

$$\frac{1}{N-1}\sum_{j\neq i}\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j\mathbf{y}_j \simeq \mathcal{E}[\mathbf{y}^{\mathrm{T}}\mathbf{y}\mathbf{y}] = \mathbf{0}$$

Thus, under the conditions outlined above, the equation governing the minimum SSTRESS

solution is given by:

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{y}_i} &\propto \left(\|\mathbf{x}_i\|^2 - \|\mathbf{y}_i\|^2\right)\mathbf{y}_i - 2\sigma_y^2\,\mathbf{y}_i + \left(p\,\sigma_x^2 - q\,\sigma_y^2\right)\mathbf{y}_i \\ &= \left(\|\mathbf{x}_i\|^2 - \|\mathbf{y}_i\|^2 + p\,\sigma_x^2 - (q+2)\,\sigma_y^2\right)\mathbf{y}_i \\ &= \mathbf{0}\end{aligned}$$

Ignoring the trivial solution that $\mathbf{y}_i = \mathbf{0}$, then we have:

$$\|\mathbf{x}_i\|^2 - \|\mathbf{y}_i\|^2 + p\,\sigma_x^2 - (q+2)\,\sigma_y^2 = 0$$

Summing over all points i, gives:

$$\begin{aligned}&\sum_{i=1}^{N}\left(\|\mathbf{x}_i\|^2 - \|\mathbf{y}_i\|^2 + p\,\sigma_x^2 - (q+2)\,\sigma_y^2\right) = 0 \\ \Rightarrow\quad &\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2 - \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^2 = p\,\sigma_x^2 - (q+2)\,\sigma_y^2 \\ \Rightarrow\quad &q\,\sigma_y^2 + (q+2)\,\sigma_y^2 = 2p\,\sigma_x^2 \\ \Rightarrow\quad &\sigma_y^2 = \frac{p}{q+1}\,\sigma_x^2 \qquad (3.23)\end{aligned}$$

Thus the variance of the map points is related to the variance of the input points by a factor of $\frac{p}{q+1}$. Table 3.1 below shows the values of the observed and predicted variances for the two-dimensional configurations ($q = 2$) generated from data uniformly randomly distributed ($\sigma_x^2 = 0.0835$) in unit hypercubes of varying dimensions p, as displayed in figure 3.3. The results indicate that the accuracy of the above relationship between σ_y^2 and σ_x^2 improves as the dimension p of the input space increases. This is not surprising however since the proof assumes the input data to be of a high dimensionality.

Although equation (3.23) does not give an explicit form for the map configuration, it is possible to give a heuristic justification for the emergence of artefactual "annular" structure with random high-dimensional input data. Whilst (3.23) shows that the map variance must be very small relative to the data variance, it is also required that data points lying in opposite corners of the hypercube be kept far apart in the map space. One way then of balancing these two contrasting requirements is to position the map points onto a circle.

Dimension p	Number of Points N	σ_y^2 observed	σ_y^2 predicted	Percentage Error
5	1000	0.166	0.139	16.4%
10	1000	0.303	0.278	8.1%
30	1000	0.864	0.835	3.4%
100	1000	2.823	2.783	1.4%

Table 3.1: A comparison of the predicted and observed variances.

3.4 Extending the Results to Metric MDS

The various experiments considered so far have all utilised input data in the form of a set of high-dimensional data points. The distances between the individual data points are computed and the purpose of the mapping algorithm is to find a set of two-dimensional map points whose spatial configuration matches these inter-point distances as closely as possible. In this sense then, Sammon's Mapping is the most natural interpretation of the mapping algorithm.

As noted in Section 2.4.2, Metric MDS and Sammon's Mapping are closely related. Whereas Sammon's Mapping operates on a set of input vectors (which in turn describe a set of objects), Metric MDS is designed to work with proximity data which details the dissimilarities between the objects. Clearly it is possible to produce a dissimilarity matrix from a set of input vectors, simply by computing the inter-point distances, and then "use" Metric MDS to generate a low-dimensional configuration space. Although this will produce an identical solution to that resulting from a Sammon Mapping of the raw input data; the motivation behind MDS is to elucidate structure from data which does not live in an explicit data space. Thus it is more appropriate to think of performing Sammon's Mapping on data composed of input vectors and Metric MDS on data detailing proximity values.

Given this interpretation of the two techniques, it is desirable to see what insight can be gained about the nature of artefactual structure in Metric MDS from the results of the previous sections. It is useful then to first compare the form of the SSTRESS measure in both Sammon's Mapping and Metric MDS. This is given by:

$$\text{SSTRESS} = \sum_i \sum_{j>i} (\| \mathbf{x}_i - \mathbf{x}_j \|^2 - d_{ij}^2)^2 = \sum_i \sum_{j>i} (\delta_{ij}^2 - d_{ij}^2)^2$$

It has already been shown that if the vectors $\mathbf{x}$ are uniformly (or normally) randomly distributed in a high-dimensional space then the resulting SSTRESS mapping will exhibit a strong annular structure. Hence from an MDS perspective we can expect a similar result if the dissimilarities δ correspond to the inter-point distances between such vectors.

It is also useful to consider what interpretation can be given to the different distance metrics investigated so far, within a framework of MDS. For a particular distance metric $\| \ldots \|$, the SSTRESS function is given by:

$$\text{SSTRESS} = \sum_i \sum_{j>i} \left(\delta_{ij}^2 - d_{ij}^2\right)^2 = \sum_i \sum_{j>i} \left(\delta_{ij}^2 - \| \mathbf{y}_i - \mathbf{y}_j \|^2\right)^2$$

This function can be generalised to any arbitrary power n of the distance metric, to give the *generalised* SSTRESS function::

$$\text{SSTRESS}_n = \sum_i \sum_{j>i} \left(\delta_{ij}^n - d_{ij}^n\right)^2 = \sum_i \sum_{j>i} \left(\delta_{ij}^n - \| \mathbf{y}_i - \mathbf{y}_j \|^n\right)^2$$

where the dissimilarities δ are assumed to be spatial distances derived from the metric $\| \ldots \|$. An alternative interpretation however, is to consider this measure as being a special case of the standard STRESS function, given by:

$$\text{STRESS} = \sum_i \sum_{j>i} (\delta_{ij} - d_{ij})^2 = \sum_i \sum_{j>i} (\delta_{ij} - \| \mathbf{y}_i - \mathbf{y}_j \|)^2$$

where the class of possible metrics is now taken to include any valid distance metric raised to a power n. In this way then, the SSTRESS measure can be viewed as a particular form of STRESS - employing the squared Euclidean distance metric.

