Feed-Forward Neural Networks and Topographic Mappings for Exploratory Data Analysis

David Lowe and Michael Tipping

Neural Computing Research Group
Aston University
Aston Street
Aston Triangle
Birmingham B4 7ET
United Kingdom.
Tel: (+44/0) 121 333 4631
Fax: (+44/0) 121 333 6215

Email: lowed@aston.ac.uk tippinme@aston.ac.uk

June 10, 1996

Abstract

A recent novel approach to the visualisation and analysis of datasets, and one which is particularly applicable to those of a high dimension, is discussed in the context of real applications. A feed-forward neural network is utilised to effect a topographic, structure-preserving, dimension-reducing transformation of the data, with an additional facility to incorporate different degrees of associated subjective information. The properties of this transformation are illustrated on synthetic and real datasets, including the 1992 UK Research Assessment Exercise for funding in higher education.

The method is compared and contrasted to established techniques for feature extraction, and related to topographic mappings, the Sammon projection and the statistical field of multidimensional scaling.

1 Introduction

The visualisation and analysis of high-dimensional data is a difficult problem and one that may be helpfully viewed in the context of *feature extraction*, which provides a useful common ground for exploring the relationships between neural network and more traditional approaches. There is a well established repertory of techniques for the derivation of appropriate feature spaces (usually of reduced dimension), with the nature and utility of these spaces depending significantly on the criteria for their extraction. Amongst such methods, there is a natural dichotomy between those features whose purpose is *representation* of the data, and those whose purpose is its *classification* [4]. There is a clear parallel here with the division into *unsupervised* and *supervised* learning in the neural network domain.

Feature extraction is perhaps the most generic of the pattern processing capabilities of neural networks, and its importance at the centre of developments is two-fold. Firstly, there is a significant advantage to be gained by accompanying *dimensionality reduction*, and secondly it permits the construction of *nonlinear* representations of the data which may then be exploited for information analysis.

These two points are exhibited in the following section, where we present a novel hybrid neural network approach – one which combines both unsupervised and supervised characteristics – to data visualisation and analysis.

2 A Feed-forward Neural Network Topographic Transformation

We seek a dimension-reducing, *topographic* transformation of data for the purposes of visualisation and analysis. By ‘topographic’, we imply that the geometric structure of the data be optimally preserved in the transformation, and the embodiment of this constraint is that the inter-point distances in the feature space should correspond as closely as possible to those distances in the data space. The implementation of this principle by a neural network is very simple. A Radial Basis Function (RBF) neural network is utilised to predict the coordinates of the data point in the transformed feature space. The locations of the feature points are indirectly determined by adjusting the weights of the network. The transformation is determined by optimising the network parameters in order to minimise a suitable error measure that embodies the topographic principle.

Note that this approach is in contrast to the Kohonen network methodology of producing a topographic transformation which exploits an explicit lateral network connectivity and an additional neighbourhood function which is modified heuristically as part of the training process.

The specific details of this alternative approach are as follows. Given a p-dimensional input space of N data points $\mathbf{x}_i$, a q-dimensional feature space of points $\mathbf{y}_i$ is generated such that the relative positions of the feature space points minimise the *stress* term:

$$E = \sum_{i<j}^{N} (d^*_{ij} - d_{ij})^2, \tag{1}$$

where the d^*_{ij} are the inter-point Euclidean distances in the data space:

$$d^*_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)}, \tag{2}$$

and the d_{ij} are the corresponding distances in the feature space:

$$d_{ij} = \sqrt{(\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_j)}. \tag{3}$$

The points $\mathbf{y}$ are generated by the RBF, given the data points as input. That is, $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i; \mathbf{W})$, where $\mathbf{f}$ is the nonlinear transformation effected by the RBF with parameters (weights) $\mathbf{W}$. The distances in the feature space may thus be given by

$$d_{ij} = \| \mathbf{f}(\mathbf{y}_i) - \mathbf{f}(\mathbf{y}_j) \|$$

and so more explicitly by

$$d^2_{ij} = \sum_{l=1}^{q} \left(\sum_k w_{lk} \left[\phi_k(\| \mathbf{x}_i - \boldsymbol{\mu}_k \|) - \phi_k(\| \mathbf{x}_j - \boldsymbol{\mu}_k \|) \right] \right)^2, \tag{4}$$

where $\phi_k()$ are the basis functions, $\boldsymbol{\mu}_k$ are the centres of those functions, which are fixed, and w_{lk} are the weights from the basis functions to the output.

The topographic nature of the transformation is imposed by the stress term which attempts to match the inter-point Euclidean distances in the feature space with those in the input space. This mapping is *relatively supervised* because there is no specific target for each $\mathbf{y}_i$; only a relative measure of target separation between each $\mathbf{y}_i, \mathbf{y}_j$ pair is provided. In this form it does not take account of any additional information (for example, class labels) that might be associated with the data points, but is determined strictly by their spatial distribution. However, as well as a measure of spatial dissimilarity (the inter-point distances), there may also be an additional notion of *subjective dissimilarity* which may be exploited in the transformation.

By 'subjective dissimilarity', we refer to the additional prior knowledge of dissimilarity that may be attributed to each pair of data points. For example, in the extreme case, this may be a simple binary ascription, such that data points representing differing classes have a constant dissimilarity, while those of the same class have zero dissimilarity. This notion, discussed in [10], has been exploited in [1, 30] for generating useful feature spaces that separate classes.

This idea of dissimilarity is only basic, and there may often be more useful prior knowledge available. As illustration, consider the problem of concentration coding in the artificial nose, described in [16]. There, the data is derived from a set of chemical vapour sensors for discrete varying concentrations of ethanol and water vapour. The distribution of data according to the sensor response implies a certain topology and metric, due to the characteristics of the sensors. However our subjective notion of 'concentration coding' typically has an alternative metric. For example it would be natural to consider that the metric in concentration space is linear so that the position corresponding to 30% concentration should be subjectively twice as far away from that of 10% or 50% than that of 20% or 40%. Indeed any such dissimilarity scaling might be chosen consistent with the available prior knowledge. It is then helpful to define for each pair of points both a spatial dissimilarity, d^*_{ij}, and a subjective dissimilarity, s_{ij} – the latter assigned according to the the prior knowledge available concerning points i and j.

