Chapter 4

Feature Extraction and the 1992 Research Assessment Exercise

4.1 Introduction

This chapter is a study of the application of NEUROSCALE to a subset of data from the 1992 United Kingdom Research Assessment Exercise (RAE), which has recently been made publicly available [Higher Education Funding Council for England 1994]. This is a 'difficult', high-dimensional, real-world database and offers a good illustration of the potential for visualisation and exploratory analysis provided by the feed-forward neural network topographic approach introduced in the previous chapter.

In the next section some background and details concerning the RAE database are given, followed by a consideration of the NEUROSCALE technique from the perspective of *feature extraction*. This will then be the basis for a comparison of various projections of the dataset obtained both by NEUROSCALE, and by more established methods. To accompany this analysis, there is a description and discussion of classification experiments undertaken on the data presented in Sections 4.4.2 and 4.5.

4.2 The RAE Dataset

4.2.1 Background to the Research Assessment Exercise

One of the major factors determining centralised research funding for university departments[1], through the higher education funding councils in the United Kingdom, is the *research rating*, awarded to each such department as a result of a Research Assessment Exercise (RAE). Such an exercise has been held in 1989 and 1992, with the next one taking place in 1996. In the last exercise, in 1992, there were five integer research ratings, ranging from 1 (lowest quality) to 5 (highest quality). (For the 1996 exercise there will be an expansion to 7 ratings.) Funding varies across subject areas as, generally, the humanities receive less financial support than science or engineering, and is also in proportion to the

[1] The word 'department' is used here as a convenience. The actual individual entities that are assessed are termed 'Units of Assessment', and may, at some institutions, encompass more than one department or alternatively, a single department may have more than one relevant Unit of Assessment.

department size. These factors being equal, research funding is then allocated according to that department's research rating. A department with a research rating of '1' receives no funding at all, a rating of '5' is rewarded with four times the funding of a '2'. So, taking these influences into account, for a given subject area, approximately, $funding \propto (rating - 1) \times size$.

For the purposes of the 1992 RAE, each higher educational institution supplied numerous quantitative measures of their research activity for 72 different subject areas, and each such department was required to supply the variables listed in table 4.1. The submitted data was then intended to be a comprehensive record of the research activity subsequent to the previous exercise in 1989.

Institution No.	No. of student[ship]s from others
Unit of assessment No./No.+letter	No. of grants from ABRC res. councs et al
No. of selected staff	No. of grants from UK based charities
No. of staff not selected	No. of grants from UK central government
No. of postdocs	No. of grants from UK local government
No. of postgrads	No. of grants from UK public corporations
No. of technicians	No. of grants from UK industry & commerce
No. of scientific officers	No. of grants from UK health & HAs
No. of experimental officers	No. of grants from EC
No. of other staff	No. of grants from other overseas
No. of selected staff	No. of grants from PCFC/NAB initiatives
FTE of selected staff	No. of grants from TC schemes - RCs
No. of staff funded from general income	No. of grants from TC schemes - company
No. of staff funded from other income	No. of grants from other
No. of staff in post throughout	Value of grants from ABRC res. councs et al
No. of category A staff	Value of grants from UK based charities
No. of category B staff	Value of grants from UK central government
No. of category C staff	Value of grants from UK local government
No. of category D staff	Value of grants from UK public corporations
No. of publications	Value of grants from UK industry & commerce
No. of research assistants	Value of grants from UK health & HAs
No. of postgrad research students	Value of grants from EC
No. of cited authored books	Value of grants from other overseas
No. of cited edited books	Value of grants from PCFC/NAB initiatives
No. of cited short works	Value of grants from TC schemes - RCs
No. of cited refereed conferences	Value of grants from TC schemes - company
No. of cited other conferences	Value of grants from other
No. of cited editorships	Usage of reseach council facilities (pounds)
No. of cited articles for academic journals	No. of published authored books
No. of cited articles for professional journals	No. of published edited books
No. of cited articles for popular journals	No. of published short works
No. of cited reviews of academic books	No. of published refereed conferences
No. of cited other publications	No. of published other conferences
No. of cited misc publications	No. of published editorships
No. of cited pubs classified as applied research	No. of published art.s for academic journals
No. in FT research	No. of published art.s for profess. journals
No. in PT research	No. of published art.s for popular journals
No. of Doctorates	No. of published reviews of academic books
No. of Masters	No. of published other publications
No. of FTE postgrad students	No. of published misc publications
No. of student[ship]s from ABRC res. counc.s etc	No. of staff producing publications
No. of student[ship]s from UK based charities	No. of selected staff on payroll
No. of student[ship]s from UK central government	No. of staff on payroll
No. of student[ship]s from UK local authorities	Proportion of staff submitted (letter)
No. of student[ship]s from UK public corporations	FTE staff submitted
No. of student[ship]s from UK industry & commerce	Research rating for applied research
No. of student[ship]s from UK health & HAs	Research rating
No. of student[ship]s from other overseas	

[Note: FT=full-time; PT=part-time; FTE=full-time-equivalent; TC=Teaching Company]

Table 4.1: The variables supplied by each department for the 1992 RAE. Note that some of the fields had to be supplied for individual years.

There are in the order of 95 distinct data types, and with some types (such as the numbers of publications) being supplied for individual years, there is a total of nearly 150 fields in the database. The research rating for each specific subject area at each institution is also supplied with the data.

The award of the research rating is partly dependent on this quantitative data, but critically, the final decision depends on the judgement of an assessment panel, composed of a number of experts in the relevant fields. In addition to the numerical data, each institution supplied two specimen publications per academic for consideration. This peer assessment component may be regarded as adding considerable noise to the data, with respect to the prediction of research ratings from the quantitative indicators alone.

Some statistical studies of this data have been previously published, both for the 1989 exercise [Johnes, Taylor, and Francis 1993] and for that of 1992 [Taylor 1994; Taylor 1995]. The data available from the 1989 exercise is less comprehensive and reliable than that from 1992. Johnes, Taylor, and Francis [1993] nevertheless observe that size has a positive influence on research quality, and also note a 'halo' effect, where a particular department at an institution is awarded a higher research rating than might otherwise be justified by the quantitative data alone and where other departments at that same institution have obtained justifiably high ratings. The analysis of Taylor [1994] was based upon the 1992 data and was restricted to Business and Management schools. This was later extended to all subject areas [Taylor 1995]. Again, size was found to be a statistically significant explainer of rating, along with three other variables: the number of articles in refereed academic journals per capita, the number of full-time equivalent postgraduates per capita and the amount of research council grants per capita.

It should be emphasised that the research rating is intended as a measure of the *quality* of research, and not its *quantity*. For a given research rating, departmental funding is proportional to the number of active research staff, so *funding* will be related to the size of the department even though the award of the *research rating* should not be. There is therefore no inherent reason for a large department to score more highly than a smaller one. Nevertheless, the observations of the previous paragraph still apply, that "research quality improves with size" [Taylor 1994] and suggested explanations for this effect are offered by Taylor [1995]. Further evidence for this assertion was also derived through a factor analytic study by Barrett [1995].

From an academic perspective, the RAE has aroused great interest, debate and not inconsiderable controversy, with much comment passed that it is wasteful of resources, unfair, will ultimately erode quality through its quantitative emphasis, and encourages short-termism. Indeed, there has been considerable discussion in the national press, for example [Crace 1995; MacLeod 1995], on its implications. This chapter is not able to address these wider issues, but can question the value of the quantitative data and whether it is truly informative given the various panels' ultimate qualitative assessments through the ratings awarded.

4.2.2 Extracting Experimental Data from the RAE Database

It is of natural interest to investigate the relationships and structure within the RAE database. The prospect of being able to predict the value of research rating from the quantitative data alone, and thus infer which indicators are salient, has been the prime motivation for its statistical study. Furthermore, it is in the interest of the assessment bodies to confirm that the data that is being collected (at the cost of considerable time and expense) is both valid and a useful contribution to the exercise. Indeed, the funding councils themselves have commissioned research on the data.

The challenge of elucidating structure from within the raw RAE dataset is actually quite demanding. If all the departments are included together, there are over 4000 data points in close to 150 dimensions. However, treating these points as one large monolithic dataset is not a sensible approach to data analysis. Firstly, the distribution of the data will vary significantly across subject areas, which is evident from simple examination of the raw data values. For example, consider the likely contrast between "number of published authored books" and "value of grants from health authorities" for subjects such as Celtic Studies and Biological Sciences. Secondly, the importance of these individual variables in determining ratings will also exhibit considerable differences from department to department, as a separate assessment panel sits in each case. Statistical evidence for these effects has been offered by Taylor [1995].

It is thus appropriate to study the database on a department-by-department basis. Unfortunately, for each subject are there are limited exemplars (typically 50-90 in most cases), which represents fewer data points than dimensions. This is a major disadvantage, particularly if it is desired to build a classifier of the data. A sensible approach, therefore, would be to seek a number of subject areas which exhibited similar structure in the data, and where the relationships between the variables and the awarded research ratings were fairly constant. Correlation analysis by the Joint Performance Indicators Working Group [1994] indicates that one such appropriate grouping would be to combine the data from Physics, Chemistry and Biological Sciences, which gives a dataset of 217 examples that has been chosen for analysis within this chapter. In addition, data from Applied Maths is also well correlated (although to a lesser degree), and these 67 examples will also be utilised as a test sample for the purposes of assessing generalisation of various models.

Although it was desired that no prejudice be exerted over the choice of variables to be retained for any subsequent projection, it was possible to remove several redundant and repeated indicators. It was also determined to agglomerate those variables given for individual years over the period they encompassed. This approach reduced the dimensionality of the input data to 80, which is nevertheless high, particularly in comparison with only 217 available examples.

The data was standardised for size by dividing throughout by the number of staff in the department (specifically CATEGORY_A + CATEGORY_B staff). This is intuitively sensible and has been the approach adopted in other studies, including that of the Joint Performance Indicators Working Group [1994]. The variables thus become of the form 'number of published short works per researcher', for example. This is an imposition of the prior knowledge that research ratings are quality-based. Although empirical evidence exists that research rating is related to size, this is likely to be an indirect effect, as hinted at by Taylor [1995]. For example, larger departments may be more productive due to economies of scale, or may have expanded as a result of research success attracting new funding and boosting staff levels.

Finally, because of the inhomogeneity of the input variables, and the wide ranges thereof, each such variable was normalised to zero mean and unit variance.

