Chapter 8

Conclusions

8.1 Overview

The introductory chapter of this thesis began by considering the generic problem of *information processing* and more specifically, how dimension-reduction techniques could ease the interpretation of complex data sets. Building on this theme, the objective of subsequent chapters has been two-fold — firstly to motivate the development of a particular neural network approach to dimension-reducing topographic mapping in the information-processing context, and secondly, to improve the theoretical understanding of the design, training and application of such models.

8.2 Why NeuroScale?

In Chapter 2 it was reasoned that in data-analytic contexts, the Sammon Mapping was the most effective strategy for topographic dimension reduction. This is an important conclusion in itself, considering the widespread and often inappropriate application of the Kohonen self-organising feature map. Established theoretical properties and experimental evidence, both in existing literature and presented within this thesis, combine to support this argument. The main advantage of Kohonen's approach is computational, because realistically, application of the Sammon Mapping is restricted to fewer than 1000 data points. However, the computational tractability of Kohonen's algorithm should not be allowed to disguise its flaws as an information processing paradigm.

The feed-forward neural network topographic mapping technique introduced in Chapter 3, NeuroScale, was thus based upon the Sammon Mapping and utilises a radial basis function neural network. Because of this neural network element, it offers the capability of generalisation to new data — a feature absent from Sammon's original algorithm which is effectively a look-up table approach.

An important extension embodied in NeuroScale is the capacity to exploit additional information in the mapping process. In standard approaches to topographic mapping, the geometry of the output space is determined solely according to some conventional metric (generally Euclidean) defined over the data space. If alternative information is available — such as class labels — then this may be allowed to influence the mapping (in order to emphasise clustering, for example). Previous implementations of this concept have been largely heuristic, whereas within NeuroScale, the extra information is embodied as an *additional metric*. This dual metric approach then allows the two classes of information to be combined variably, and importantly, in a *consistent* manner. At one extreme is the purely unsupervised mapping, based exclusively on the geometry of the data in the input space, while at the other extreme is a supervised variant based exclusively on the additional metric, which can be considered to represent the 'preference' of the mapping process. As a result of this mechanism, enhanced visualisation spaces may be derived which can be considerably more informative than those extracted utilising conventional topographic criteria.

That there is genuine merit in the generation of such hybrid supervised/unsupervised feature spaces was demonstrated in Chapter 4, which comprised a study of data taken from the 1992 Research Assessment Exercise. When compared with other established feature extraction approaches on this difficult high-dimensional real-world dataset, the NEUROSCALE mapping that included an element of class information gave the most useful visualisation. Instead of the confusion of the five classes of interest, the use of a metric which preferred a linear ordering permitted considerably more structure to be elucidated. However, because the mapping still contained a significant geometric element, it was possible to discover apparent inconsistencies in the awarding of research ratings from the NEUROSCALE projection of the training data — inconsistencies that were supported by independent classification experiments. The supervised component of the mapping also implied that the generalisation projection to previously unseen data offered better potential for subsequent prediction. Indeed, a classifier based on this feature space was able to correctly predict the rating awarded to 70% of the test set, which was superior to other illustrative classification examples.

On the basis of the results presented in Chapter 4 and elsewhere in the first half of this thesis, NEUROSCALE offers considerable potential as a tool for the visualisation and exploratory analysis of data. That this proposed topographic model utilises a feed-forward neural network naturally raises certain specific questions concerning its application; questions which relate closely to neural networks employed in more conventional rôles. Of particular relevance is the need to determine effective training techniques, appropriate model complexity and values of other model parameters, and these issues were considered in the second half of the thesis.

8.3 Theoretical Issues

While the presented theoretical analysis should apply to general topographic RBF transformations, all experiments utilised *Gaussian* radial basis functions. This is a reflection of the ease of differentiability of that function, rather than any property specific to its use in the NEUROSCALE model. Also, the emphasis of the final chapters was on exclusively objective mappings, although some results are applicable to the case where subjective information is incorporated and these are indicated in the text.