This knowledge implies an alternative topology which to some extent should influence the topology induced by the objective spatial data. One convenient way which allows a mixing between the objective and subjective elements and which permits a smooth transition from one to the other is to modify the stress measure from equation (1). The term d^*_{ij} may be replaced with the alternative δ_{ij} defined by:

$$\delta_{ij} = (1 - \alpha).d^*_{ij} + \alpha.s_{ij} \tag{5}$$

The parameter α (where $0 \leq \alpha \leq 1$) can thus be considered as an interpolating parameter between unsupervised and supervised transformations, and is a more general method of including subjective information than that utilised in [1]. With $\alpha = 0$, the the transformation is purely objective, relying solely upon the measured input data distribution. With $\alpha = 1.0$, the transformation is no longer explicitly dependent on the distribution of the data, but is determined by the assigned subjective dissimilarities. Figure 1 depicts the relationships between

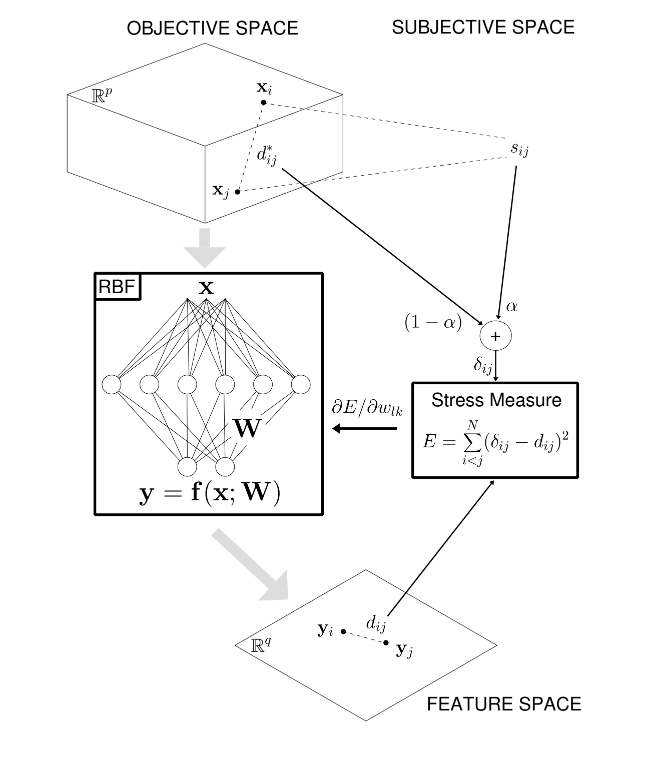

Figure 1: A pictorial representation of the NEUROSCALE model

the various spaces and the rôle of the neural network model.

Combining equations (1), (3) and (4) and differentiating with respect to the weights in the network allows the partial derivatives of the stress $\partial E/\partial w_{lk}$ to be derived for each pattern pair. These may be accumulated over the entire pattern set and the weights adjusted by an iterative procedure to minimise the stress term E. Note that the objective function for the RBF is no longer quadratic, and so a standard analytic matrix-inversion method for fixing the final layer weights cannot be employed. Instead, we used a conjugate-gradient routine [21] for the minimisation, having initialised the weights so as to perform a principal co-ordinates transformation (which is related to the principal components transformation, see section 5.2.3). This has the property of being an optimal *linear* transformation in terms of one specific distance-preserving criterion ([19], p.406).

We refer to this overall procedure as 'NEUROSCALE', with the interpretation that it should be viewed as a tool for visualisation and exploratory analysis of data. As a parameterised Sammon mapping, this approach offers a generalising topographic visualisation of data, but the exploitation of additional subjective information, via the subjective metric, permits the extraction of 'enhanced', more informative, feature spaces. This concept will be illustrated and discussed further in Section 6.2. However, the utility of the procedure is not limited solely to classification problems alone, and is applicable to other domains such as interpolation or time-series analysis. For example, in many data sets of the type described in the concentration coding experiment, there may be no convenient, discrete class encodings (e.g. 30%,40%,50% . . .); instead the concentration variables may be experimentally measured over some continuous range (e.g. 36.9%, 66.2% . . . .). With such data, there is an intuitive measure of dissimilarity between pairs of data points despite the absence of explicit class groupings. In such instances, the NEUROSCALE approach to supervised visualisation is one of the few accessible techniques available.

To illustrate the ideas behind, and effects of, NEUROSCALE, the following section presents a simple and effective demonstration of the approach for class-based problems.

3 A SYNTHETIC EXAMPLE – DATA ON 3 CONCENTRIC SPHERES

To illustrate the principle of the NEUROSCALE method, it was applied to 150 data points in 3-dimensional space, comprising 3 sets of 50 points, each set lying on one of three concentric spheres, with added Gaussian noise. All spheres were centred at the origin with radii 0,1 and 2 units respectively (so that the innermost sphere is effectively a cluster). The data points $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ were generated by the formula

$$\mathbf{x}_i = (r_k + \nu_i).\begin{bmatrix}\cos\theta_i \sin\phi_i \\ \sin\theta_i \cos\phi_i \\ \sin\phi_i\end{bmatrix}, \tag{6}$$

where r_k is the radius (either 0,1 or 2), ν_i is a Gaussian random variable with zero mean and variance 0.05, and θ_i,ϕ_i are uniform random variables in the range $[0, 2\pi)$ and $[0, \pi)$ respectively.

All points on each sphere were considered to belong to a single class and two different schemes for subjective dissimilarities were incorporated. In the first, each sphere is a distinct class with the subjective dissimilarities simply characterised by the absolute difference in radii. So, the matrix of subjective dissimilarities between spheres is naturally given by

$$\mathbf{C}_1 = \begin{bmatrix}0 & 1 & 2\\ 1 & 0 & 1\\ 2 & 1 & 0\end{bmatrix},$$

where the columns are ordered from the innermost sphere to the outermost sphere. In the second case the innermost and outermost spheres are considered to be the same class, so the matrix becomes

$$\mathbf{C}_2 = \begin{bmatrix}0 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 0\end{bmatrix}.$$

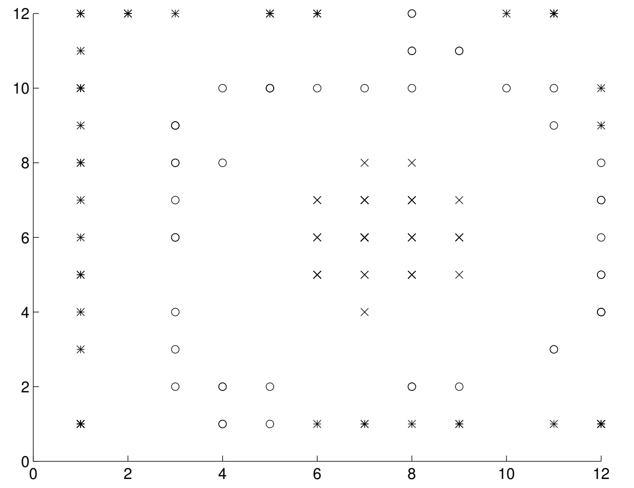

Figure 2: 12x12 Kohonen feature map of the 3-Spheres data.

For the purposes of the NEUROSCALE procedure, we require a value of subjective dissimilarity, s_{ij}, for every pair of data points – giving a subjective matrix $\mathbf{S}$ which is isomorphic to the Euclidean distance matrix $\mathbf{D}^*$. Values of s_{ij} can therefore be determined for every pair of points, given the knowledge of which spheres they lie on, by referring to one of the matrices $\mathbf{C}_1$ or $\mathbf{C}_2$.