4.3 Feature Extraction, Neural Networks and NEUROSCALE

In practical application neural networks can fulfil a variety of rôles. Common examples are for classification, nonlinear regression, function approximation, auto-encoding, topographic mapping (as in the case of NEUROSCALE) or time series prediction. In all of these diverse incarnations, the neural network can be viewed as performing some variety of *feature extraction* process. This interpretation provides a helpful unifying perspective when comparing neural network approaches with more classical techniques.

'Feature extraction' is a term for the process of deriving alternative (and usually lower-dimensional) representations of data that are more appropriate for a given specific application. While it might appear intuitive that access to more input features should improve performance, the inclusion of noisy or redundant features in a model can easily lead to a degradation thereof [Weiss and Kulikowski 1991]. For example, in the case of limited data, an additional feature that is irrelevant (in terms of classification, say) will simply exacerbate over-fitting as it can be interpreted as acting as a key to a look-up table.

One example of feature extraction is in a classification scenario where data may be projected linearly onto those axes that best separate the classes according to some criterion. This is a case of a *linear*, *supervised* extraction process, and it is convenient to group feature extraction techniques in general according to their linear–nonlinear and supervised–unsupervised nature. This categorisation has been adopted in three recent studies [Mao and Jain 1995; Lowe and Tipping 1995; Lowe and Tipping 1996].

The classical linear, unsupervised feature extraction technique is *principal components analysis* (PCA) [Jolliffe 1986]. The first q principal components of a set of data points are those obtained by the projection of the p-dimensional data onto the q orthogonal axes that retain the maximum variance. The projection vectors may be easily found as the dominant eigenvectors of the data (sample) covariance matrix, or alternatively, there are numerous neural network architectures specifically designed to extract principal components and subspaces (see Chapter 5 for some examples). There are also nonlinear extensions of PCA, both within the neural network domain [Saund 1989; Kramer 1991] and from a statistical viewpoint [Hastie and Stuetzle 1989; LeBlanc and Tibshirani 1994], as well as a hybrid of the two [Dong and McAvoy 1996]. A more complex class of unsupervised methods are those incorporated in *projection pursuit* [Friedman and Tukey 1974; Huber 1985]. These schemes look to maximise alternative criteria under projection, for example, higher-order statistics such as skew and kurtosis.

All the techniques mentioned in the above paragraph utilise the data in isolation, without any explicit class information, and are thus considered to be *unsupervised*. If alternative information is available, most commonly in the form of class labels, then this may be exploited in the extraction process to produce *supervised* feature spaces more appropriate for subsequent classification. This concept was originally considered by Fisher [1936] in the search for a linear projection of the measured variables that maximised the separation of two classes, or groups, from within the Iris dataset. His empirical approach was to maximise the quotient of the squared group-mean difference and within-group variance in the projection. This may be formulated as $(\mathbf{a}^{\mathrm{T}}\mathbf{S}_b\mathbf{a})/(\mathbf{a}^{\mathrm{T}}\mathbf{S}_w\mathbf{a})$, where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the *weighted between-group covariance matrix* and *within-group pooled covariance matrix* respectively, with $\mathbf{a}$ the projection vector. This ratio may be maximised by setting $\mathbf{a}$ to the dominant eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$. The resultant projection is known *Fisher's linear discriminant function*, and although it was designed simply to be "sensible" and makes no explicit assumptions of the form of the class conditional densities, it may be shown to be equivalent to a *maximum likelihood* discriminant rule for two classes characterised by normal distributions with identical covariance matrices [Mardia, Kent, and Bibby 1979].

Fisher's approach was generalised to the discrimination of more than two groups by Rao [1948] and Bryan [1951]. In the case of generalised *linear discriminant analysis* (LDA), an appropriate criterion for discrimination is $\mathrm{tr}\left[\hat{\mathbf{S}}_w^{-1}\hat{\mathbf{S}}_b\right]$, where $\hat{\mathbf{S}}_w$ is now the within-group pooled covariance in the *transformed* space, and likewise, $\hat{\mathbf{S}}_b$ is the transformed counterpart of $\mathbf{S}_b$. Such a criterion naturally increases as the within-group variances become smaller and the between group variances become larger. For this measure, the projection vectors $\mathbf{a}_i$ may be found by solution of the generalised eigenvector equation $\hat{\mathbf{S}}_b\mathbf{a}_i = \lambda\hat{\mathbf{S}}_w\mathbf{a}_i$, and the resultant discriminant axes are often referred to as the *canonical variates*. For a k-class problem, all the discriminatory information, as measured by $\mathrm{tr}\left[\hat{\mathbf{S}}_w^{-1}\hat{\mathbf{S}}_b\right]$, may be retained by projection onto $\min(p, k-1)$ eigenvectors.

In the neural network domain, it has been demonstrated by Gallinari et al. [1991] that the hidden layer space of a *linear* neural network trained as a classifier maximises the criterion $|\hat{\mathbf{S}}_b|/|\hat{\mathbf{S}}_t|$, where $\hat{\mathbf{S}}_t = \hat{\mathbf{S}}_w + \hat{\mathbf{S}}_b$ is the *total covariance matrix* measured in the feature space, which generates an equivalent projection to $\mathrm{tr}\left[\hat{\mathbf{S}}_w^{-1}\hat{\mathbf{S}}_b\right]$ [Fukunaga 1990]. Again, there are nonlinear extensions of this approach, with the hidden unit space of a linear output-layer multi-layer perceptron having been shown to maximise the *network discriminant function* $\mathrm{tr}\left[\hat{\mathbf{S}}_t^{+}\hat{\mathbf{S}}_b\right]$ for a particular target coding scheme [Webb and Lowe 1990]. This again gives an equivalent projection to the $\mathrm{tr}\left[\hat{\mathbf{S}}_w^{-1}\hat{\mathbf{S}}_b\right]$ criterion [Fukunaga 1990].

Topographic mappings, such as those generated by the Kohonen or Sammon Maps, may be viewed as nonlinear, unsupervised feature extraction processes. Here the criterion for selection of features is not to maximise variance or class separability, but rather that the topology, or geometric structure, of the data be preserved in the feature space. Naturally, NEUROSCALE with $\alpha = 0$ also fits into this category. Alternatively, as discussed in Chapter 3, NEUROSCALE with $\alpha = 1$ may be considered a supervised technique, as full subjective (or class) information is exploited in the mapping. However, only with $\alpha = 1$, and in certain special cases, may it be interpreted as a form of discriminant analysis. For example, consider a mapping with subjective dissimilarities set to zero for data points in the same class ($\delta_{ij} = 0 : \omega_i = \omega_j$) and set to some constant value for data points in different classes ($\delta_{ij} = \lambda :$

$\omega_i \neq \omega_j$). In this case, the STRESS measure is

$$\sum_{ij}(\delta_{ij} - d_{ij})^2 = \sum_{\omega_i=\omega_j} d_{ij}^2 + \sum_{\omega_i\neq\omega_j} (\lambda - d_{ij})^2, \tag{4.1}$$

$$= \sum_{\omega_i=\omega_j} d_{ij}^2 + \sum_{\omega_i\neq\omega_j} (\lambda^2 - 2\lambda d_{ij} + d_{ij}^2), \tag{4.2}$$

$$= \sum_{ij} d_{ij}^2 - 2\lambda \sum_{\omega_i\neq\omega_j} (d_{ij} - k), \tag{4.3}$$

$$= 2N\mathrm{tr}\,[\mathbf{S}_t] - 2\lambda \left(\sum_{\omega_i\neq\omega_j} [(\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_j)]^{1/2} - k \right), \tag{4.4}$$

where $k = \frac{1}{2}\sum_{\omega_i\neq\omega_j} \lambda$ and is constant. Equation (4.4) is very similar to one particular discriminant criterion, $\mathrm{tr}\,[\mathbf{S}_t] - \mu(\mathrm{tr}\,[\mathbf{S}_b] - c)$, where μ and c are constants [Fukunaga 1990, pp447]. The presence of the square root in equation (4.4) unfortunately prevents any further simplification with respect to this standard measure, but it may be heuristically interpreted as minimising the total covariance while retaining some measure of the between-group spread at a constant value.

However, in general even for $\alpha = 1$, there is not such a close relationship to discriminant analysis, as the subjective dissimilarities will not be as simplistic as in the above example.

It can therefore be intuitive to view the parameter α as interpolating in some manner between an unsupervised mapping and a supervised variant — a variant which is related closely to the philosophy of multidimensional scaling because the topography of the feature space is then determined by the subjective information, and not explicitly by the spatial distribution of the data.

The collection of techniques discussed above may be related by the taxonomy in figure 4.1, which categorises some of the aforementioned feature extraction schemes and places the NEUROSCALE approach in that context also. The techniques illustrated in that diagram will be applied to the analysis of the RAE dataset in the following section.

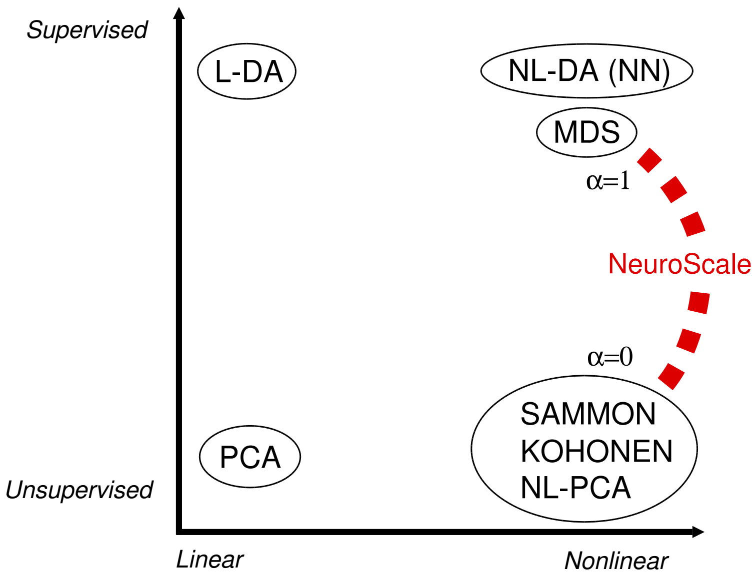

Figure 4.1: A schematic of feature extraction approaches.

4.4 Experiments on the RAE Data

For the purposes of brevity in this section, the combined Physics, Chemistry and Biological Science set of 217 data points will be referred to as the RAE_PCB dataset, with the 67 example data points from Applied Maths referred to as RAE_AM.