The advantage of this interpretation is threefold. Firstly it provides a more principled approach to the use of squared distances since they can be considered to be the result of using a *squared metric*, as opposed to an ad-hoc modification (to the STRESS function) to enable the optimisation to be performed by alternating least squares methods. Secondly it reduces the problem of Metric MDS to one of choosing just a distance metric (rather than a distance metric *and* an optimisation function), allowing the comparison of configurations produced by different metrics to be performed more easily. Finally no additional manipulation of the dissimilarity data is required (such as squaring, cubing, etc) since the dissimilarity values are assumed to be derived from the chosen distance metric.

It is worth noting that this interpretation does not arise in Sammon's Mapping since the spatial distances between the data points can be computed explicitly as $\| \mathbf{x}_i - \mathbf{x}_j \|$. With MDS, it is the fact that only the dissimilarities are known, and not the input vectors $\mathbf{x}$, that results in the various possible ways of interpreting the role of the metric in the procedure. This is an important but subtle distinction between the two techniques.

3.5 Application to Neuroanatomical Connection Data

3.5.1 Introduction to Brain Connectivity and Scaling

Research into neuroanatomy has established that the various gross structures of the brain are divided into a large number of different processing regions. Although the spatial position of these regions in the brain is now reasonably well known, the overall processing architecture defined by the inter-connection of the regions is less well understood. Knowledge of this processing architecture for a particular brain structure would reveal much important information about its operation and function. Therefore in recent years neuroanatomists have begun to catalogue a large number of the connections between the various regions in the brain. Such *connectional data*, as it is known, is often complex and uncovering structure in this data is an important and challenging problem.

One technique which has been used for such analysis is Multidimensional Scaling. In particular Nonmetric Multidimensional Scaling, or NMDS, has been applied to connection data derived from experimental investigations into the pattern of connections between regions in the macaque monkey visual cortex. This connection data typically details the *type* of the connection between two regions. Inherent in the use of Multidimensional Scaling to understand this data is the assumption that the type of the connection between two regions provides a measure of information about the proximity of these regions within the gross processing architecture of the structure under consideration. In particular, regions which are linked by a bi-directional or reciprocal connection are assumed to be closer together than regions which are linked by a one-way connection. Similarly, regions linked by a one-way connection are assumed to be closer together than regions which have no connection between them.

Given this assumption then, the role of Multidimensional Scaling in analysing brain connectivity data is to derive a two-dimensional configuration which reflects the proximity relationships contained within the data. Such a configuration would hopefully reveal the topological organisation of the different regions and the overall processing architecture of the structure.

However, there has been much recent controversy in the scientific literature as to the validity of the configurations derived with MDS, and in particular the extent to which the resulting maps are corrupted by artefactual structure. Given the preceding analysis of the artefactual structure problem and the results gained into the importance of the choice of distance metric in the error function, the aim of this section then is to investigate the configurations derived for one particular application of MDS to connectivity data (that of the primate cortical visual system) and research the extent to which previous results have been corrupted by artefactual structure.

3.5.2 Overview of Previous Work on the Primate Cortical Visual System

The first person to apply techniques from Multidimensional Scaling to connection data was Young in 1992. He applied NMDS, with the SSTRESS measure, to a matrix of connections between thirty areas of the macaque monkey visual cortex [Young 1992]. This matrix of connections (which forms the input dataset) is shown in table 3.2, where each value refers to a particular type of neuroanatomical connection [2]. The resulting two-dimensional configuration as derived from NMDS is shown in figure 3.7.

By connecting areas which have a one-way or a bi-directional connection (as given by the connection matrix) with a straight line on the configuration map, it is possible to reveal the details of the underlying topological organisation of the macaque visual cortex. When Young carried this procedure out, he found strong evidence for the "two-streams" hypothesis of visual processing. In particular, he concluded that, starting from area V1, visual information flows out in two highly distinct and hierarchically organised streams (known as the *dorsal* and the *ventral* streams), as defined by the division and the ordering of the areas in the two arms of the configuration. The two streams then reconverge in areas A46 and STPa, and the visual information is recombined here. In addition connections between

[2] The values in table 3.2 represent *similarity* values from which *dissimilarities* were generated as the input to the NMDS procedure.