The 3-Spheres data is a problem for which a topographic projection based on a Kohonen network is unsuitable. The unsupervised Kohonen feature map of this data is given in figure 2, and illustrates the difficulty of projecting the three distinct surfaces within the data.

The NEUROSCALE transformation was trained for both class models and for values of α of 0, 0.5, 0.75 and 1.0. The resulting projections are given below.

3.1 PROJECTIONS OF THE 3-SPHERES DATA

Two-dimensional projections of the data are illustrated in figures 3 and 4, for each subjective dissimilarity matrix respectively. These results were obtained using a network with 50 Gaussian basis functions.

3.2 DISCUSSION

The plot for $\alpha=0$ in figure 3, displaying the 'opening out' of the spheres, is characteristic of such structure preserving transformations. Although no subjective class information has been exploited, there is a natural separation of the spheres. As α is increased, the spheres are gradually 'folded' until at $\alpha=1$, the RBF has optimally mapped all the data points in each sphere to a single point. A similar phenomenon is evident in figure 4, where the middle sphere is extracted and the other spheres eventually merged. The combination of both topographic and subjective constraints is clear in the $\alpha=0.5$ plot.

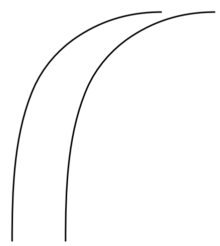

Figure 5: Cross section of the two adjacent planes.

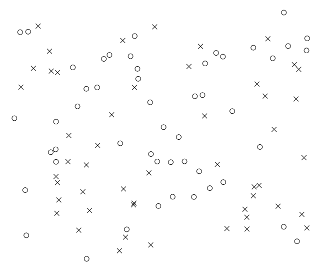

Figure 6: An RBF topographic projection of two adjacent planes with $\alpha = 0$.

4 A SYNTHETIC EXAMPLE – DATA ON ADJACENT PLANES

For this example, 50 data points were distributed uniformly at random over each of two adjacent planes. Both planes were of height 5 units and width 2 units, and were offset by 0.5 units. In addition, each plane curved through an angle of 30°. A cross-sectional illustration of this arrangement is shown in figure 5. Figure 6 shows the unsupervised ($\alpha = 0$) mapping. Naturally both planes are confused. Figure 7, however, gives the projection for $\alpha = 0.5$ where each plane is considered a separate class in a similar manner to the spheres data. This resulting feature space exhibits both a good separation between classes *and* retention of the local topology in each plane, as can be seen by the two overlaid outlines.

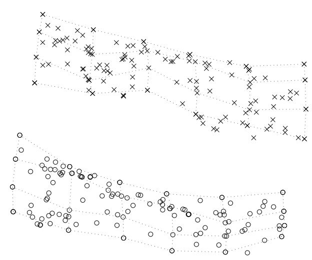

Figure 7: An RBF topographic projection of two adjacent planes with $\alpha = 0.5$. A grid is superimposed.

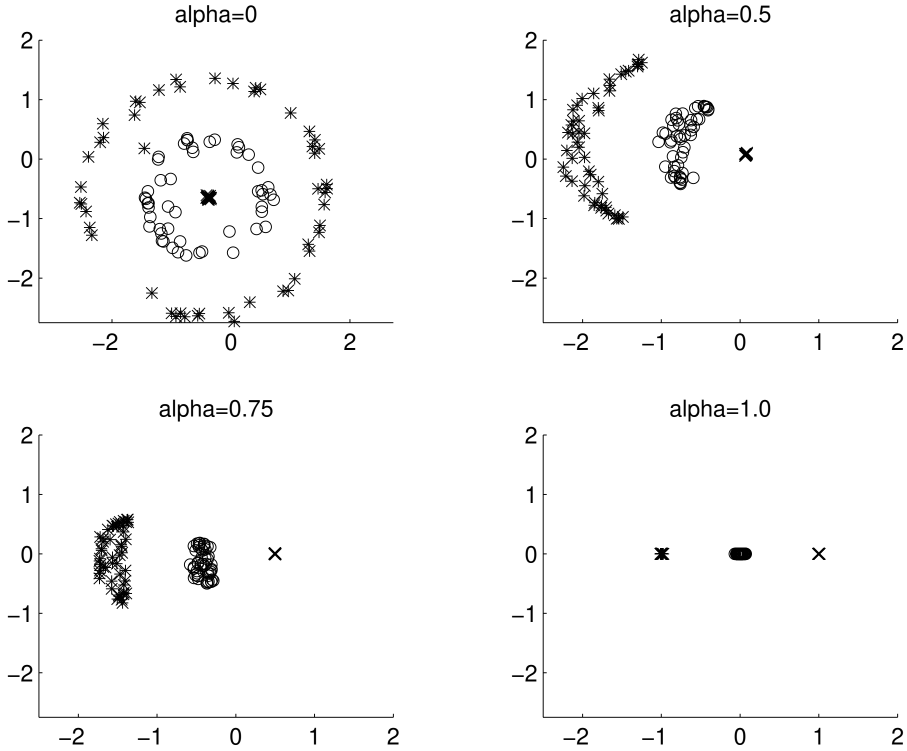

Figure 3: Projections of the 3-Spheres data for subjective matrix $\mathbf{C}_1$.

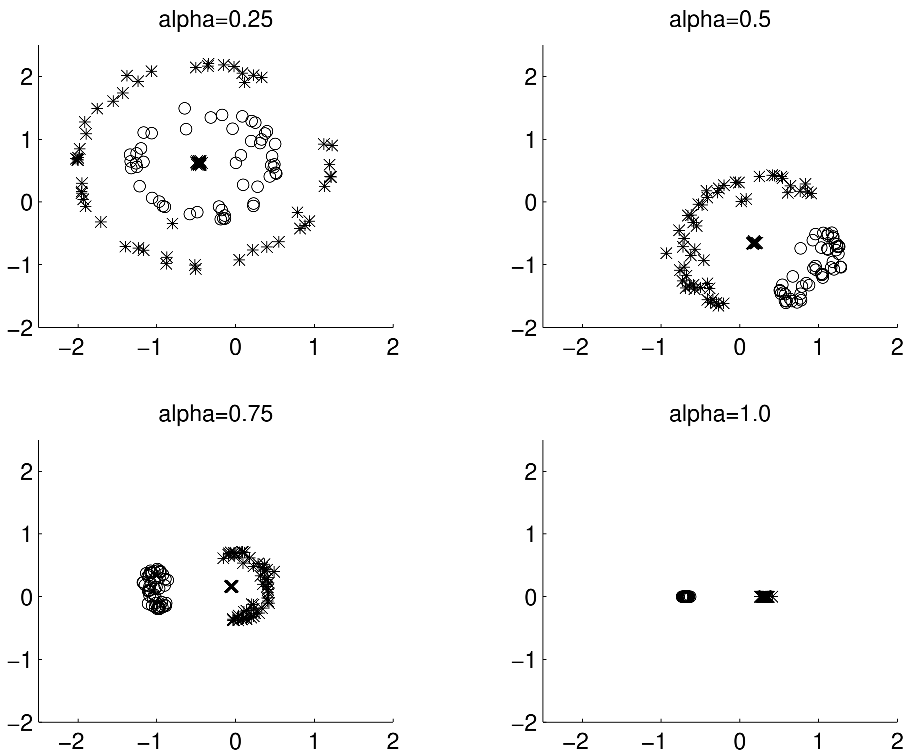

Figure 4: Projections of the 3-Spheres data for subjective matrix $\mathbf{C}_2$.