4.4.1 Principal Components Analysis

As the data in RAE_PCB is 80-dimensional, a sensible first step is to perform a principal components analysis (PCA). If the data is actually residing for the most part in a lower dimensional linear subspace, then this will be exposed by PCA, and the data may be projected down onto a reduced number of principal axes, without significant loss of information. These axes are the principal eigenvectors of the sample covariance matrix of the data, with the variance projected onto each axis given by the corresponding eigenvalue. Axes along which the projected variance is deemed negligible, where the eigenvalues are small compared with the principal eigenvalues, may be judiciously discarded.

The eigenvalues of the covariance matrix of the RAE_PCB dataset are given in figure 4.2.

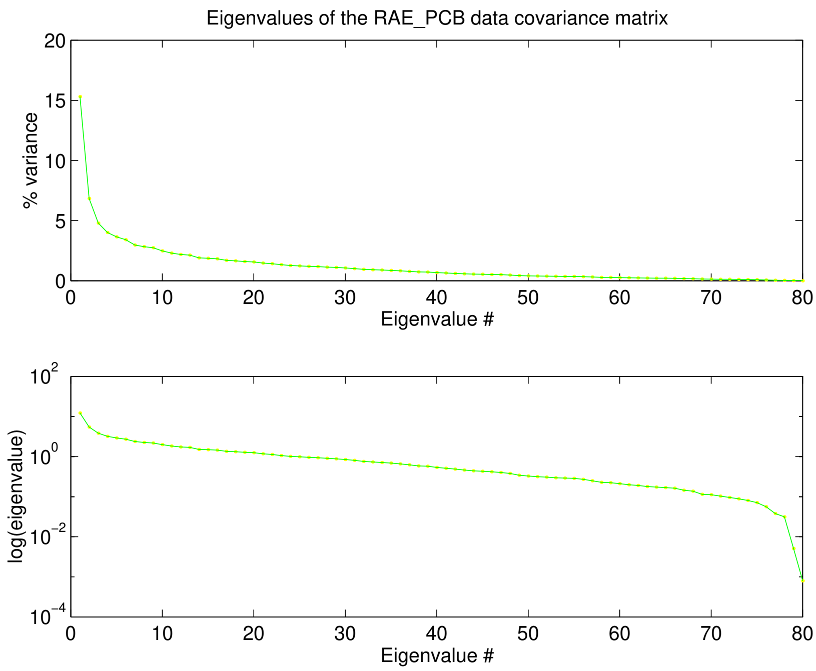

Figure 4.2: The eigenvalues of the RAE_PCB covariance matrix.

After the first two principal eigenvalues, there is a gradual decay in magnitude, with even the 32-nd dimension contributing greater than 1% of the overall variance. The large principal eigenvalue is characteristic of covariance matrices where the majority of variables are positively correlated, and can be considered a measure of the *size* of those variables [Chatfield and Collins 1980]. This is the case for the RAE_PCB dataset, as can be seen by the illustration of the sample covariance matrix in figure 4.3. This effect aside, in both the linear and logarithmic eigenvalue plots, the data does not exhibit the characteristic fall-off associated with inherently lower-dimensional data and there is no suggested cut-off point for discarding dimensions.

The feature space defined by a projection onto the two principal axes is given in figure 4.4. Very little

Figure 4.3: A schematic of the covariance matrix of the RAE_PCB dataset. White squares are positive elements, black squares negative.

information concerning the data can be elucidated from this plot, apart from the existence of some outliers with rating '1' at one extreme of the distribution. The linear nature of the projection implies that most of the structure in the data will be lost, particularly as over 78% of the variance is present in the remaining components. This is the major restriction of PCA, particularly when applied to such high-dimensional data, and in general it is best employed to seek degeneracies in the data as a preliminary to other techniques.

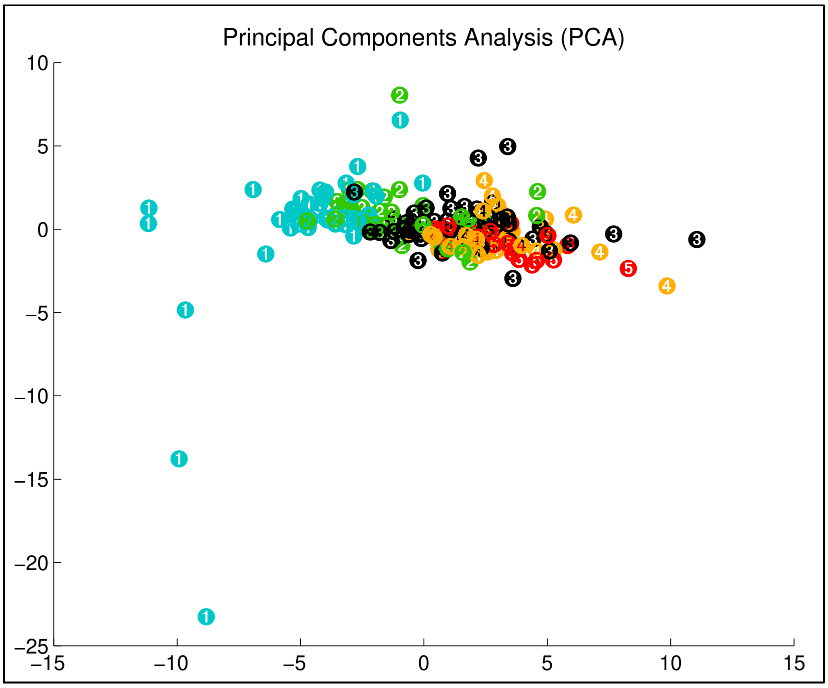

Figure 4.4: Projection onto first two principal axes of RAE_PCB.

4.4.2 Classification of the Data

Before considering alternative spatial representations of the data, it is of interest to consider a classification of RAE_PCB with respect to the research rating. Is it possible to make reliable predictions of the research rating awarded by the panel from the quantitative data alone? Even more importantly, is there any relationship at all between research rating and the gathered data?

Three different prediction models were considered, all fully connected to the 80 inputs and trained to produce a 1-of-5 output coding:

① Linear Model,

② Multilayer Perceptron Neural Network (MLP),

③ Radial Basis Function Neural Network (RBF).

The results for each model are given below, in terms of prediction error and misclassification matrices for both training and test sets. Each row of a misclassification matrix represents the actual class, and each column the prediction. So in the first such matrix below, the '1' in row 2 column 4 indicates a university that actually received a research rating of 2 but was predicted to gain a 4. Note that, for accurate estimating of prediction error of a model, the test set should be sampled from the same distribution as the training set [Lowe and Webb 1990; Lippmann 1994]. (That is, the prior probabilities for the respective classes should be identical for each set.) This assumption will be made for the RAE_PCB and RAE_AM datasets here, although study shows that, compared to RAE_PCB, RAE_AM has an exaggerated number of rating '1's and a reduced proportion of '2's. The class sizes are, in order, for RAE_PCB: (42,42,62,44,27) and for RAE_AM: (6,21,18,14,8). However, to alter the model based on this knowledge (through a weighting of the error function) would be "cheating" in this instance.

① Linear Model

The classification results were:

□ Training Set : LINEAR MODEL □

```
MSE per output:          0.0360696
Patterns Misclassified: 33 (out of 217)
Percentage Correct:     84.8%

Classification Matrix
=====================
 40  2  0  0  0
  4 35  2  1  0
  3  5 51  2  1
  0  2  3 36  3
  0  0  1  4 22
```

■ **Test Set : LINEAR MODEL** ■

```
MSE per output:          0.106999
Patterns Misclassified: 41 (out of 67)
Percentage Correct:      38.8%

Classification Matrix
=====================
  5  1  0  0  0
  6  6  7  1  1
  1  6  7  4  0
  1  2  5  3  3
  0  0  1  2  5
```

② *Multilayer Perceptron*

The architecture of the MLP was 80-h-5, where h is the number of hidden units in the network. The hidden-unit activation functions were hyperbolic tangents (tanh) and the output neurons were linear. To determine a near-optimal number of hidden units, training and test errors were evaluated for 1-12 units, with the errors averaged over 25 runs with different random initial weight configurations. Weights were optimised with a conjugate-gradient routine. Plots of sum-of-squared training errors and test misclassification rate against number of hidden units is given in figure 4.5 below. Standard deviation error-bars are also shown.

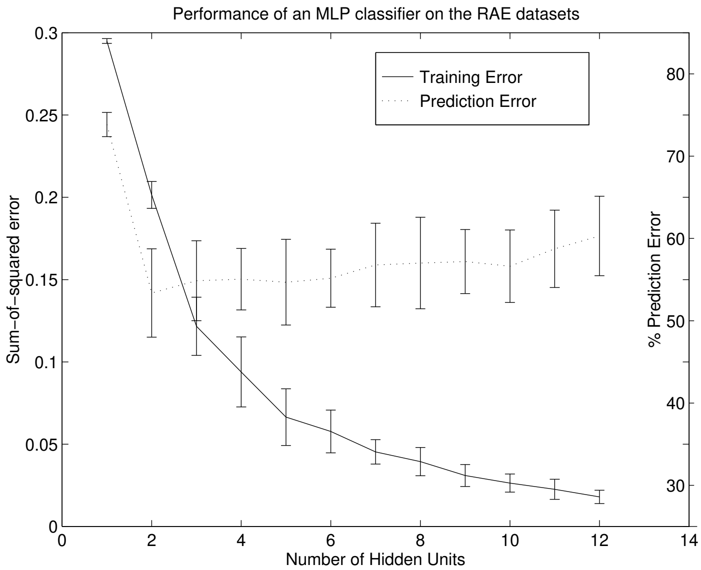

Figure 4.5: An MLP classifier trained on the RAE_PCB data and tested on the RAE_AM set.

In terms of misclassification rate, a network with 2 hidden units is optimal in this experiment:

☐ Training Set : MLP (2) ☐

```
MSE per output:         0.0402512
Patterns Misclassified: 37 (out of 217)
Percentage Correct:     82.9%

Classification Matrix
=====================
 42  0  0  0  0
  3 39  0  0  0
  0  1 61  0  0
  0  0  6 38  0
  0  0  0 27  0
```

■ Test Set : MLP (2) ■

```
MSE per output:         0.0802902
Patterns Misclassified: 35 (out of 67)
Percentage Correct:     47.8%

Classification Matrix
=====================
  5  0  0  1  0
  7 12  0  2  0
  5  4  8  1  0
  0  1  6  7  0
  0  0  1  7  0
```

For a network with 12 hidden units, to illustrate over-fitting.