	V1	V2	V3	Vp	V3a	V4	VOT	V4t	MT	MSTd	MSTl	FST	PITd	PITv	CITd	CITv	AITd	AITv	STPp	STPa	TF	TH	PO	PIP	LIP	VIP	DP	A7a	FEF	A46
V1	-	2	2	0	2	2	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0
V2	2	-	2	2	2	2	2	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
V3	2	2	-	1	2	2	0	2	2	2	0	2	0	0	0	0	0	0	0	0	2	0	1	2	2	2	0	0	0	0
Vp	0	2	1	-	2	2	2	0	2	2	0	1	0	0	0	0	0	0	0	0	2	0	1	2	1	1	0	0	0	0
V3a	2	2	2	2	-	2	0	0	2	2	2	2	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2	0	0	0
V4	2	2	2	2	2	-	1	2	2	0	0	2	2	2	2	2	0	2	0	0	2	2	0	2	2	0	2	0	0	1
VOT	0	2	0	2	0	1	-	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V4t	1	1	2	0	0	2	0	-	2	1	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
MT	2	2	2	2	2	2	0	2	-	2	2	2	0	0	0	0	0	0	0	0	0	0	1	2	2	2	0	0	1	1
MSTd	0	2	2	2	2	0	0	1	2	-	0	2	1	1	0	0	0	0	2	0	1	0	2	0	2	2	2	2	2	0
MSTl	0	2	0	0	2	0	0	1	2	0	-	2	0	0	0	0	0	0	2	0	0	0	2	0	1	2	1	0	2	0
FST	0	1	2	1	2	2	0	2	2	2	2	-	1	1	0	0	0	0	2	0	2	0	0	0	2	2	1	1	2	0
PITd	0	0	0	0	0	2	1	0	0	1	0	1	-	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1
PITv	0	0	0	0	0	2	1	0	0	1	0	1	0	-	1	2	1	2	0	0	1	1	0	0	1	0	0	0	1	1
CITd	0	0	0	0	0	2	0	0	0	0	0	0	0	1	-	0	1	1	1	0	0	1	0	0	0	0	0	0	2	2
CITv	0	0	0	0	0	2	0	0	0	0	0	0	1	2	0	-	2	2	1	0	1	1	0	0	0	0	0	0	2	2
AITd	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	-	0	0	1	1	1	0	0	0	0	0	0	2	2
AITv	0	0	0	0	0	2	0	0	0	0	0	0	1	2	1	2	0	-	0	0	2	2	0	0	0	0	0	0	0	0
STPp	0	0	0	0	0	0	0	0	0	2	2	2	0	0	1	1	0	0	-	2	2	2	0	0	0	0	0	2	2	2
STPa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	-	2	2	0	0	0	0	0	2	0	2
TF	0	0	2	2	0	2	0	0	0	1	0	2	0	1	0	1	1	2	2	2	-	0	0	0	1	0	0	2	0	2
TH	0	0	0	0	0	2	0	0	0	0	0	0	0	1	1	1	1	2	2	2	0	-	0	0	0	0	0	2	0	2
PO	2	1	1	1	1	0	0	1	1	2	2	0	0	0	0	0	0	0	0	0	0	0	-	1	2	2	2	2	1	0
PIP	2	1	2	2	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-	0	0	2	1	0	0
LIP	0	0	2	1	1	2	0	0	2	2	1	2	0	1	0	0	0	0	0	0	1	0	2	0	-	2	2	2	2	1
VIP	0	1	2	1	0	0	0	0	2	2	2	2	0	0	0	0	0	0	0	0	0	0	2	0	2	-	0	1	0	0
DP	0	0	0	0	2	2	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	2	2	2	0	-	2	1	2
A7a	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	2	0	2	2	2	2	2	1	2	1	2	-	2	2
FEF	0	0	0	0	0	0	0	0	1	2	2	2	1	1	2	2	2	0	2	0	0	0	1	0	2	0	1	2	-	2
A46	0	0	0	0	0	1	0	0	1	0	0	0	1	1	2	2	2	0	2	2	2	2	0	0	1	0	2	2	2	-

Table 3.2: A matrix of connections between areas of the macaque visual cortex. Connections coded as '2' represent reciprocal or bi-directional connections, those coded as '1' represent one-way connections (direction not indicated), and those coded as '0' represent connections which have been explicitly tested for and found absent or connections which are not presently known. In addition the matrix is symmetric.

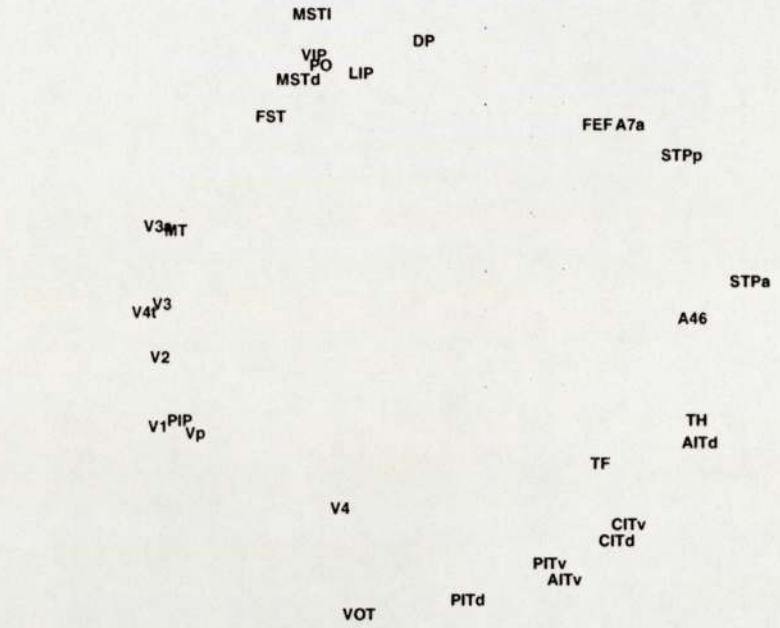


Figure 3.7: Young's configuration after applying NMDS (with SSTRESS) to the primate data.

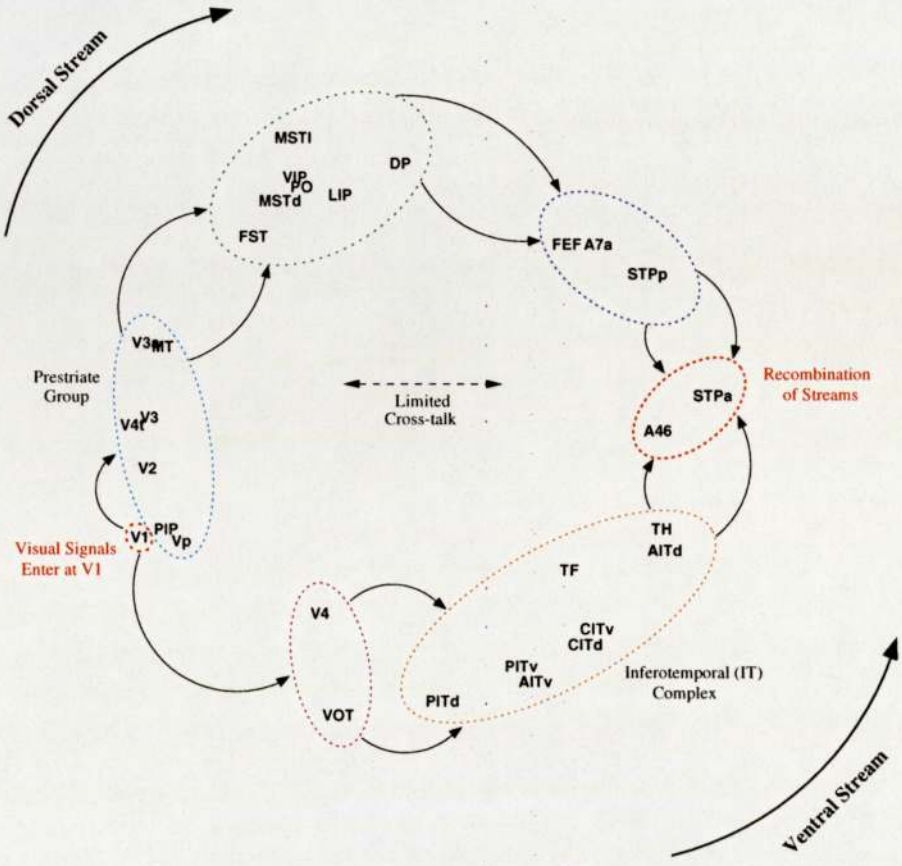


Figure 3.8: Support of the configuration for the "two streams" hypothesis.

the two streams are much less dense than those within each stream. These observations are illustrated in figure 3.8.