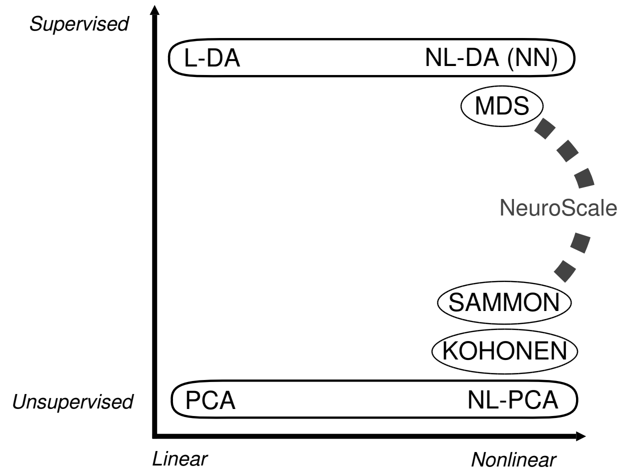

Figure 8: Relational diagram showing the links between several common techniques used for feature extraction.

5 FEATURE EXTRACTION AND TOPOGRAPHIC MAPPINGS

The NEUROSCALE projection is a topographic feature extraction technique that employs additional preferential information. Since the method has links to several other techniques in statistical pattern processing, this section presents a brief overview of some of the more traditional approaches to feature extraction and how they relate to the 'NEUROSCALE' model. The section also discusses the concept of a topographic map, and describes in greater detail two related methods upon which our work builds – the *Sammon mapping* and *multidimensional scaling*.

Figure 8 shows the domain of influence of these methods categorised by their linearity and the degree of exploitation of explicit target information. The interpolation from nonlinear-unsupervised to nonlinear-supervised is also emphasised for the RBF topographic mapping.

5.1 ESTABLISHED FEATURE EXTRACTION TECHNIQUES

The classical unsupervised feature extraction technique is *principal components analysis* (PCA), which is also known as the *Karhunen-Loève expansion*. This is an orthogonal linear projection (usually dimension reducing) which is optimal in the sense of preserving the variance in the transformed data. There has been much exploration of the links between PCA and neural networks (e.g. [20]) and this has been extended to nonlinear forms of PCA both within the neural network context (e.g. [23, 11]) and without (e.g. the method of 'principal curves' [7]).

PCA and related methods make no use of class information that may be available for each given data point. If it is desired to classify the data from the features, then it is advantageous to exploit such information in their extraction. This approach is used in *linear discriminant analysis* (related to *canonical variate analysis*), where a class separation criterion is maximised under linear transformation. One common such criterion is $|\mathbf{S}_B|/|\mathbf{S}_W|$, the ratio of the determinants of the between-class and within-class scatter matrices of the transformed data, and is closely related to Fisher's linear discriminant function [3]. It has been shown that this criterion is maximised in the hidden layer space of a *linear* feed-forward network trained to perform a classification task [5]. Again, this can be generalised to a *nonlinear* neural network, with linear outputs trained to effect classification. The hidden unit space is then shown to be constructed so as to optimise the trace criterion $Tr[\mathbf{S}_B\mathbf{S}_T^+]$, where $\mathbf{S}_T$ is the total scatter matrix of the hidden data [31].

5.2 TOPOGRAPHIC MAPPINGS

The methods outlined in Section 5.1 are concerned with the preservation of variance or class separability under the transformation to the feature space. *Topographic* maps represent another class of unsupervised transformations where the preservation criterion is the *structure* of the data. Use of the term 'structure' implies that the geometric neighbourhood relations between data points are preserved – that is, points that are nearby in the data space will be similarly distributed in the feature space, and equally, points that are more distant should be likewise more distant after the transformation.

5.2.1 *The Kohonen Map*

Although there are several topographic neural network models (for instance, Willshaw's Elastic Net model [32]), the archetypal neural network topographic map is the *Kohonen self-organising feature map* [9], where the structure is imposed on the mapping by means of the neighbourhood function. However, the (usually two-dimensional) topology of the Kohonen net itself imposes a constraint on the resulting projection, and this can limit its utility in many applications (See, for example, [15] or the 3-Spheres data in Section 3).

A non–neural structure-preserving map can be generated by the *Sammon Mapping* [22].

5.2.2 *The Sammon Mapping*

The Sammon mapping is a topographic mapping that seeks to retain structure by maintaining the correspondence between inter-point distances in the data space and the feature space. Given a prior choice of feature space dimension, the Sammon map is generated by minimisation of the *Sammon Stress* error term, similar to that given earlier in equation(1):

$$SS = \frac{1}{\sum_{ij} d^*_{ij}} \sum_{ij} \frac{(d_{ij} - d^*_{ij})^2}{d^*_{ij}}, \tag{7}$$

where d^*_{ij} is the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the data space, and d_{ij} is the distance between their corresponding images $\mathbf{y}_i$ and $\mathbf{y}_j$ in the feature space. The extra terms, with respect to equation (1), serve to reduce the sensitivity of the mapping to the scale of the original data, and also render the stress measure dimensionless. Since $d_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||$, the mapping can be determined by adjusting the points $\mathbf{y}$ iteratively (by a gradient descent method, for example).

The Sammon mapping thus attempts to keep points that are close together in the input space close together in the feature space, and similarly for distant points, and so will approximately preserve any clustering. The extent to which the integrity of this structure from input space can be retained under the mapping is dependent upon both the intrinsic dimensionality of the data, and also on its topology. In the 3-Spheres data from Section 3, neighbourhood relations are distorted by the 'peeling' apart of the spheres and resulting in the artefactual circular structure evident in figure 3. These effects should be borne in mind when interpreting such topographic projections.

In Sammon's original paper [22], several examples are given to illustrate the efficacy of the mapping, where PCA techniques confuse multiple clusters of data, whilst the Sammon map visibly retains their separation in the feature space.

There are, however, some disadvantages to the technique:

- The mapping is generated iteratively and is prone to local minima.
- The computational requirements scale with the square of the number of data points.
- The map is generated as a 'look-up table' – that is, there is no way to project new data since there is no *transformation* defined.
- There is no method to determine the dimensionality of the feature space *a priori*.