☐ Training Set : MLP (12) ☐

```
MSE per output:         0.00207232
Patterns Misclassified: 0 (out of 217)
Percentage Correct:     100%

Classification Matrix
=====================
 42  0  0  0  0
  0 42  0  0  0
  0  0 62  0  0
  0  0  0 44  0
  0  0  0  0 27
```

■ Test Set : MLP (12) ■

```
MSE per output:         0.190793
Patterns Misclassified: 42 (out of 67)
Percentage Correct:     37.3%

Classification Matrix
=====================
  4  1  0  1  0
  5  6  8  0  2
  1  5  8  4  0
  0  0  7  4  3
  0  0  2  3  3
```

③ *Radial Basis Function Network*

For the radial basis function network, the linearity of the training problem implied by choosing fixed centres permits an efficient implementation of 'leave-one-out' cross-validation [Weiss and Kulikowski

1991] for model-order selection. In figure 4.6 below, the leave-one-out cross-validation error on the RAE_PCB dataset is calculated for networks with 1 to 20 basis functions. The error when generalising to the data in RAE_AM is also plotted. The basis functions used were 'thin-plate splines' ($r^2 \log r$), and they were positioned at random over points in the dataset and then adjusted by the batch version of the K-means algorithm [MacQueen 1967; Moody and Darken 1989]. In this algorithm, the centres are initially located at random on the data points. At each iteration, every data point is 'assigned' to its nearest centre, after which the centres are then adjusted such that each lies at the mean of the data points which were previously assigned to it. This procedure is then repeated until there is no change in the centre positions. K-means is often described as a "greedy" approach to vector quantisation, in that it is fast and generates a fairly good, although generally not globally optimal, solution, by performing an equivalent local Newton minimisation of the quantisation error [Bottou and Bengio 1995].

All error values were averaged over 100 runs. To reduce confusion, error-bars have not been plotted on the graph, but the standard deviations were in the order of 0.008 for the training set and 0.011 for the test set.

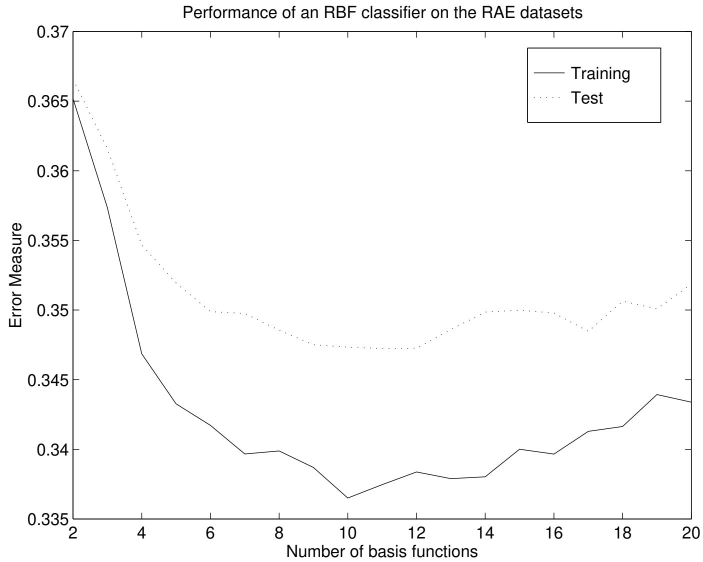

Figure 4.6: An RBF classifier trained on the RAE_PCB data and tested on the RAE_AM set. Training error is calculated using leave-one-out cross-validation.

The classification performance of a typical RBF network, with 10 basis functions, is:

☐ Training Set : RBF ☐

```
MSE per output:          0.0620856
Patterns Misclassified: 97 (out of 217)
Percentage Correct:      55.3%

Classification Matrix
=====================
 40  1  1  0  0
 16  7 14  3  2
  4  1 44 13  0
  1  2 19 20  2
  0  0  8 10  9
```

■ **Test Set : RBF** ■

```
MSE per output:          0.0712626
Patterns Misclassified: 37 (out of 67)
Percentage Correct:      44.8%

Classification Matrix
=====================
  6  0  0  0  0
  9  2  8  2  0
  0  1 14  3  0
  0  0  7  5  2
  0  0  3  2  3
```

Discussion of Classification Results

The linear model appears to perform well on the training dataset, classifying nearly 85% of the patterns correctly. However, there are very few available examples compared to their dimensionality (217 against 80), so the linear problem is not well-constrained and an apparently 'good' classification might be expected regardless of the suitability of the model. The considerably degraded performance on the test set confirms this. Note that simply 'guessing' a rating of '3' (the most prevalent class in the training set) on the test set would give 27% accuracy and is thus a good baseline for comparison.

A notable feature of the training/test error curves in figures 4.5 and 4.6 is the very low implied optimal classifier complexity. The number of hidden units in the MLP for lowest test set error was the minimum, 1, and the optimal number of basis functions appears to be 10. This apparently low number is a manifestation of a problem often referred to as *the curse of dimensionality* (a term originally coined by Bellman [1961], and an excellent exposition of its significance is given by Friedman [1994]). For genuinely high dimensional data (which the PCA results indicate is certainly the case for the RAE data considered here), the number of sample data points required to allow a faithful approximation of any underlying function is massive. In cases of inadequate data, such as for the RAE_PCB dataset, generalisation effectively becomes *extrapolation* (rather than *interpolation*), as it is highly improbable that any training data should be in the vicinity of any given test data point. In the absence of sufficient data, considerable smoothness constraints must be imposed on the model — hence the limited number of basis functions, and MLP hidden units, found to give optimal generalisation. This behaviour is an example of what is often termed the *bias-variance dilemma* [Geman, Bienenstock, and Doursat 1992]. It is necessary to bias the model (through the smoothness constraints), or the resulting function will be highly sensitive to the noise, and thus exhibit high variance over individual datasets.

The (nonlinear) neural network models perform better than their linear counterpart, classifying 45-48% of the test set correctly, compared to that of 39% of the linear classifier, which indicates that a more faithful model of the underlying relationship between data and research rating has been constructed. However, the generally poor observed performance may be seen as reinforcing the suspicion that the operation of the assessment panel would effectively introduce noise into the data, when viewed from a classification perspective. This will be considered further in Section 4.5, where an explicit feature extraction stage will be incorporated prior to classification.

4.4.3 Sammon Mapping

A Sammon mapping of the RAE_PCB dataset is illustrated in figure 4.7. The equivalent NEUROSCALE projection with $\alpha = 0$ is also illustrated (figure 4.8). The two plots should be identical if one basis function per point was used in the RBF projection; in this example, only 70 basis functions were utilised, and consequently, the STRESS of the NEUROSCALE projection is marginally greater. For this dataset, it was observed that 'thin-plate-spline' basis functions offered better (lower-STRESS) results than Gaussians, and these were used for all the NEUROSCALE projections within this chapter.

There is some notable similarity between the Sammon Mapping and the PCA projection, particularly with regard to the position of outliers. The nonlinearity of the topographic mapping algorithm enables significantly more structure to be visible, particularly along the direction of the vertical axis. Some coarse clustering of research ratings is now evident, although the classes overlap to the extent that prediction would be highly unreliable.

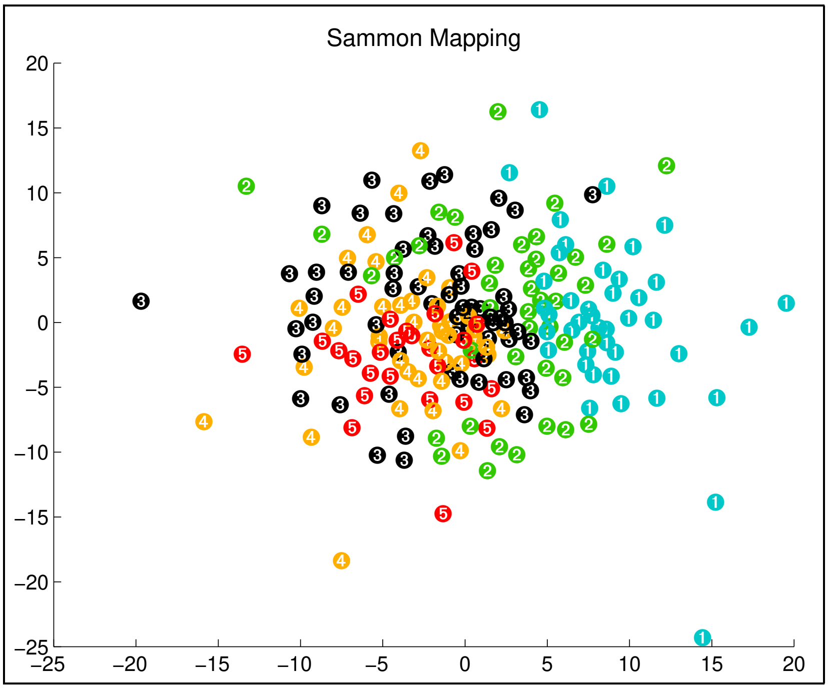

Figure 4.7: A Sammon Mapping of the RAE_PCB data. Final STRESS value = 1.118

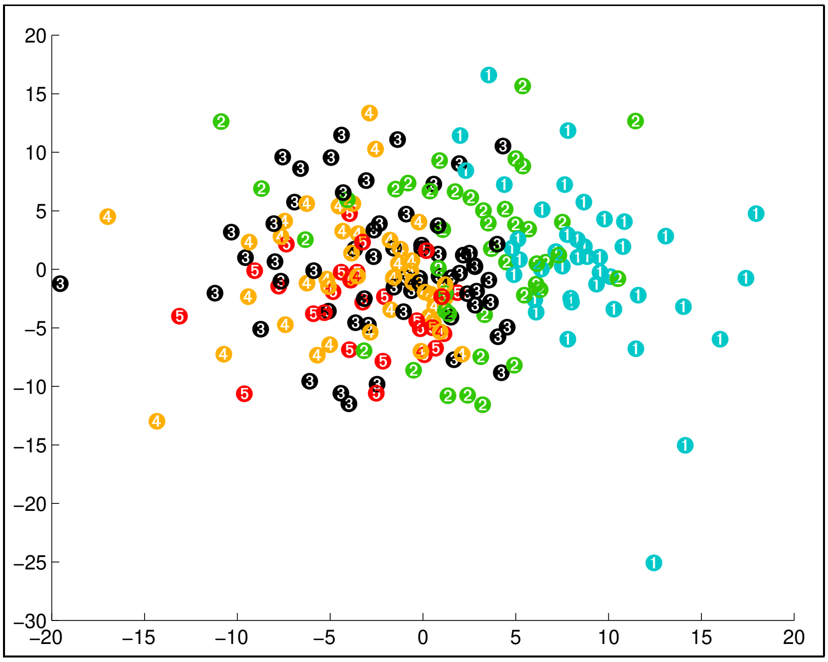

Figure 4.8: The NEUROSCALE equivalent projection for the Sammon mapping. Final STRESS value = 1.181

4.4.4 The Kohonen Self-Organising Feature Map

A 10x8 Kohonen map was applied to the RAE_PCB dataset, and the visualisation obtained using the SOM_PAK code [SOM Programming Team 1995] is displayed in figure 4.9. The nearest-neighbours to each data point in the two dimensional sheet are plotted, with multiple classes at any one node having an added displacement for the purposes of improving clarity.