Local minima. In terms of training the NEUROSCALE model, it is already well known that Sammon's STRESS measure exhibits many sub-optimal local minima for even the simplest of problems. Further evidence of this effect was presented in Chapter 6, where it was seen that NEUROSCALE models with significant values of basis function width 'σ' were seemingly unaffected by these minima. In fact, it was shown that the local minima found by Sammon's algorithm were generally *unsmooth* transformations of the input data and unrealisable by the RBF within the NEUROSCALE architecture. This result is very significant as the proliferation of sub-optimal local minima is a major practical disadvantage of Sammon's technique.

However, by way of balance, NEUROSCALE was seen to introduce a new class of minima as a result of those very smoothness properties of the network. When adopting a standard nonlinear optimisation approach to network training, the attained final STRESS minimum occurred at some higher value than the best obtainable by the Sammon Mapping (although, in that latter approach, many runs might have been necessary to find that particular minimum). This effect is directly related to the smoothness of the RBF transformation, and as $\sigma \to 0$, it vanishes and NEUROSCALE tends to approximate the Sammon mapping more directly and so becomes susceptible to all its associated local minima also.

This reveals a clear trade-off. Large σ gives few local minima problems, but the STRESS of the ultimate solution is inferior to the best obtained by the Sammon Mapping. As σ decreases, this optimal STRESS value improves, but additional sub-optimal minima as exhibited by the Sammon Mapping are then introduced.

Model complexity. A significant feature of the NEUROSCALE model is its facility for generalising to previously unseen data. When applying neural network models in other domains, it is known that over-complex models generalise poorly in the presence of noise, as they tend to over-fit the data. By contrast, in Chapter 6, it was seen that test STRESS of NEUROSCALE models was relatively insensitive to the number of basis functions within the network — even to the point where as many basis functions as data points were utilised. There are two key underlying points which combine to explain this counter-intuitive behaviour. Firstly, it is possible to derive significant insight into the necessary form of the network function independent of the data. Secondly, there is a regularising component automatically incorporated in the NEUROSCALE training algorithm which, given this knowledge of the desired functional form, proves to be highly appropriate.

Taking the first point, for topographic mappings it was reasoned that for good generalisation the transformation effected by the neural network (or indeed, any functional model) should ideally be smooth (have zero second- or higher-order derivative terms) and an expression was determined for the necessary gradient of the network function. To underline this, experimental evidence was presented illustrating that for networks exhibiting identical training STRESS on a data set, those with higher curvature (second-order terms) gave correspondingly higher errors on unseen test data. Most compelling was an example where networks of differing curvature, producing apparently identical configurations from a training set, exhibited considerable difference in error on an identical test set. The key to this seemingly inconsistent behaviour is to understand that there are many configurations of points that give rise to identical measures of STRESS, and these simply correspond to arbitrary rotations and/or translations of any one particular configuration. Furthermore, the smoothness of any particular transformation depends, nonlinearly, on this rotation and translation.

This is a highly relevant observation, as one existing approach to producing a transformational topographic mapping is to generate a configuration using Sammon's standard procedure, and then fit a parameterised model to the resulting solution *a posteriori*. However, the smoothness, and therefore the quality of generalisation that might be expected, of such a network is generally arbitrary. By contrast, NEUROSCALE networks exhibited lower curvature and test STRESS than these *a posteriori* models in *all* examples. An important result therefore of Chapter 6 was to explain this behaviour in terms of the learning dynamics of the relative supervision algorithm. It was seen that NEUROSCALE models trained in this way automatically tend to generate output configurations that reduce the sum-of-squared weight values within the network. This effective *weight decay* or *self-regularisation* explains why NEUROSCALE models had appeared largely insensitive to their complexity and were observed to generalise better that identical *a posteriori*-trained networks. While the extent of this implicit regularisation is not explicitly controllable (although it increases with σ), it is important to underline that its effect is not to reduce the squared-weights at the expense of increasing the training error, but rather to minimise training error while at the same time seeking a rotation and translation of the final output configuration with associated lower values of the weights.