In the sense that the Sammon map is concerned with generating a configuration of points in order to fit a matrix of distance measures, it is closely related to the statistical technique of *multidimensional scaling*.

5.2.3 Multidimensional Scaling (MDS)

Multidimensional scaling [2] is a statistical method for generating a configuration of points from a set of *dissimilarity* measurements. A prototypical illustration of this concept given in many MDS texts (and applicable also to the Sammon mapping) is the generation of a map of cities from a table of the measured road distances between them (e.g. [19], p.410). This becomes equivalent to the Sammon mapping if the dissimilarity measurements are an explicit set of inter-point Euclidean distances.

However, in contrast to the Sammon mapping, MDS has been much exploited in psychology and the social sciences where the dissimilarity measurements are more abstract and are usually obtained from experimental human subjective judgements of various stimuli. The assumption is made that these observations can be meaningfully fitted to a set of points in some Euclidean space, where the distance between the points representing each pair of stimuli corresponds to their perceived dissimilarity. It is then hoped that this configuration will aid visualisation of the data and/or provide insight to the processes that generated it. Several good examples of MDS applied to psychological data can be found for various datasets in [2]. More recently, MDS has been applied to analysis of the connectivity of regions within the primate visual cortex [33]. However, in response to this, some controversy has arisen surrounding the interpretation of MDS feature spaces as to whether such maps imply genuine structure in the original data, or alternatively, whether such apparent structure is an artefact of the method [6]. Recall our previous comments that care should be taken in interpreting any method of dimension-reducing topographic mapping. It may be the case that the original data has an intrinsic dimensionality and topology which is in conflict with the constraining topographic projection.

There are two main branches to MDS: the original *metric* method, and the more commonly used *non-metric* method.

In metric MDS, the distances in the configuration are intended to directly correspond to the given dissimilarities. In the original, and largely dominant, metric model, *classical* MDS [28], the configuration is obtained by an analytic method (via a spectral decomposition of the centred inner-product matrix), and, if the dissimilarities correspond to a matrix of Euclidean distances between a set of points in some space, it can be shown to be equivalent to a PCA of those original points. (This method is therefore also known as principle *co-ordinates* analysis.) It can further be shown that classical MDS is an optimal *linear* dimension-reducing transformation of those points with respect to one particular distance-retaining stress measure ([19], p.406).

In non-metric, or *ordinal*, MDS [24, 25, 12, 13] the requirement that distances in the projected space optimally fit the dissimilarities is relaxed so that only the *ordering* of distances is retained. That is, the two most dissimilar stimuli should also be the two most distant points in the configuration and the second most dissimilar pair of stimuli be the second most distant points etc. It is therefore not necessary for all pairs of values to be identical. Indeed, the ordinal constraint implies that it is only necessary that the dissimilarities be some arbitrary monotonically increasing function of the distances. This concept preserves many intuitive psychological properties and in many cases permits the generation of more useful, lower-dimension, lower-stress mappings. In contrast to the classical method, these ordinal configurations must be generated iteratively, usually via a gradient descent procedure, and are computationally expensive due to the requirement of a monotonic regression step.

The Sammon mapping is effectively a *metric* scaling method, but derived by an iterative procedure (and so implicitly nonlinear) rather than by the classical technique (which would simply effect a PCA). As such, its exact analogue does not exist in the MDS domain, although the parallels were pointed out in [14].

5.2.4 Parameterised Maps

One major restrictive disadvantage of Sammon mappings and MDS methods is the look-up table nature of the generated configuration. There is no defined *transformation* that permits the mapping of unseen data – the entire configuration must be re-generated with the new data included. There is therefore no notion of *generalisation*.

With this in mind, several researchers have proposed utilising a parameterised transformation to effect the mapping. It can be seen that instead of directly adjusting the configuration points $\mathbf{y}$ from equation (7), if those points are defined as a (nonlinear) function of the original points $\mathbf{x}$ by $\mathbf{y} = \mathbf{f}(\mathbf{x}; \mathbf{W})$, then the parameters $\mathbf{W}$ may be adjusted in the optimisation procedure to indirectly adjust the points $\mathbf{y}$ and so to minimise the Sammon stress or similar measure. This has several advantages. It enables the transformation of unseen data by $\mathbf{f}(\cdot)$, and also implies that the complexity (in terms of the number of parameters) of the transformation can now depend on the complexity of the data, rather than simply the number of data points. (Note that the number of parameters in the Sammon mapping is the number of data points $\times$ the dimension of the feature space.)

Parameterised Sammon mappings have been outlined in [8, 18], where the transformation is performed by a multilayer perceptron neural network (MLP) and in [16, 17], using a radial basis function (RBF) network. Similar RBF approaches have been exploited in [29, 30] from a MDS perspective. Although there is no specific target information for the training of the networks, given the Sammon stress or similar error measure, expressions for updating weights for each pair of applied patterns may be easily derived. This procedure is intuitively referred to in [16] as *relative supervision*.

As described in Section 2, we adopted a parameterised RBF Sammon mapping approach for our transformation, but with the modified stress criterion given in equation (5) which permits the exploitation of additional subjective knowledge. Hence the NEUROSCALE method may be viewed as a technique which is closely related to Sammon mappings and nonlinear metric MDS, with the added flexibility of producing a generalising transformation which also allows the incorporation of varying degrees of subjective knowledge.

6 Real Dataset Example: The 1992 RAE Database

In 1992, the Universities Funding Council undertook a Research Assessment Exercise (RAE) of 72 separate subject areas in all higher education institutions in the United Kingdom. Institutions supplied numerous quantitative indicators of their respective research activity, such as the number of active researchers, postgraduate students, the values of grants awarded and various numbers of publications. These variables, along with some qualitative input such as example publications, formed part of the input into committees which provided a peer-assessment of the *research rating*, on a scale of 1 to 5, to each subject area at each institution. It should be stressed that the peer review of each submission also incorporated additional non-quantitative information (such as the panel's subjective assessment as to the 'quality' and current activity of research in each unit of assessment). Hence there is not likely to be a simple relationship between the objective database values and the attributed research ratings. However we would expect that there should be some measure of correlation and hence possible structure in this database that may be elucidated by feature extraction techniques. Some statistical analyses have already been published on this database, for example [26, 27].

The data presents a challenge due to both its high dimensionality, which is of the order of 150 in its raw form, and due to the "noise" present in the values of the research rating class labels due to their subjective assignment in the peer review process.

Although there are over 4000 records in the database for all subjects, the distribution of the explanatory variables and the correlation between them and the research rating varies widely between subject areas. For this study, we chose to analyse three related and well correlated subjects – Physics, Chemistry and Biological Sciences. In addition, generalisation performance of the methods was checked by retaining data for the Applied Mathematics subject area.