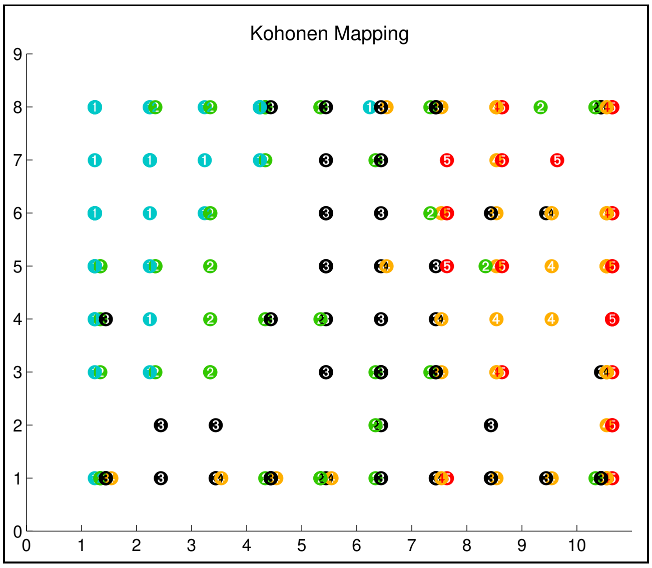

Figure 4.9: A visualisation of a Kohonen map applied to the RAE_PCB data.

The Kohonen map, like the Sammon mapping, displays little structure and considerable confusion of classes. At some nodes on the array, departments with four different research ratings are co-incident.

4.4.5 Discriminant Analysis

While the class labels of each data point are illustrated on both the principal component and Sammon projections, this information has not been exploited in the extraction of the feature spaces. If the separation of these classes is desired in feature space, then the methods of discriminant analysis may be applied. The plot in figure 4.10 is a plot onto the first two *linear* discriminant axes (these axes may also be termed the first two *canonical variates*). This represents a non-rigid linear projection which maximises the discriminant criterion tr $\left[\mathbf{S}_w^{-1}\mathbf{S}_b\right]$, as defined in Section 4.3 earlier.

A nonlinear extension of this technique may be illustrated by considering the hidden unit space of a multi-layer perceptron trained as a 1-of-5 classifier [Webb and Lowe 1990]. This plot is given in figure 4.11. Note that while the feature space was constructed to maximise the criterion tr $\left[\mathbf{S}_t^{+}\mathbf{S}_b\right]$, the matrix $\mathbf{S}_b$ is not the traditional weighted between-group covariance matrix for the target coding employed here. In fact, $\mathbf{S}_b = \sum_{k=1}^{5} n_k^2(\mathbf{m}_k - \bar{\mathbf{m}})(\mathbf{m}_k - \bar{\mathbf{m}})^{\mathrm{T}}$, where $\mathbf{m}_k$ is the mean of the hidden unit outputs for patterns from class k, $\bar{\mathbf{m}}$ is the overall mean of the hidden unit outputs, and n_k is the number of patterns in class (with research rating) k. The conventional $\mathbf{S}_b$ matrix is only weighted by the factor n_k, and not the squared value.

The linear discriminant plot reveals some considerable structure in the data. Significantly more distinct clustering is evident, when compared with previous unsupervised feature spaces, although there is still considerable mixing of research ratings at the borders. Furthermore, there is a visible structural *ordering* of classes within the plot.

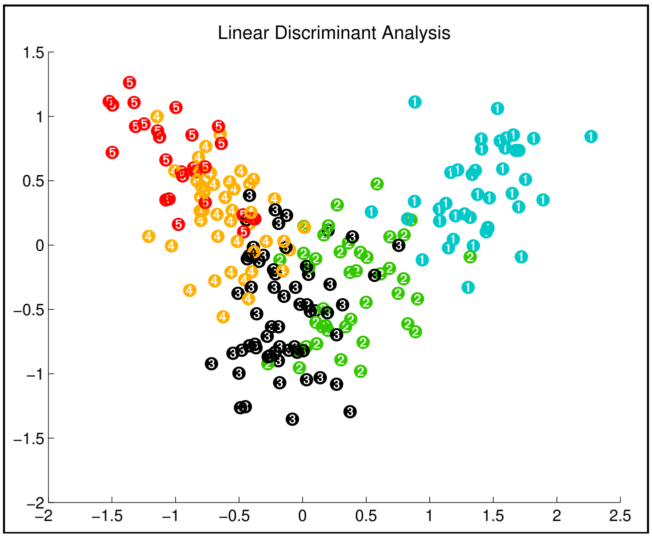

Figure 4.10: A plot of the first two canonical variates of the RAE_PCB data.

In the nonlinear case, to maximise the discriminant criterion the points have been approximately mapped to the corners of the -1/1 square. Some general misclassification is evident and points for research rating '4' and '5' occupy the same corner. The network discriminant function seeks to maximise the total covariance, while minimising a measure of the between-class covariance. Total covariance will be maximised by placing the classes as mutually distant as possible — in the corners — but, because there are only two hidden units in the MLP, only four such classes can be so placed without reducing the between-class covariance. The need to maximise the total covariance dominates in this case and hence the confusion of two classes. It is interesting to note that one of the pair is the least represented class, the research rating '5', because, according to Webb and Lowe [1990], "networks trained with a one-from-n coding bias strongly in favour of those classes with largest membership".

Note that the interpretation of this plot can be aided by referring to the equivalent misclassification matrix based upon the illustrated hidden unit space which is given in Section 4.4.2.

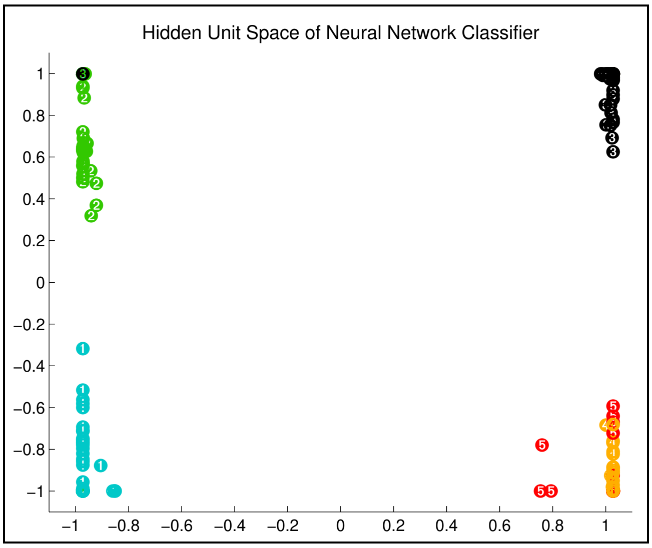

Figure 4.11: The hidden unit space of a MLP classifier trained on the RAE_PCB data.

4.4.6 NEUROSCALE with Supervisory Information

To incorporate some supervisorial, or preferential, input to the topographic projection, it is necessary to define the set of subjective dissimilarities for every pair of data points. These may in turn be derived from some measure of dissimilarity of the respective research ratings. The dissimilarities between research ratings are chosen to reflect both the prior knowledge concerning the assessment process and what is considered to be the preferred structure of the feature space. Based on the information available regarding the criteria for the assignment of ratings, it is intuitive to consider that a department with a rating of '2' is as dissimilar from a '1' as it is from a '3', and furthermore, a '3' is twice as dissimilar from a '1' as it is from a '2'. These assumptions imply a geometrically linear ordering of research ratings thus:

which in addition, also reflects the resultant funding, since *funding* $\propto$ (*rating* -1). This in turn implies that a simple matrix may serve to characterise the subjective dissimilarities between research ratings:

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

so that $c_{ij} = |i - j|$ is the subjective dissimilarity between points of rating i and j. Note that the scaling of this matrix is arbitrary, as it is only the *relative* differences that are of interest. Thus for example, entries $c_{24} = 2$ and $c_{25} = 3$ 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 $\mathbf{C}$ such that the average inter-point subjective dissimilarity is equal to the average inter-point Euclidean distance. The matrix $\mathbf{C}$ then provides sufficient information to calculate a value of subjective dissimilarity, s_{ij}, between every pair of input data points, thus defining the subjective metric.

To illustrate a feature space influenced by both the objective (Euclidean distance in input space) and subjective metrics, figure 4.12 is a plot of the NEUROSCALE projection with $\alpha = 0.5$. Because of the scaling of the matrix $\mathbf{C}$, a choice of $\alpha = 0.5$ implies a (very) approximate balance between the twin, objective and subjective, metrics.

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 topology 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 seeks to explicitly separate, nonlinearly, points with different research ratings.