Of particular interest is the fact that the configuration has a strongly annular form. That such a shape is potentially indicative of artefactual structure in the configuration was first noted by Simmen, Goodhill and Willshaw. They produced NMDS representations of ternary dissimilarity matrices (ie. containing just the values 0, 1 and 2) in which the entries were assigned at random [Simmen, Goodhill, and Willshaw 1994]. As expected the resulting configurations revealed an annular form. A performance measure was then computed to provide some insight into the effectiveness of the configurations for capturing the underlying structure (or lack of) in the proximity data. This measure, termed RSQ, is the squared correlation between the input dissimilarities and the corresponding spatial distances in the configuration[3]. The values obtained with the configurations derived from random dissimilarity matrices were found to be slightly less than the value obtained with Young's configuration. Thus the authors concluded that although the visual system data is not entirely due to a random process, it is most likely that "Young's configuration reflects a mixture of both genuine and artefactual structure".

3.5.3 Analysis of the Organisation of the Primate Cortical Visual System

Given the results contained earlier in this chapter concerning the role of the distance metric in the derivation of configurations that exhibit artefactual structure, it is natural to see what insight can be gained into the degree of artefactual structure in Young's configuration. Although a thorough comparison is not possible, since the previous results apply strictly to Metric MDS, it is nevertheless worthwhile examining the configurations generated with Metric MDS as the nature of the connection data may allow for the dissimilarity values to be considered approximately at the interval level of measurement and thus suitable for analysis by metric techniques.

Configurations were therefore derived with Metric MDS using the standard Euclidean metric and also this metric raised to a number of different integer powers. For each particular metric, one thousand randomly initialised configurations were generated and these were then adjusted in order to minimise the error function as defined by the metric. The configuration with the minimum final error value was then chosen. The proximity matrix of dissimilarity values was produced by taking the matrix of connection values as given in table 3.2 and subtracting each value from the constant 2. In this way areas which were unconnected were assigned a dissimilarity of 2 and areas which were reciprocally connected were assigned a dissimilarity of 0 (the areas which were 'one-way' connected retained a value of 1).

[3] An RSQ of 1 indicates a perfect solution and a value of 0 a completely uncorrelated solution.

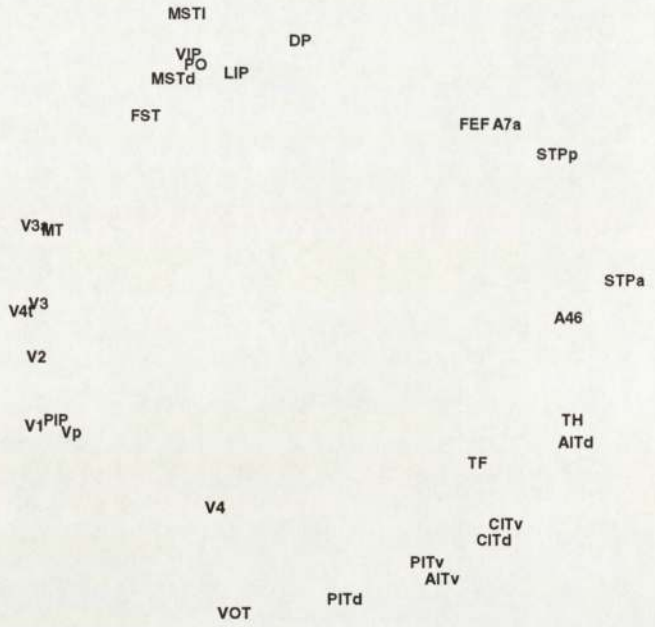
In order to accurately compare each of the resulting configurations with that derived by Young, it was necessary to linearly transform the configurations using a Procrustes Rotation [Mardia, Kent, and Bibby 1997, Chapter 14]. This technique provides a method for aligning two configurations in an optimal least-squares sense. The Procrustes method allows for three types of transformation, namely: rigid rotation, reflection, and isotropic scaling. If both configurations under comparison had been derived from a Metric MDS procedure, then this latter operation of scaling would not have been appropriate since the procedure explicitly matches the spatial distances with the dissimilarities. However since Young's configuration is the result of a nonmetric procedure (which only places importance on the *ordering* of the spatial distances), it could be isotropically scaled in the alignment process. Once the two configurations had been aligned, the residual sum-of-squares error (RSS), which gives a measure of the *goodness of fit* between the two configurations, was computed.

Figure 3.9 shows the resulting configuration derived from Metric MDS trained to minimise the SSTRESS function, together with Young's original configuration for ease of comparison. Since Young's configuration was generated from the ALSCAL implementation of NMDS, which is designed to minimise the SSTRESS measure, this is the most accurate comparison between the metric and nonmetric techniques. The metric configuration is very similar to Young's configuration and it displays the same annular structure and hierarchical formations. Some regions are slightly displaced in the metric configuration (eg. V4, TF, PO, VIP) but this is not unexpected since the metric technique is less flexible than its nonmetric counterpart. This result therefore represents a reasonable justification for a comparison of configurations derived from metric techniques with the configuration Young derived with NMDS. The RSS value for the two configurations was 1.352.