Initial data preprocessing included: removing redundant and repeated variables; accumulating some indicators which were given for a number of years; all variables were standardised for size by dividing by the number of research staff at the institution and then pre-whitened. The eventual training data set consisted of 217 examples each with 80 explanatory variables, and the following techniques were applied for the analysis of the data:

1. Principal Components Analysis
2. Generalised Linear Discriminant Analysis
3. Sammon Mapping
4. Kohonen Mapping
5. Neural Network (MLP) Classifier
6. RBF topographic transformation ('NEUROSCALE') for values of $\alpha = 0.5$ and 1.0.

In the case of the the final, NEUROSCALE, method, we chose a set of class dissimilarities consistent with a linear relationship between research ratings, and this choice of metric is intended to reflect our preference for the structure of the extracted feature space. For example, we prefer projections of '4'-rated departments to cluster closer to those that are '3'-rated than to those which are '2'-rated, and this linear preference additionally reflects the resultant funding. (A department with a rating of '1' receives no funding, and a '5' receives four times that of a '2', so *funding* $\propto$ (*rating* $- 1$).) This preferential linear structure is illustrated in figure 9.

Figure 9: The preferential structure of the feature space for the five classes of research rating.

This then implies that the dissimilarities between research ratings may be defined by the matrix:

$$\mathbf{C}_{rr} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

and in a similar manner to the 3-Spheres data of Section 3, a value of subjective dissimilarity, s_{ij}, for every pair of departments in the dataset may be determined by reference to the above matrix $\mathbf{C}_{rr}$. Note that the scaling of this matrix is arbitrary, as we are only interested in the *relative* differences. Thus for example, entries $s_{24} = 2$ and $s_{25} = 3$ mean that the relative difference between departments with ratings '2' and '4' is 2, while the relative difference between departments with ratings '2' and '5' is 3. This considered, it is then sensible to scale the values in the matrix $\mathbf{C}_{rr}$ such that the average inter-point subjective dissimilarity is equal to the average inter-point Euclidean distance. This then implies that the NEUROSCALE plot with $\alpha = 0.5$ represents an approximate balance between the twin, objective and subjective, metrics.

6.1 PROJECTIONS OF THE RAE DATA

For the purposes of visualisation and comparison, all projections were into two dimensions, and the feature spaces thus extracted are illustrated in figures 10 and 11.

6.2 DISCUSSION

The projections in figures 10 and 11 illustrate a range of features extracted under assumptions of unsupervised to supervised, and linear to nonlinear. Although a Kohonen network was also applied to this database, the characteristics of the data render the Kohonen representation ineffective and hence the results were not presented. From figure 10a it is clear that the features extracted by the PCA are not likely to be of any exploratory use, since no significant structure could be extracted. However by exploiting explicit class information, the LDA space (10c) manages to separate the patterns approximately into class clusters. The same is true for the hidden unit space of the multilayer perceptron (11d), except that the data separation is more severe (and two of the classes, 4 and 5, are 'confused' by the network as there are only 2 hidden nodes). The use of the LDA to provide a transformation rather than just a mapping is exhibited by 10d which demonstrates the effect of projecting a test set of data ('Applied Mathematics') into the LDA space.

The Sammon mapping (10b) depicts the *nonlinear* but *unsupervised* feature space. Despite not exploiting any class information, this representation produces a set of features which allows patterns in similar classes to remain in close proximity. Neighbourhood relations in the data space are approximately preserved in the feature space. This is a very similar feature space to that the 'NEUROSCALE' network would produce with $\alpha = 0$.

Figure 11a shows a feature space produced by the NEUROSCALE network, with $\alpha = 0.5$, and so incorporating an element of subjective 'class' knowledge into the transformation. The influence of the subjective metric is clearly evident by simple comparison with the Sammon Mapping. There is now a clear ordering of research ratings in a similar topography to that of the LDA projection. In contrast to that linear supervised feature space, careful examination shows that the inter-class boundaries are more pronounced in the NEUROSCALE plot. This may be expected, as the subjective metric element attempts to separate, nonlinearly, points with different research ratings.

The observations in the above paragraph concern the effect on the projection of the subjective metric. In addition to this subjective element, because of the choice of an intermediate value of α (0.5), some of the geometric struc-