The observations in the above paragraph concern the effect on the projection of the additional, supervisorial, knowledge. In addition to this subjective element, and because of the intermediate value of α (0.5), some of the geometric structure of the original data is retained in the feature space. This implies that useful information may be inferred from the locations of individual points in that space, as this structure reflects to some degree the topology of the input space. For example, in the plot of RAE_PCB in figure 4.12, it may illuminate potential anomalies in the awarding of research ratings. This considered, the plot in figure 4.13 highlights four particular departments in the projection. Each department 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:

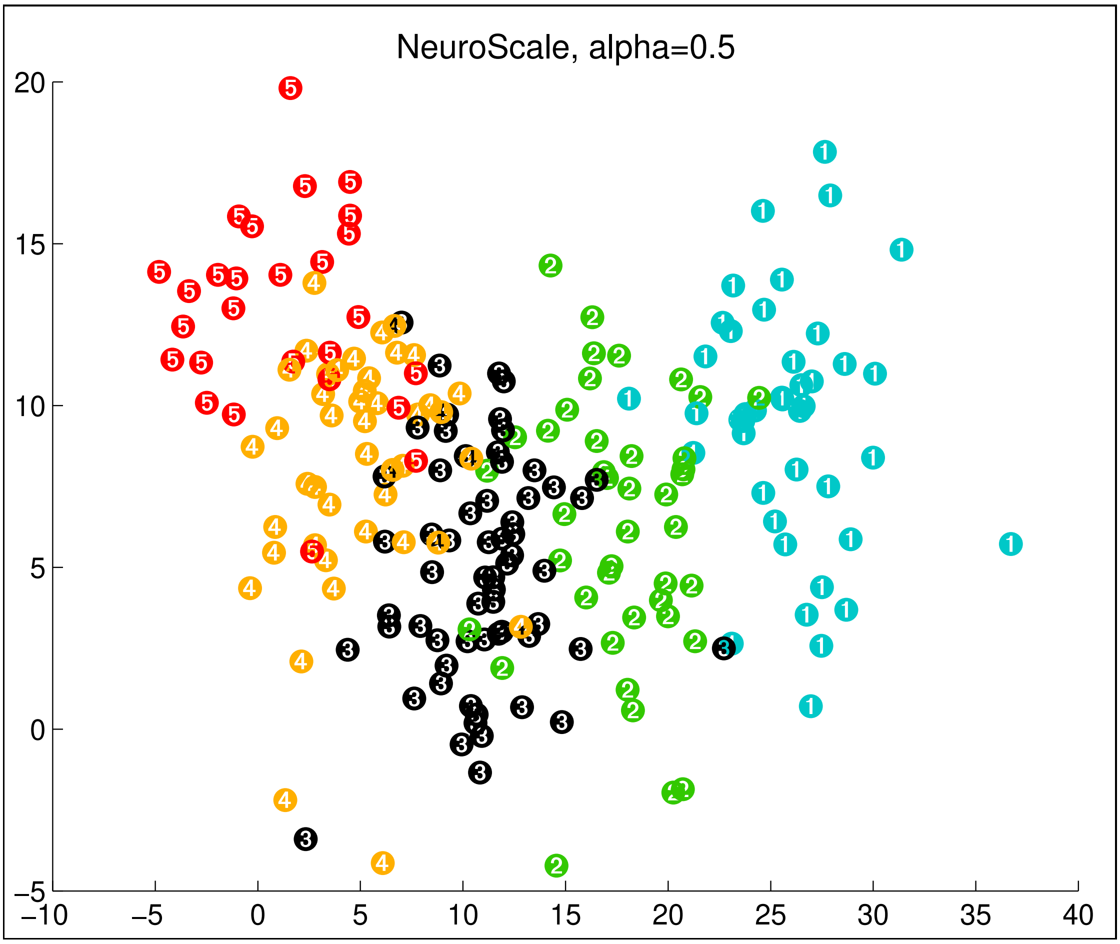

Figure 4.12: A feature space extracted by the NEUROSCALE technique with supervisorial influence.

① Physics at Heriot Watt University, Edinburgh, which has received a '5' while lying amongst a cluster of '4's.

② Physics at Queen's University, Belfast, which has also received a '5' while lying amongst departments awarded '4's and '3's.

③ Physics at Stirling University, which received a '4' while lying amongst a cluster of '3's.

④ 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 4.11, the location of individual points with respect to their neighbours is largely artefactual. Due to the topographic constraint upon the feature space of figure 4.12, there will be an element of structural information therein. In the cases of the individual points highlighted above, further evidence for the structural meaning can be elucidated by considering the classification results of section 4.4.2. Table 4.2 below shows the predicted ratings of the four departments above according to the best linear, MLP and RBF models.

In general, and particularly for the neural networks which demonstrated lower error on the test set, 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. Although these four particular classifications may appear anomalous, there may well be good explanations. Firstly, it is noticeable

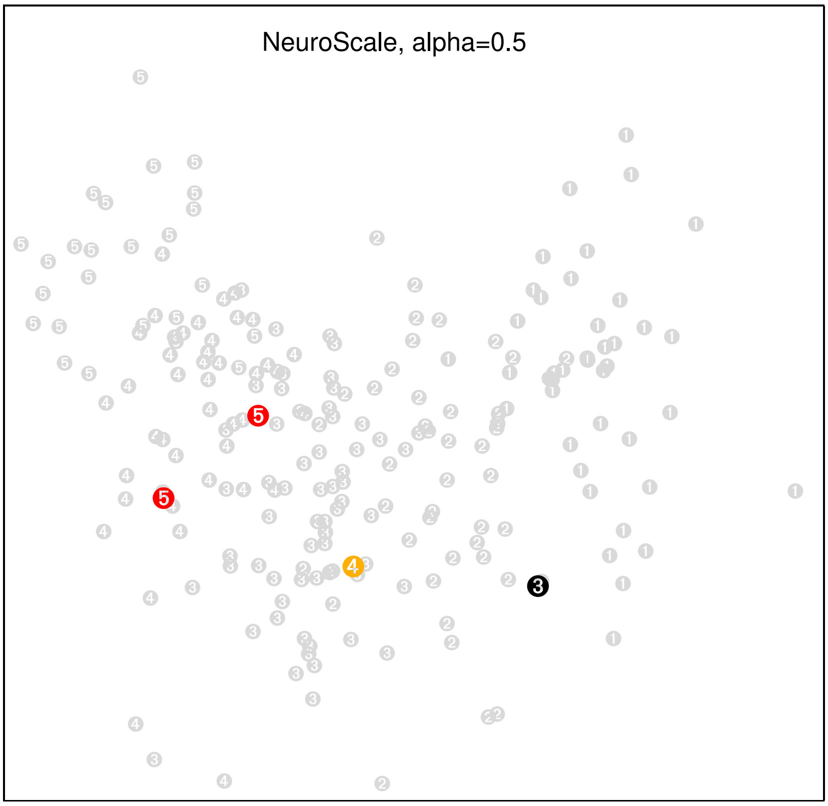

Figure 4.13: The $\alpha = 0.5$ feature space extracted by NEUROSCALE, with four departments highlighted whose awarded ratings appear anomalous.

that all four departments are of the Physics unit of assessment. The panel which awarded the ratings for this subject may have had slightly different criteria to those for Chemistry and Biological Sciences. Equally, the panel has access to additional qualitative information which, in the case of these four departments, may have influenced its judgement.

As a final example feature space generated by the NEUROSCALE technique, the illustration in figure 4.14 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 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/scaling procedure. As the number of basis functions in the transformation is considerably fewer than the number of points, there is not sufficient flexibility in the model to precisely locate the points and satisfy the subjective metric constraint.

	Actual Rating	Linear	MLP	RBF
Physics, Heriot-Watt	*5*	5	4	3
Physics, Queen's	*5*	4	4	3
Physics, Stirling	*4*	4	4	3
Physics, Westminster	*3*	1	2	1

Table 4.2: Predicted and actual ratings for 'inconsistent' departments.

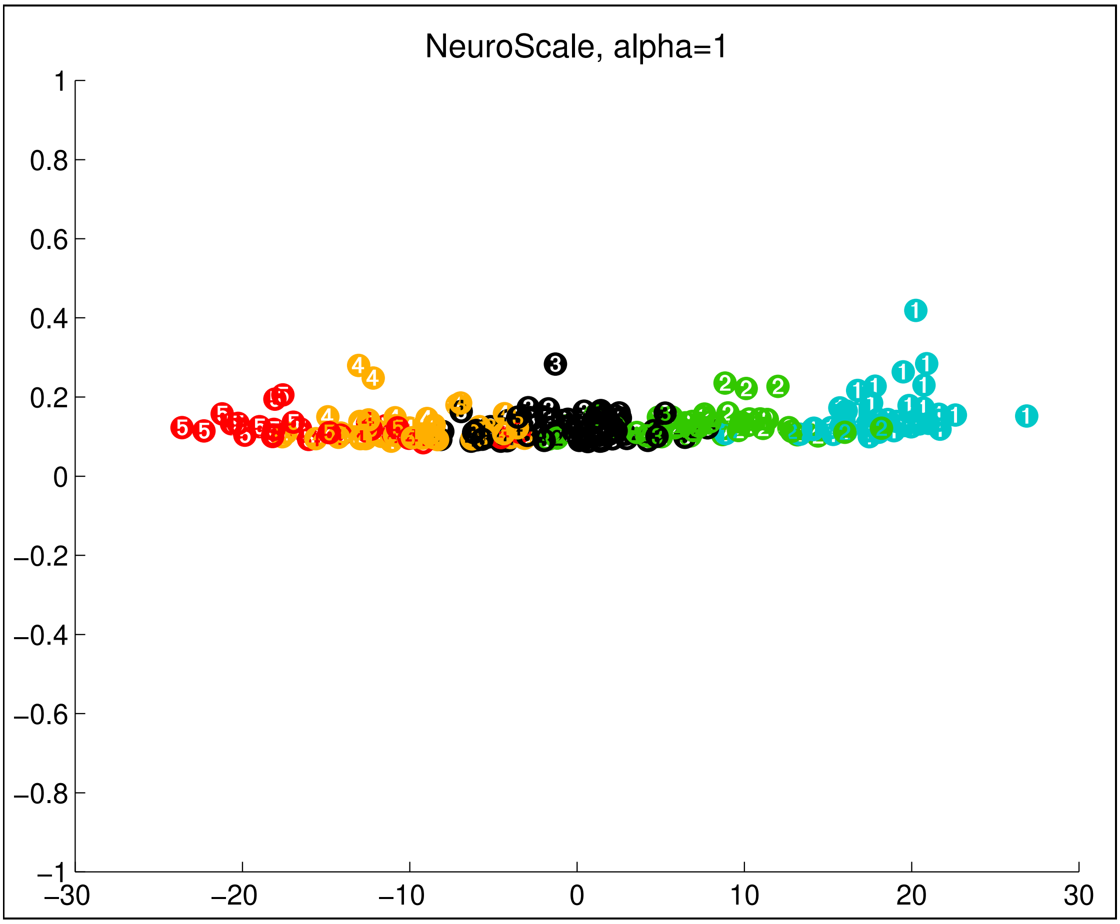

Figure 4.14: A feature space extracted by the NEUROSCALE technique determined by the subjective metric alone. This may be considered a fully supervised variant of the projection.

4.4.7 Generalisation to the Test Dataset

One of the key benefits of the NEUROSCALE approach to both topographic mapping and Sammon's projection is that the definition of the transformation, as effected by the RBF, permits new, unseen, data to be projected. Such a projection is illustrated in figure 4.15, for the $\alpha = 0.5$ experiment trained on the RAE_PCB dataset and tested on the RAE_AM set. As comparison, in figure 4.16, a projection of the test data onto the linear discriminant axes of figure 4.10 is also given.