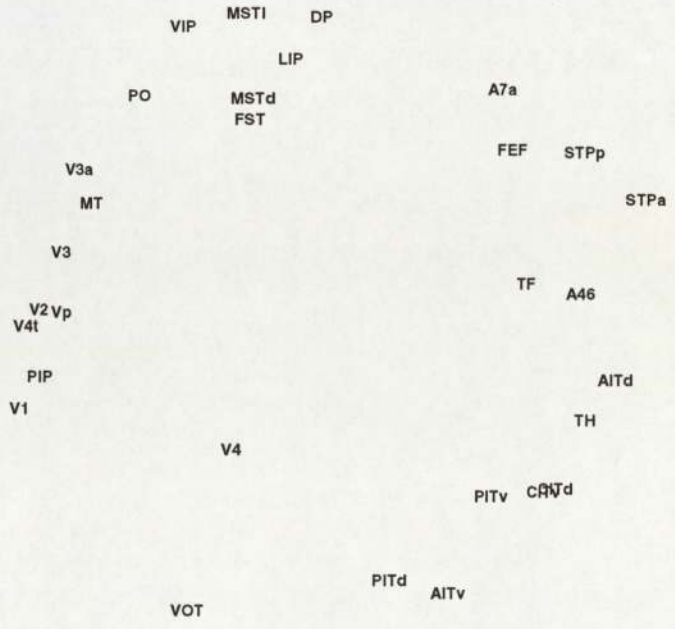
Figure 3.10 shows the resulting configuration derived from Metric MDS trained to minimise the STRESS function. As expected the configuration is less annular in nature and the tight clustering of the hierarchies observed in Young's configuration is less evident here. In addition there is more opportunity for cross-talk across the centre of the configuration, particularly between the regions Vp, V4 and TF, which are all connected reciprocally. The RSS value between this configuration and Young's was 2.606.

Figure 3.11 shows the configurations derived with the Euclidean distance metric raised to the powers 3, 4, 5, and 6. For the powers 3 - 5 of the metric, the regions begin to cluster near to the corners of an equilateral triangle, as observed more directly with random input data in figure 3.5. For the power 6 metric, the region A46 is centred whilst the other regions tend towards the corners of the triangle. Due to the low number of regions and hence configuration points ($N = 30$), an accurate comparison with the maps of figure 3.5 is not possible. However the general structure of these configurations does appear to indicate the presence of a random component in the primate visual cortex connection data.

Overall these results indicate that there is a strong possibility that Young's configuration contains a degree of artefactual structure and this artefactual structure manifests itself in the annular nature of the configuration. In addition the use of the STRESS measure with the standard Euclidean distance metric is likely to result in a more accurate spatial configuration for this problem, and this configuration provides less favourable evidence for the "two streams" hypothesis of visual processing than Young's SSTRESS derived configuration. The use of the NMDS technique with the standard Euclidean distance metric would provide further insight into this problem and is an area for further research.

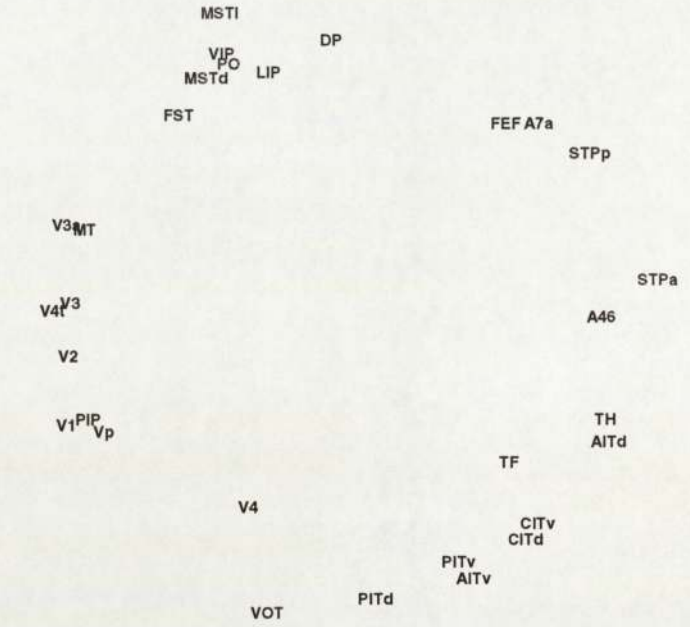


(a) NMDS Configuration with SSTRESS

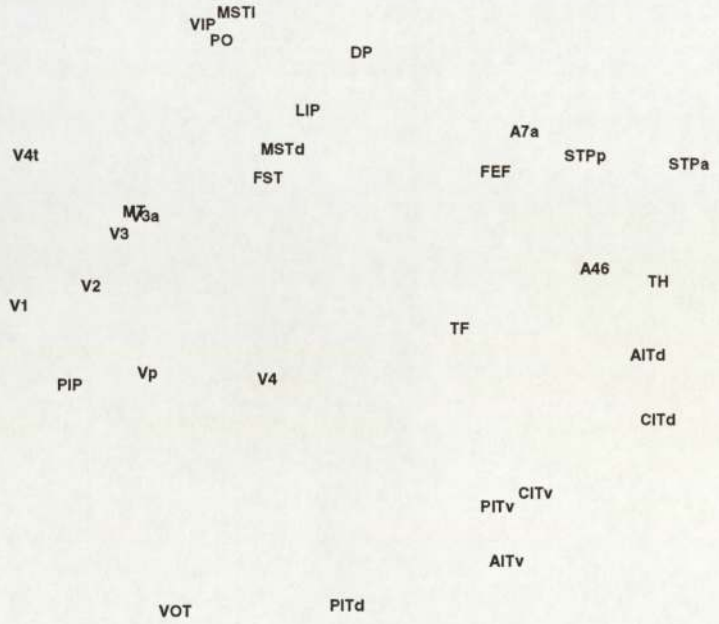


(b) Metric MDS Configuration with SSTRESS

Figure 3.9: Comparison of Young's NMDS configuration with that derived from Metric MDS trained with the SSTRESS measure.