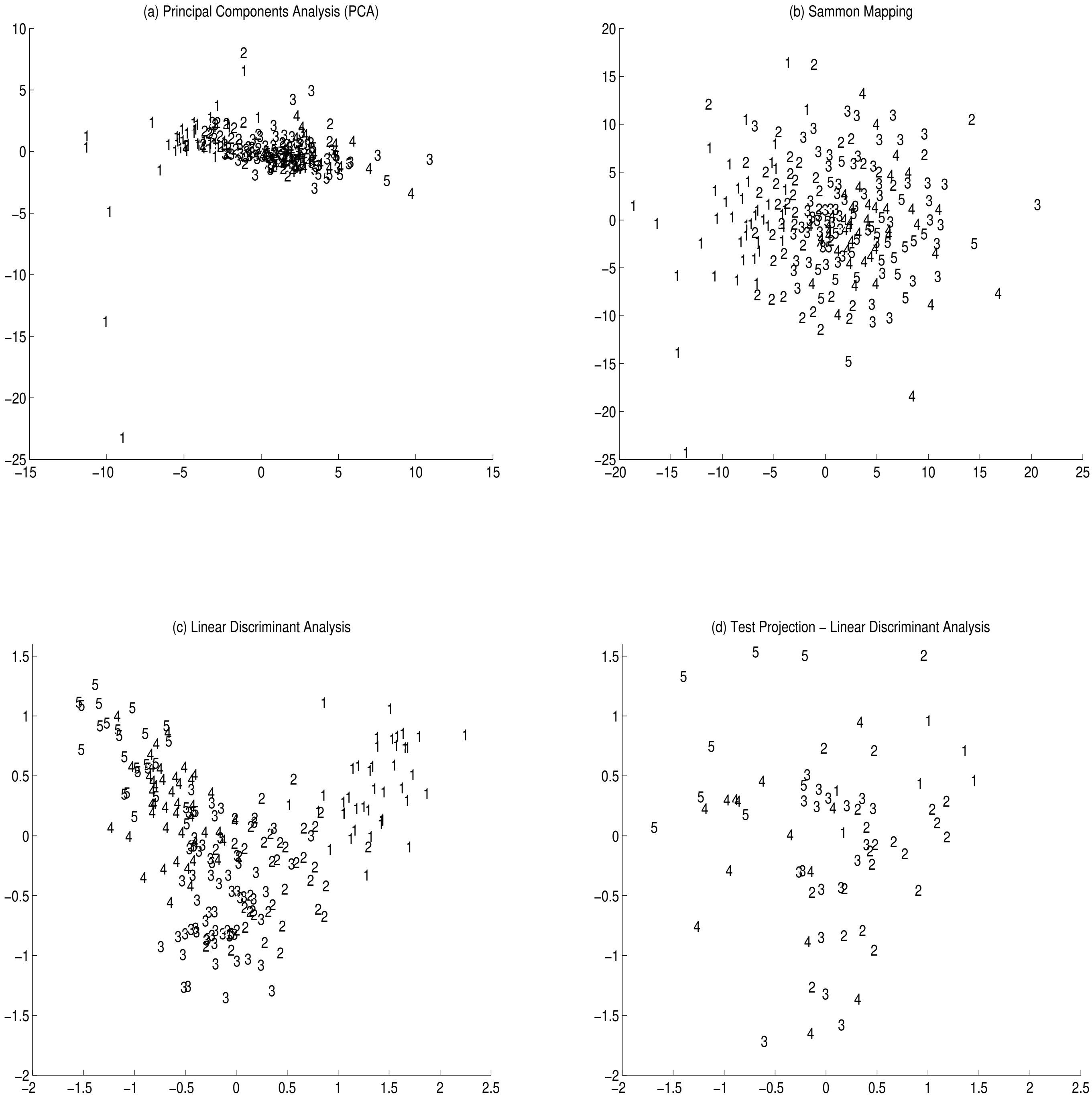

Figure 10: Extracted 2D feature spaces for the RAE data

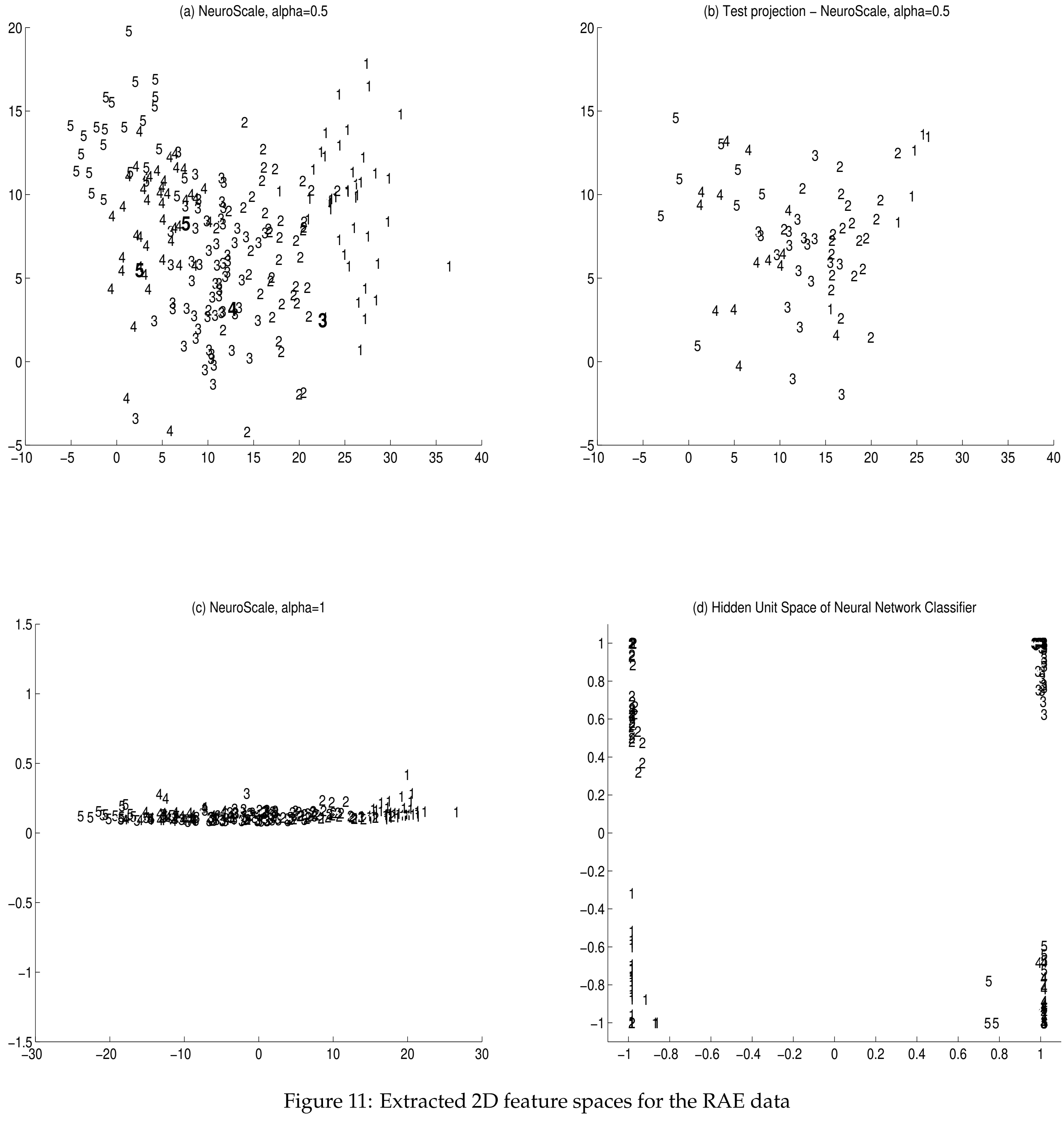

Figure 11: Extracted 2D feature spaces for the RAE data

ture of the original data is retained in the feature space. This implies that useful information might be inferred from the locations of individual points in the feature space, as the structure therein reflects, to some degree, the corresponding topology of the input space. For example, in this plot of the RAE data, it may illuminate potential anomalies in the awarding of research ratings. In figure 11a, four particular departments have been highlighted (in enlarged bold type) on the projection. Each of these departments appears to have received a rating incompatible with its position on the map, judged by consideration of the ratings awarded to its immediate neighbours in the feature space.

These departments are, from left to right on the plot:

- Physics at Heriot Watt University, Edinburgh, which has received a '5' while lying amongst a cluster of '4's.
- Physics at Queen's University, Belfast, which has also received a '5' while lying amongst departments awarded '4's and '3's.
- Physics at Stirling University, which received a '4' while lying amongst a cluster of '3's.
- Physics at the University of Westminster which was awarded a '3', while apparently located on the border between ratings '1' and '2'.

In the case of a purely supervised plot, for example the nonlinear discriminant analysis in figure 11d, the location of individual points with respect to their neighbours in the feature space is largely artefactual. Due to the topographic constraint upon the feature space of figure 11a, there will be an element of structural information therein. In the cases of the individual points highlighted above, further evidence for the structural significance can be elucidated by considering the predictions of a RBF network, trained to classify each department from the input data. The table below shows the actual and predicted ratings of the four departments above.