As in the case of the training plot, for the RAE_PCB dataset, the NEUROSCALE projection exhibits better clustering and separation of research ratings. While there is still significant confusion at the borders, there is sufficient structure present in the plot to allow judicious inference of research ratings. Subjectively, the NEUROSCALE test projection appears to retain more of the structure of its respective training plot than the linear discriminant counterpart, indicating that this latter linear technique had not constructed such a good representation of the data. This contrast will be made more explicit in the next section.

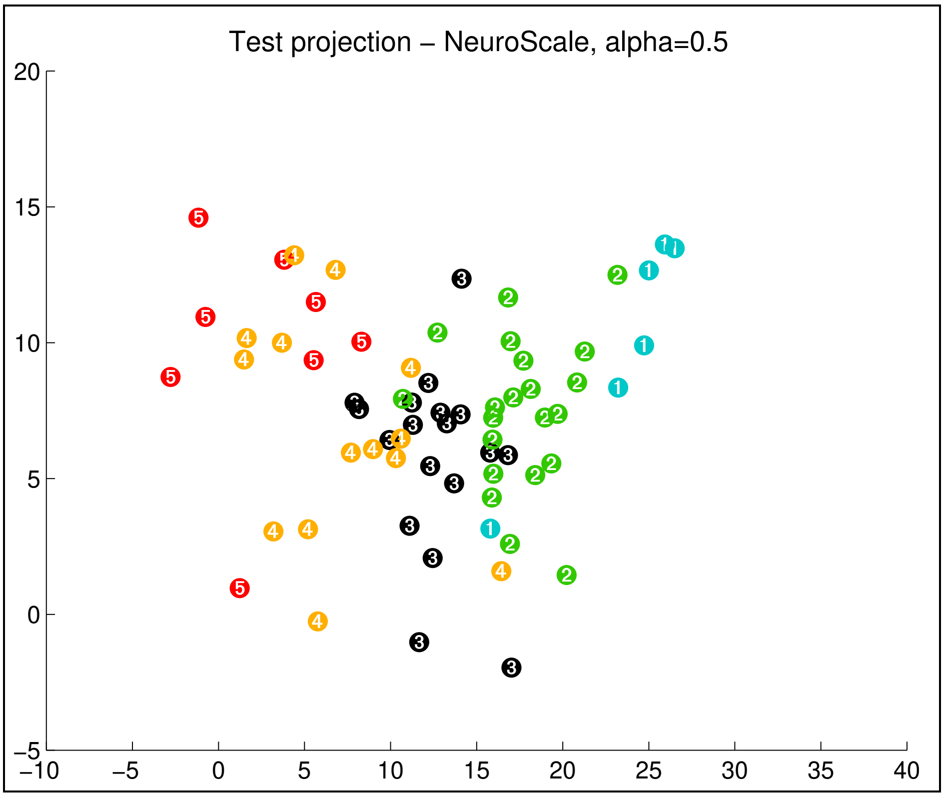

Figure 4.15: The RAE_AM test dataset transformed by the NEUROSCALE, $\alpha = 0.5$, technique previously trained on the RAE_PCB data.

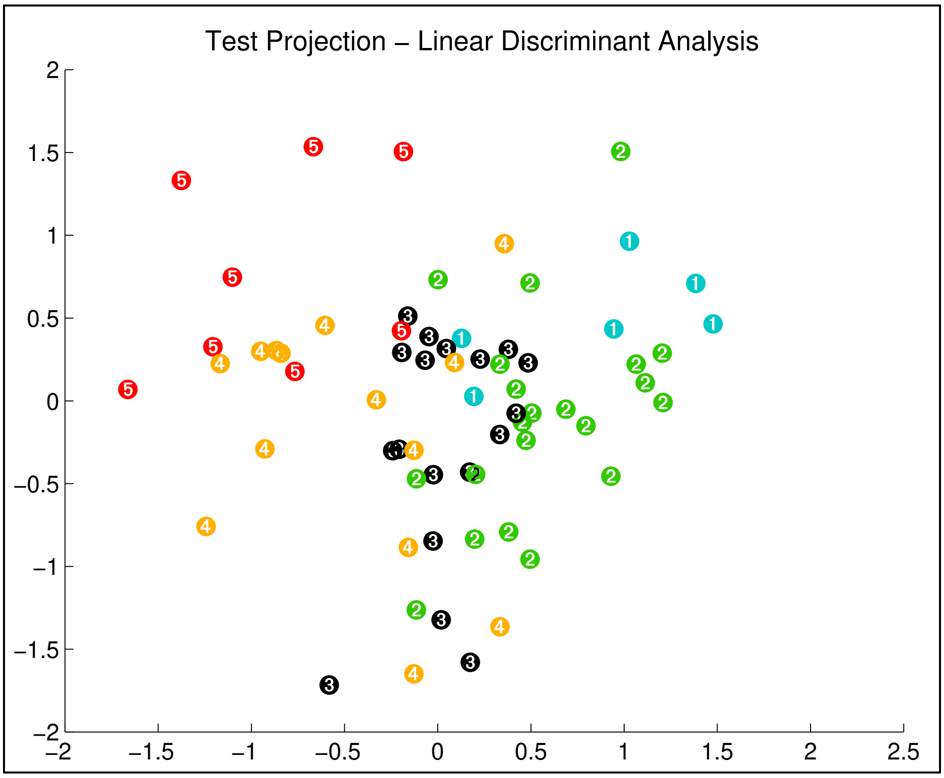

Figure 4.16: The RAE_AM test dataset projected onto the first two linear discriminant axes.

4.5 NEUROSCALE Pre-Processing for Classification

The emphasis of the previous two subsections was on the use of NEUROSCALE as a data analytic tool. However, in the discussion of Section 4.3, it was mentioned that in classification problems, for optimal performance, it is often better to extract, or select, an appropriate subset of features from the data prior to the classification stage. In fact, many texts include a schematic of a conventional prediction system similar to that shown in figure 4.17 below [Duda and Hart 1973; Fukunaga 1990; Weiss and Kulikowski 1991; Bishop 1995], which incorporates an explicit *data pre-processing* stage.

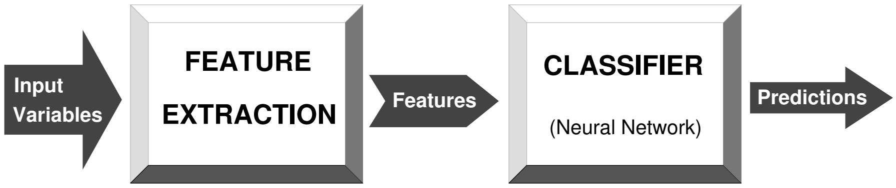

Figure 4.17: Schematic of a general classifier system.

This section considers the construction of an RBF classifier for the RAE data again, whose inputs are no longer taken from the complete set of raw variables as in Section 4.4.2, but instead from the feature spaces illustrated in the previous section as extracted by the following techniques:

① Generalised Linear discriminant analysis

② NEUROSCALE: $\alpha = 0$

③ NEUROSCALE: $\alpha = 0.5$

④ NEUROSCALE: $\alpha = 1$

4.5.1 Experimental Prediction Models

The neural network used for classification was an RBF with thin-plate spline basis functions ($r^2 \log r$), with the centres located within the data by the k-means algorithm, detailed in Section 4.4.2. The number of centres was chosen individually for each prediction model using leave-one-out cross-validation. For each pre-processing method, classification results are shown for 'typical' networks (specifically, the network that gave the median error over 10 runs of each model).

① Generalised Linear Discriminant Analysis

□ Training Set

```
MSE per output:          0.0359165
Patterns Misclassified:  59 (out of 217)
Percentage Correct:      72.8%

Classification Matrix
=====================
 41  1  0  0  0
  4 27 11  0  0
  0  7 49  6  0
  0  1 11 28  4
  0  0  0 14 13
```

■ Test Set

```
MSE per output:          0.0641893
Patterns Misclassified:  32 (out of 67)
Percentage Correct:      52.2%

Classification Matrix
=====================
  4  2  0  0  0
  6 11  3  1  0
  0  6  8  4  0
  1  1  5  7  0
  0  0  0  3  5
```

② NEUROSCALE : $\alpha = 0$

□ Training Set

```
MSE per output:          0.0592697
Patterns Misclassified:  107 (out of 217)
Percentage Correct:      50.7%

Classification Matrix
=====================
 32  8  2  0  0
 11 16 12  0  3
  4  3 42  8  5
  1  0 23 14  6
  0  0 11 10  6
```

■ Test Set

```
MSE per output:          0.0657081
Patterns Misclassified:  35 (out of 67)
Percentage Correct:      47.8%

Classification Matrix
=====================
  6  0  0  0  0
  5  9  6  1  0
  1  2 11  3  1
  0  0  9  4  1
  0  0  3  3  2
```

③ NEUROSCALE : $\alpha = 0.5$

☐ Training Set ☐

```
MSE per output:         0.0293056
Patterns Misclassified: 38 (out of 217)
Percentage Correct:     82.5%

Classification Matrix
=====================
 38  4  0  0  0
  2 32  8  0  0
  0  4 52  6  0
  0  0  5 38  1
  0  0  0  8 19
```

■ Test Set ■

```
MSE per output:         0.0484208
Patterns Misclassified: 20 (out of 67)
Percentage Correct:     70.1%

Classification Matrix
=====================
  5  1  0  0  0
  1 18  2  0  0
  0  4 14  0  0
  0  1  7  6  0
  0  0  0  4  4
```

④ NEUROSCALE : $\alpha = 1.0$

☐ Training Set ☐

```
MSE per output:         0.0323057
Patterns Misclassified: 51 (out of 217)
Percentage Correct:     76.5%

Classification Matrix
=====================
 39  3  0  0  0
  3 32  7  0  0
  0  6 49  7  0
  0  0  8 30  6
  0  0  2  9 16
```

■ Test Set ■

```
MSE per output:         0.0497819
Patterns Misclassified: 26 (out of 67)
Percentage Correct:     61.2%

Classification Matrix
=====================
  4  2  0  0  0
  0 18  3  0  0
  0  4 14  0  0
  0  1  7  3  3
  0  0  0  6  2
```

4.5.2 Discussion

The results of the various feature extraction approaches for pre-processing the data can be summarised by the histogram in figure 4.18 below, which illustrates the classification performance of the techniques ①-④ above and compares them with the direct methods of Section 4.4.2.