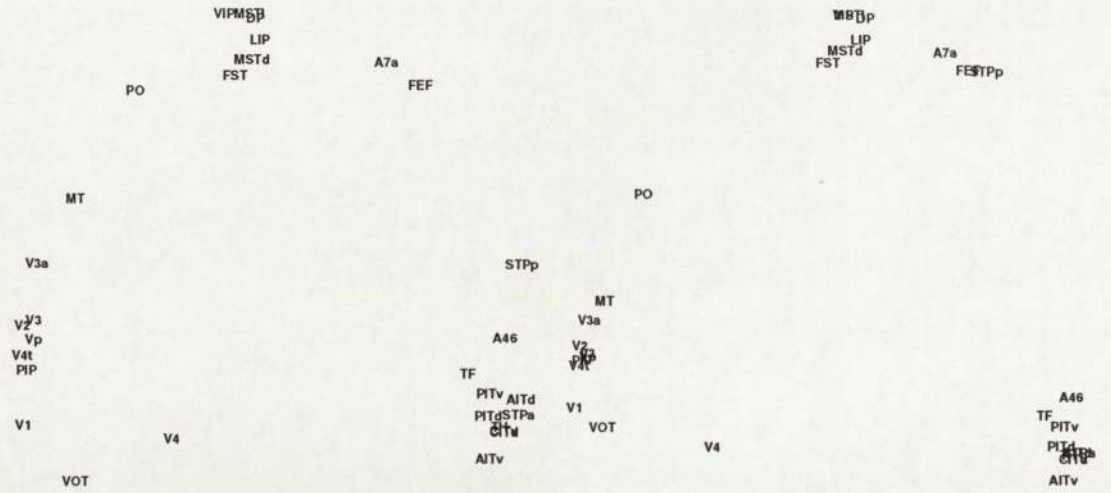


(a) NMDS Configuration with SSTRESS



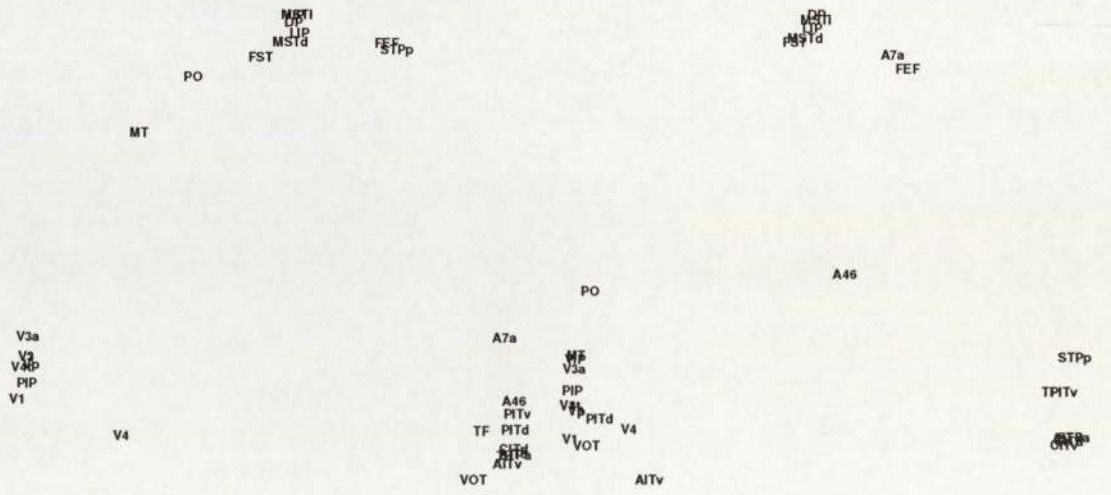
(b) Metric MDS Configuration with STRESS

Figure 3.10: Comparison of Young's NMDS configuration with that derived from Metric MDS trained with the STRESS measure.



(a) Power 3

(b) Power 4



(c) Power 5

(d) Power 6

Figure 3.11: Configurations of the Primate Visual Cortex data derived from Metric MDS trained with Euclidean distance metrics raised to different powers.

3.6 Conclusions

This chapter has presented a detailed study into the problem of artefacts from topographic mappings (in particular Sammon's Mapping and Metric Multidimensional Scaling). It was shown initially that the presence and degree of artefactual structure is determined by the choice of distance metric. The SSTRESS measure was shown to give rise to an annular structure in the map space when trained on uniformly randomly distributed input data. In addition the map points around this ring were observed to become more tightly clustered as the dimension of the input data increased.

However this error measure can be viewed as a particular variant of the standard STRESS measure which employs the squared Euclidean distance metric and it is this metric which is responsible for annular configurations from random input data. When analysed theoretically it was found that the variance of uniformly randomly distributed input data is related to variance of the corresponding map data by a factor of $\frac{p}{q+1}$, where p is the dimension of the input space, q is the dimension of the map space and $p \gg q$.

Mappings utilising the STRESS measure (with the standard Euclidean distance metric) were found to produce a more accurate spatial representation of random input data, although some slight curvature of the map spaces was observed. Higher powers of the euclidean distance metric were shown to result in clusters of map points located on the corners of an equilateral triangle. In addition it was noted that from purely theoretical considerations, the standard Euclidean distance represents the most natural choice of metric for the purposes of data visualisation.

These results were then used to investigate a prominent and controversial use of techniques from Multidimensional Scaling in the analysis of the connectivity of regions in the macaque monkey visual cortex. It was found that the configuration derived by Young through Non-metric MDS with the SSTRESS measure was likely to contain a degree of artefactual structure which would not be present in a configuration derived with the STRESS measure. Thus it was concluded that the primate cortical visual connectivity data was likely to contain a random component which gave rise to the annular configurations observed with the SSTRESS measure, and as a consequence such configurations were not reliable sources of evidence for the "two-streams" hypothesis of visual processing in the primate visual cortex.

Chapter 4

Conclusions

4.1 Overview

This thesis began by considering the properties that are necessary for a technique to be useful as a tool for the visualisation and exploration of high-dimensional data. It was noted that such a technique must be capable of accurately representing structure present in the data in the low-dimensional visualisation space, and in addition any structure which is present in the visualisation space should be representative of true underlying structure in the data space. Much research has gone into addressing the former point yet very little work has been performed concerning this latter (and equally important) point. The work contained within this thesis therefore is an attempt to redress this balance.

4.2 Summary of the Key Results

Since a detailed discussion of the main results of this thesis is presented in the conclusions to Chapter 3, only a brief summary of the key results will be provided here. These are as follows:

- Low-dimensional maps derived from the minimisation of the SSTRESS measure with randomly distributed input data within a high-dimensional hypercube will result in configurations which exhibit an annular or circular artefactual structure in the map space.
- The use of the standard STRESS measure with a Euclidean distance metric provides the most accurate topographic representation of the original data and reduces the possibility of corruption by artefacts.