	Actual	Predicted
Physics, Heriot-Watt	5	3
Physics, Queen's	5	3
Physics, Stirling	4	3
Physics, Westminster	3	1

In general, the classifier predictions support the evidence from the NEUROSCALE plot. In this example, the relative location of points in the feature space has proved informative. With respect the the research exercise itself, although these four particular classifications may appear anomalous, there may well be good explanations due to the operation of the peer assessment process. Firstly, it is noticeable that all four departments are of the Physics unit of assessment. The panel which awarded the ratings for this subject may have had different criteria to those for Chemistry and Biological Sciences. Equally, the panel has access to additional information which, in the case of the four departments in question, may have influenced its judgement.

As a final example feature space generated by the NEUROSCALE technique, the illustration in figure 11c shows a plot for $\alpha = 1$. This feature space is no longer influenced (explicitly) by the spatial distribution of the input data, but is determined by the subjective metric alone. Thus the feature space should represent the preferential knowledge embodied in that metric, and should take the form of a five point clusters distributed along a straight line. The smearing out of the points along that line is a result of the RBF approximation to the Sammon mapping. As the number of basis functions in the transformation is considerably fewer than the number of points (recalling that the data labels are likely to be subject to some considerable noise), there is not sufficient flexibility in the model to precisely locate the points and satisfy the subjective metric constraints.

The main advantage of the NEUROSCALE approach to a Sammon mapping is the ability to generalise. Figure 11b shows the projection of the test set of data ('Applied Mathematics') into the derived NEUROSCALE feature space for the case of $\alpha = 0.5$, exhibiting a 'sensible' projection which could be used for subsequent decision making or inference.

7 CONCLUSION

By comparison, review and example, this paper has attempted to make more accessible a technique for topographic feature extraction which is not widely known. The suggested approach has several advantages over the standard topographic feature map:

- There exist close relationships to several traditional techniques in the statistics and pattern processing literature.
- The technique produces a transformation of the data, rather than just a simple mapping.
- It permits the incorporation of varying degrees of subjective knowledge which can be allowed to influence the extracted feature space.
- Extracted feature spaces are often more 'representative' of the problem than the space extracted by a Kohonen network (e.g. the 3-Spheres problem).
- The number of parameters in the non-linear optimisation process scales only with the size of the network, rather than with the number of patterns. This

is of particular benefit when employing memory-hungry optimisation routines (such as BFGS [21]). Furthermore, we observed a reduction in training time of some 40% (compared to a standard Sammon mapping) for 200 patterns projected to two dimensions using the conjugate gradient optimisation routine. Such improvements are more exaggerated as the number of patterns increases.

Limitations include:

- The computational requirement of the technique still scale with the square of the number of patterns (although the RBF component of the procedure, in terms of the transformation of patterns and calculation of derivatives $\partial \mathbf{y}/\partial w_{lk}$, only scales linearly). This limits the number of training patterns that can be used to produce a transformation. A sequential processing method may alleviate this problem.

- Problems of local minima. Note that in this paper we initialised the RBF weights to perform a principal co-ordinates analysis of the data prior to the nonlinear optimisation. In general, from an arbitrary initialisation the resulting feature space configuration could represent a poor local minimum. In fact, in our experiments and compared to the standard Sammon mapping procedure, random initialisation of the RBF weights was also observed to result in a dramatic reduction of final configurations that represented sub-optimal local minima. The most likely explanation of this phenomenon is that the majority of poor minima in which the Sammon mapping may be trapped are highly *unsmooth* projections that are unrealisable by the RBF network we employed.

- A choice of parameter α is necessary. Appropriate values can only be ascertained on a trial and error basis. The effect of a particular value of α is very much dependent on the order of magnitude of distances in the input space and of the scaling of the subjective dissimilarities applied.

References

[1] Trevor F. Cox and Gillian Ferry. Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153, 1993.

[2] Mark L. Davison. *Multidimensional Scaling*. Wiley, New York, 1983.

[3] R. A. Fisher. The use of multiple measurements on taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, second edition, 1990.

[5] P. Gallinari, S. Thiria, F. Badran, and F. Fogelman-Soulié. On the relations between discriminant analysis and multilayer perceptrons. *Neural Networks*, 4:349–360, 1991.

[6] Geoffrey J. Goodhill, Martin W. Simmen, and David J. Willshaw. An evaluation of the use of multidimensional scaling for understanding brain connectivity. *Philosophical Transactions of the Royal Society of London Series B*, 348:265–280, 1995.

[7] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[8] Anil K. Jain and Jianchang Mao. Artificial neural network for nonlinear projection of multivariate data. In *IJCNN International Joint Conference on Neural Networks*, volume 3, pages 335–340. New York: IEEE, 1992.

[9] Teuvo Kohonen. *Self-organisation and associative memory*. Springer-Verlag, New York, third edition, 1989.

[10] Warren L. G. Koontz and Keinosuke Fukunaga. A nonlinear feature extraction algorithm using distance transformation. *IEEE Transactions On Computers*, C-21(1):56–63, 1972.

[11] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[12] J. B. Kruskal. Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[13] J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[14] J. B. Kruskal. Comments on 'A nonlinear mapping for data structure analysis'. *IEEE Transactions on Computers*, C-20:1614, 1971.

[15] X. Li, J. Gasteiger, and J. Zupan. On the topology distortion in self-organising feature maps. *Biological Cybernetics*, 70:189–198, 1993.

[16] David Lowe. Novel 'topographic' nonlinear feature extraction using radial basis functions for concentration coding in the 'artificial nose'. In *3rd IEE International Conference on Artificial Neural Networks*. London: IEE, 1993.

[17] David Lowe and Michael E Tipping. A novel neural network technique for exploratory data analysis. In *Proceedings of ICANN '95 (Scientific Conference)*, volume 1, pages 339–344. Paris: EC2 & Cie, 1995.

[18] Jianchang Mao and Anil K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317, 1995.

[19] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London, 1979.

[20] Erkki Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[21] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, second edition, 1992.

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

[23] Eric Saund. Dimensionality-reduction using connectionist networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(3):304–315, 1989.

[24] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika*, 27(2):125–140, 1962.

[25] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function, II. *Psychometrika*, 27(3):219–246, 1962.

[26] Jim Taylor. Measuring research performance in business and management studies in the United Kingdom: The 1992 Research Assessment Exercise. *British Journal of Management*, 5:275–288, 1994.

[27] Jim Taylor. A statistical analysis of the 1992 Research Assessment Exercise. *Journal of the Royal Statistical Society A*, 158(2):241–261, 1995.

[28] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[29] Andrew R. Webb. Non-metric multidimensional scaling using feed-forward networks. CSE1 Research Note 192, Defence Research Agency, Malvern, UK, 1992.

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

[31] Andrew R. Webb and David Lowe. The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3(4):367–375, 1990.

[32] D. J. Willshaw and C. von der Malsburg. How patterned neural connections can be set up by self-organisation. In *Proceedings of the Royal Society of London B*, volume 194, pages 431–445, 1976.

[33] Malcolm P. Young. Objective analysis of the topological organization of the primate cortical visual system. *Nature*, 358:152–155, 1992.