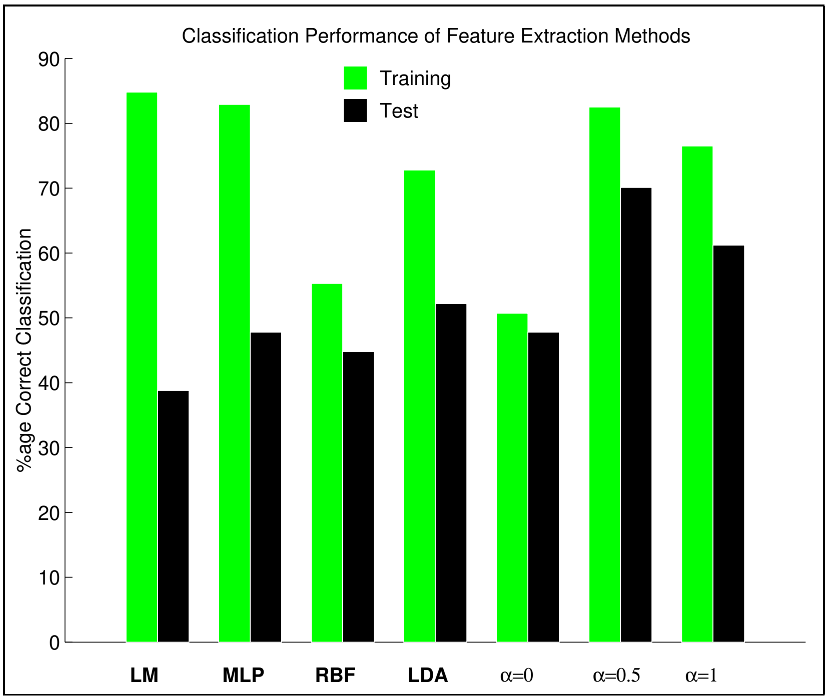

Figure 4.18: A histogram comparing the relative performance of various classifiers on the RAE data. The left-hand three models are those constructed directly from the data. The right-hand models are RBFs applied after a feature extraction stage.

The prediction rates illustrated by figure 4.18 confirm the proposition that additional input information can reduce classifier quality, particularly in the case of limited data. The classification rate on the test set for the worst predictor using pre-processed data (NEUROSCALE with $\alpha = 0$) is exactly equivalent to the best predictor applied to the entire set of input variables. Clearly, the feature extraction stage is highly appropriate for this particular task. Furthermore, effecting this pre-processing with the NEUROSCALE, $\alpha = 0.5$, network, produced the best results, classifying 70.1% of the test dataset.

That the hybrid supervised/unsupervised model ($\alpha = 0.5$) was optimal, of the three variants tested here, is particularly interesting. For the unsupervised version ($\alpha = 0$), it should be clear from studying the configuration of figure 4.8 earlier that considerable misclassification was inevitable based on those extracted features, both for the training and test sets. Note that the test error is only marginally greater than the training error for this model. Further insight into this phenomenon will be provided in Chapter 6.

The purely supervised NEUROSCALE variant ($\alpha = 1$), which might naively have been expected to generate the best results, is outperformed by a model that incorporates some element of topographic information ($\alpha = 0.5$). The solution configuration for $\alpha = 1$ is a straight line, which is closely approximated in figure 4.14, and the addition of some structural component evidently provides useful information in the second dimension. This further prompts the question of which dimensionality of feature space might be optimal in this pre-processing application, and is an obvious area for more detailed study.

Further insight into the dependence of classifier performance on the value of α is given in figure 4.19 below. This illustrates the classification rate of a typical classifier for values of α from 0 to 1 in steps of 0.1. Note that although these values are for a single run (averaging over many such runs of both the scaling and classification procedure would be highly computationally expensive), there is still an evident trend which implies that there is some mid-range value of α which leads to optimal performance.

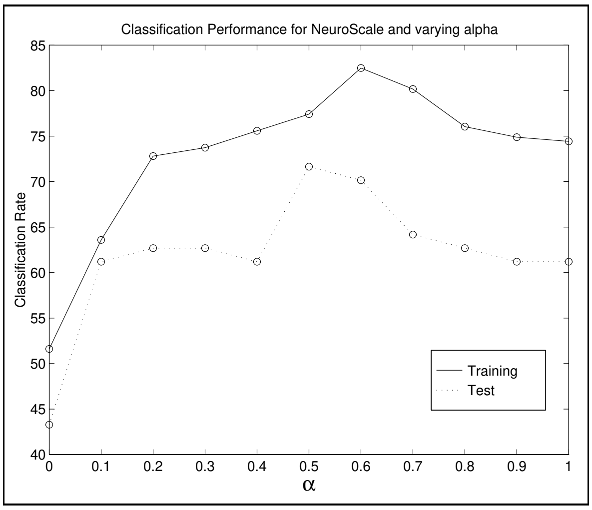

Figure 4.19: Variation of classification performance with α.

It is possible to generate a two-dimensional $\alpha = 1$ supervised feature space by altering the subjective dissimilarities. For example the matrix of inter-class (research-rating) dissimilarities

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

will produce such a space. The matrix $\mathbf{C}'$ simply attempts to cluster points from identical classes and separate those from different classes by the unit distance. There is therefore no implied *ordering*, as embodied in the matrix $\mathbf{C}$ used for earlier NEUROSCALE projections, and the resultant $\alpha = 1$ configuration is of a circular nature. However, in this experiment, there was no significant improvement in classification error observed when using such a matrix.

This implies that retention of some level of structure in the feature extraction phase is of benefit to later classification, and may in fact be more useful than a particular explicit discrimination criterion. This confirms a major advantage of a topographical transformation such as NEUROSCALE, since conventional approaches to structure-preservation, such as the Sammon mapping, do not possess a generalisation property, which precludes their inclusion in a predictive model.

The author is only aware of two examples in the literature where experiments have been conducted that exploit subjective class information within a parameterised topographic map before the discrimination process. Koontz and Fukunaga [1972] considered simple two-dimensional two-class problems, mapped also to two dimensions using their 'distance-difference' mappings. It was observed that classification performance improved as the discriminating element (J_{SE}) of the mapping was increased, for both separable and non-separable problems. Cox and Ferry [1993] also employed similar and "somewhat contrived" datasets to illustrate the efficacy of their MDS-based mapping. Additionally, they applied the method to a real 15-dimensional two-class dataset and showed that applying the MDS mapping prior to discrimination with Fisher's function improved classification accuracy compared to direct application of that function to the original data. This, however, should be expected as the pre-processing implies that additional nonlinear behaviour may be exploited in the discrimination process, whilst Fisher's discriminant is of linear character.

4.6 Conclusions

4.6.1 Exploratory Data Analysis

Analysing data taken from the 1992 Research Assessment Exercise is a particularly interesting, and difficult, problem. The data is of very high dimension, and this, coupled with the "noise" implied by the subjective peer assessment component of the exercise, makes the research ratings difficult to predict from the quantitative data alone. This was underlined in section 4.4.2, where, with the models investigated, it was found to be impossible to predict more than half of an unseen test set correctly.

With respect to visualising the data in two dimensions, several illustrative feature spaces, of linear–nonlinear and supervised–unsupervised character, produced by standard methods were generated. Of the linear techniques, the principal component projection failed to elucidate any useful structure from the data, while by exploiting the associated class information, the linear discriminant space separated the research ratings into approximate clusters. This latter technique could then project previously unseen data, although the clustering was less pronounced.

Of the nonlinear techniques, the Sammon Mapping is an improvement upon the unsupervised PCA projection, exhibiting more structure, but again offered little promise for exploratory analysis. Equally so is the feature space produced by the NEUROSCALE technique with $\alpha = 0$, although this latter approach would permit subsequent projection of the test data set. The established neural network approach to topographic mapping, the Kohonen map, proved to be ill-matched to the topology of the data and thus of limited utility in this application.

The most effective visualisation of the data presented in this chapter was that obtained from NEUROSCALE with an intermediate value of $\alpha = 0.5$, illustrated in section 4.4.6. The merit of combining both unsupervised and supervised — or structural and subjective — flavour in the feature space is made explicit. The subjective knowledge imposes a clustering and ordering constraint that facilitates analysis with respect to classification of departments. The generalisation capability of this approach was also clearly shown, illustrating a better projection than that achieved by LDA, and which showed potential (see below) for subsequent decision making or inference. Previously unseen data, which is transformed by the network, may be analysed in that space. The supervisorial content of the projection allows a prospective prediction to be made, and the structural content implies that the confidence of that prediction can be gauged according to the proximity of the test point to other clusters.

Equally, by retaining some of the structure of the original data, it is also possible to highlight apparent anomalies and inconsistencies in the assessment process, and this evidence was supported by the results of the previous classification experiment.

4.6.2 Data Pre-processing for Classification

The results from Section 4.5 confirm that improved classifier performance can be attained by extracting an appropriate set of features from the data *a priori*. Given this approach, it was possible to correctly predict the rating awarded to 70% of the Applied Mathematics departments (and 82% of those awarded ratings 1, 2 and 3). Note that these ratings were determined by a distinct assessment panel from the Physics, Chemistry and Biological Science departments (the data upon which the classifier was trained), and that it had access to additional qualitative information in making its decision. Despite this, and the large number of available input variables, it is evident that considerable structure *is* present in the data, in terms of the relationship to the research rating ultimately awarded. This conclusion supports the argument that, from the perspective of the funding councils, collection of the research data is, in part at least, a worthwhile exercise. However, as considered by Taylor [1995], this chapter has not addressed the question of *which* of the large number of variables are relevant.

It would be interesting to attempt, using the best model above, to predict the outcome of the 1996 exercise for those relevant departments. Unfortunately, changes in the assessment process and the supplied data would probably render this exercise somewhat difficult and ineffective.

Of particular interest from the pre-processing experiments, is the fact that the optimum feature extraction method, of those investigated, was the NEUROSCALE transformation with $\alpha = 0.5$. This approach outperformed its purely unsupervised and purely supervised relatives. Clearly, the structural influence upon the projection is beneficial to classification, and, counter-intuitively, more effective than supervisorial criteria alone, whether linear (LDA) or nonlinear (MLP and NEUROSCALE with $\alpha = 0$). The questions of exactly which value of α and output dimension q are optimal in this experiment remain open.

Two similar investigations into such topographic pre-processing in the literature are too simplistic for serious comparison, and the results for NEUROSCALE on the RAE data would appear significant. However, the primary focus of this thesis is concerned with topographic mappings for generic feature extraction, rather than tailored extraction for the purposes of any specific prediction or classification task. There thus remains much scope for further investigation of this aspect of NEUROSCALE.