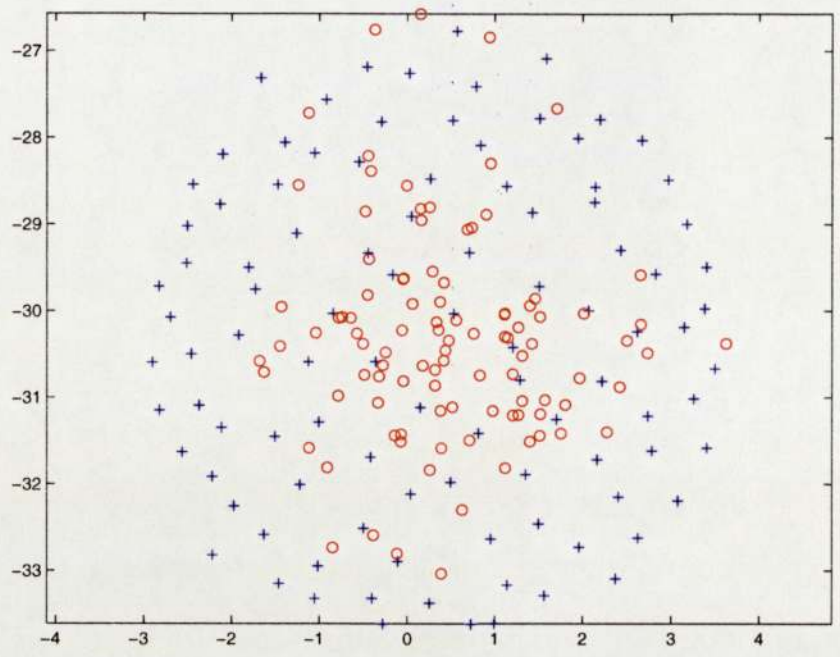
- The configuration of the gross organisation of the primate visual cortical processing system derived with Nonmetric Multidimensional Scaling (trained with the SSTRESS measure) is likely to be corrupted by a degree of artefactual structure.

4.3 Directions for Future Research

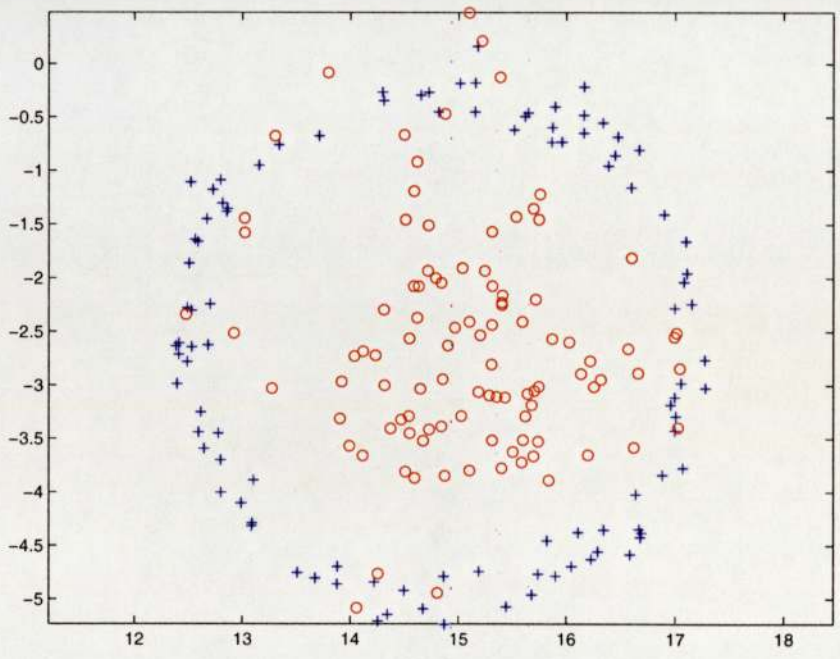
Perhaps the most significant and potentially beneficial area for future research is the extension of the results conatined within this thesis to the technique of Nonmetric Multidimensional Scaling. Since this technique is much more widely used than Metric MDS, an experimental study into the problem of artefactual structure in NMDS configurations would be a worthwhile area for future research.

Another area for future work is the extension of the results to the NEUROSCALE model, as described in section 2.4.3. Since the model is a neural network implementation of Sammon's Mapping, it might be expected that its susceptibility to artefactual structure is similar to that of Sammon's Mapping. However the influence of the relative supervision training algorithm and the generalisation property of NEUROSCALE provide new areas for research.

As a brief introduction to this topic, figure 4.1 overleaf shows the training and test set projections for two NEUROSCALE models trained on 100-dimensional uniform random input data, one with the standard Euclidean distance metric and the other with the squared Euclidean distance metric. As would be expected, the configurations produced with the training data are consistent with the configurations produced by the equivalent Sammon Mappings. However the projections of the test data are novel and worthy of further research.



(a) Euclidean distance metric



(b) Squared Euclidean distance metric

Figure 4.1: Training and test set projections for two NEUROSCALE models (each with 100 hidden units, basis function width of 7.0) trained on 100-dimensional uniformly randomly distributed input data using different distance metrics. The training points are denoted by blue crosses and the test points by red circles.

Appendix A

Derivatives of STRESS for Various Distance Metrics

A.1 Overview

This appendix details the derivative of the STRESS measure E with respect to a particular map point $\mathbf{y}_i$ for various distance metrics. For the general case, we have:

$$E = \sum_i \sum_{j \neq i} \left(d^*_{ij} - d_{ij}\right)^2$$

where: $d^*_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$ and $d_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|$.

The derivative of E w.r.t. $\mathbf{y}_i$ is then given by:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -2 \sum_{j \neq i} \left(d^*_{ij} - d_{ij}\right) \times \frac{\partial d_{ij}}{\partial \mathbf{y}_i} \tag{A.1}$$

Thus it only remains to find the value of the term $\frac{\partial d_{ij}}{\partial \mathbf{y}_i}$ for different distance metrics.

A.2 Euclidean Distance Metric

Define the metric:

$$d_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \| = \left[(y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2\right]^{\frac{1}{2}}$$

Then we have:

$$\frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \frac{1}{2} d_{ij}^{-1} \begin{bmatrix} 2(y_{i1} - y_{j1}) \\ 2(y_{i2} - y_{j2}) \end{bmatrix}$$

$$\Rightarrow \quad \frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \frac{\mathbf{y}_i - \mathbf{y}_j}{d_{ij}}$$

Therefore:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -2 \sum_{j \neq i} \left(\frac{d_{ij}^* - d_{ij}}{d_{ij}} \right) (\mathbf{y}_i - \mathbf{y}_j) \tag{A.2}$$

A.3 Squared Euclidean Distance Metric (or SSTRESS measure)

Define the metric:

$$d_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \| = (y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2$$

Then we have:

$$\frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \begin{bmatrix} 2(y_{i1} - y_{j1}) \\ 2(y_{i2} - y_{j2}) \end{bmatrix}$$

$$\Rightarrow \quad \frac{\partial d_{ij}}{\partial \mathbf{y}_i} = 2(\mathbf{y}_i - \mathbf{y}_j)$$

Therefore:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -4 \sum_{j \neq i} \left(d_{ij}^* - d_{ij} \right) (\mathbf{y}_i - \mathbf{y}_j) \tag{A.3}$$

A.4 Power *n* Euclidean Distance Metric

Define the metric:

$$d_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \| = \left[(y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2 \right]^{\frac{n}{2}}$$

Then we have:

$$\frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \frac{n}{2} d_{ij}^{n-2} \begin{bmatrix} 2(y_{i1} - y_{j1}) \\ 2(y_{i2} - y_{j2}) \end{bmatrix}$$

$$\Rightarrow \quad \frac{\partial d_{ij}}{\partial \mathbf{y}_i} = n \, d_{ij}^{n-2} (\mathbf{y}_i - \mathbf{y}_j)$$

Therefore:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -2n \sum_{j \neq i} d_{ij}^{n-2} \left(d_{ij}^* - d_{ij} \right) (\mathbf{y}_i - \mathbf{y}_j) \tag{A.4}$$

A.5 Minkowski Metric: r = 1 (Manhattan Distance)

Define the metric:

$$d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\| = |y_{i1} - y_{j1}| + |y_{i2} - y_{j2}|$$

Then we have:

$$\frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\begin{array}{llll} \textit{where}: & a = +1, & \text{if } y_{i1} > y_{j1} & \textit{and}: \quad b = +1, \quad \text{if } y_{i2} > y_{j2} \\ & a = -1, & \text{if } y_{i1} < y_{j1} & \qquad\quad b = -1, \quad \text{if } y_{i2} < y_{j2} \end{array}$$

Therefore:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -2 \sum_{j \neq i} \left(d^*_{ij} - d_{ij}\right) \begin{bmatrix} a \\ b \end{bmatrix} \tag{A.5}$$

A.6 Minkowski Metric: r = 3

Define the metric:

$$d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\| = \left[|y_{i1} - y_{j1}|^3 + |y_{i2} - y_{j2}|^3\right]^{\frac{1}{3}}$$

Then we have:

$$\frac{\partial d_{ij}}{\partial \mathbf{y}_i} = \frac{1}{3} d_{ij}^{-2} \begin{bmatrix} 3a \\ 3b \end{bmatrix}$$

$$\begin{array}{llll} \textit{where}: & a = +\left(y_{i1} - y_{j1}\right)^2, & \text{if } y_{i1} > y_{j1} & \textit{and}: \quad b = +\left(y_{i2} - y_{j2}\right)^2, \quad \text{if } y_{i2} > y_{j2} \\ & a = -\left(y_{i1} - y_{j1}\right)^2, & \text{if } y_{i1} < y_{j1} & \qquad\quad b = -\left(y_{i2} - y_{j2}\right)^2, \quad \text{if } y_{i2} < y_{j2} \end{array}$$

Therefore:

$$\frac{\partial E}{\partial \mathbf{y}_i} = -2 \sum_{j \neq i} \left(\frac{d^*_{ij} - d_{ij}}{d^2_{ij}}\right) \begin{bmatrix} a \\ b \end{bmatrix} \tag{A.6}$$

Bibliography

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M., M. Svensén, and C. K. Williams (1998). GTM: The generative topographic mapping. *Neural Computation 10*, 215 – 234.

de Ridder, D. and R. P. W. Duin (1997). Sammon's mapping using neural networks. *Pattern Recognition Letters 18*, 1307 – 1316.

Klock, H. and J. M. Buhmann (1997). Multidimensional scaling by deterministic annealing. In M. Pelillo and E. R. Hancock (Eds.), *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Proc. Int. Workshop EMMCVPR '97, Venice, Italy, pp. 246–260. Springer Lecture Notes in Computer Science.

Kohonen, T. (1990). The self-organizing map. *Proceedings of The IEEE* **78**(9), 1464 – 1480.

Lerner, B., H. Guterman, M. Aladjem, and I. Dinstein (1999). A comparative study of neural network based feature extraction paradigms. *Pattern Recognition Letters 20*, 7–14.

Lowe, D. (1993). Novel 'topographic' nonlinear feature extraction using radial basis function networks for concentration coding in the artificial nose. In *3rd IEE International Conference on Artificial Neural Networks*. London: IEE.

Mao, J. and A. K. Jain (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions On Neural Networks* **6**(2), 296–317.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1997). *Multivariate Analysis*. Academic Press.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions On Computers* **C-18**(5), 401 – 409.

Schiffman, S. S., M. L. Reynolds, and F. W. Young (1981). *Introduction To Multidimensional Scaling: Theory, Methods and Applications*. Academic Press.

Simmen, M. W., G. J. Goodhill, and D. J. Willshaw (1994). Scaling and brain connectivity. *Nature 369*, 448–450.

Svensén, M. (1998). *GTM: The Generative Topographic Mapping*. Ph. D. thesis, Aston University, Birmingham, UK.

Tarassenko, L. (1998). *A Guide To Neural Computing Applications*. Arnold.

Tipping, M. E. (1996). *Topographic Mapping and Feed-Forward Neural Networks*. Ph. D. thesis, Aston University, Birmingham, UK.

Tipping, M. E. and C. M. Bishop (1997). Mixtures of probabilistic principal component analysers. *Neural Computation 369*, 448–450.

Webb, A. R. (1995). Multidimensional scaling by iterative majorisation using radial basis functions. *Pattern Recognition* **28**(5), 753 – 759.

Webb, A. R. and D. Lowe (1990). The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks* **3**(4), 355–364.

Young, F. W. and D. F. Harris (1990). *SPSS Base System User's Guide*, Chapter Multidimensional Sacling: Procedure ALSCAL, pp. 396 – 461. SPSS Inc.

Young, M. P. (1992). Objective analysis of the topological organization of the primate cortical visual system. *Nature 358*, 152–155.