Training Algorithms. This feature of the NEUROSCALE training algorithm is highly effective, and explains the apparent ease with which test projections with low STRESS were obtained throughout the thesis. However, this regularising effect is a function of the smoothness of the network and as already observed, there is a penalty to be paid in having too large a value of σ as the minimum obtainable STRESS on the training set deteriorates. In Chapter 7, a new training method was presented — the "shadow-targets" algorithm. Its name refers to the fact that the algorithm effectively shadows Sammon's mapping algorithm, but with the smoothness of the RBF incorporated at each step. An unfortunate consequence of minimising a STRESS measure within NEUROSCALE is that the benefit of linearity in the RBF model is lost as the error measure introduces quartic terms. The shadow-targets algorithm effectively decomposes the training problem into two steps — one of which is linear. Because of this approach, the new algorithm was an order of magnitude more efficient at reaching a STRESS minimum, and was applicable to both unsupervised and supervised mappings, although it did not incorporate the implicit weight decay of previous methods. However, the algorithm does not suffer from the increased minimum problems also associated with those techniques and is therefore able to reproduce the best observed minimum generated by the Sammon Mapping. While, for equivalent values of σ, generalisation was seen to be poorer than that illustrated for models trained using relative supervision, this new approach allows large values of σ to be chosen. Such values of basis function width produce smoother transformations, and it was relatively simple to find a value of σ which gave lower training and test STRESS than that of previous approaches. From the evidence of this thesis, the shadow-targets training algorithm should therefore be the best method for training a topographic feed-forward neural network, given appropriate choice of σ.

8.4 Directions for Future Research

Applications. As underlined by Friedman [1994], many computational learning methods can be shown to be optimal under the appropriate circumstances. Therefore, while the results on the RAE data in Chapter 4 showed considerable promise for the NEUROSCALE approach, there is always scope for application to further datasets, particularly those with alternative forms of subjective information.

Relevance of theoretical results to subjective mappings. Much of the insight derived into the issue of model complexity in Chapter 6 was on the basis of purely objective, $\alpha = 0$, mappings. The assumption that maximal smoothness leads to optimal generalisation no longer holds when an alternative metric is permitted to influence the mapping. In a classification scenario, for example, curvature will be necessarily high at class boundaries, and this will affect the choice of σ in addition to requiring greater consideration of the appropriate number of basis functions to use. Nevertheless, the implicit weight decay element may still be of key benefit, but further study is necessary in this direction.

NEUROSCALE for classification. The use of particular feature spaces derived by NEUROSCALE as the basis for a subsequent data classification in Chapter 4 proved very effective, although the experiments were necessarily concise. This is an area that offers much potential for further study, particularly in terms of the effect of parameter 'α', the number of features 'q' and the form of subjective dissimilarities used. Another important question is whether the self-regularising component within relatively supervised NEUROSCALE models is relevant to the classification context, and does this lead to better prediction performance?

Radial basis functions. One adjustable element of the NEUROSCALE architecture is the type of basis function that is utilised. Various types were used in experiments (Gaussian, thin-plate splines, $z \log z$ and cubic), with some being better than others for various datasets. However, no formal comparison/evaluation of basis function type has been undertaken. Additionally, the experiments in section 6.6.3 on the optimal width of Gaussian basis functions were generally consistent with predictions but revealed an unexplained mid-range 'bump' in the test error profile, which warrants further investigation. Furthermore, while parts of Chapter 6 described how generalisation did not deteriorate with large numbers of basis functions, there was no consideration of the smallest number of such functions that might be used (which would lead to computational improvements).

The scaling problem. The major drawback of NEUROSCALE is that the training time scales in the square of the number of patterns in the dataset, and thus places a relatively low maximum limit thereon

($\approx$ 1000). Chapter 7 considered some of the previous effort directed at alleviating this problem, but it was argued that most of the approaches were unprincipled and amounted to some form of heuristic subset selection. A very recent development has been the formulation of a topographic map, related to Kohonen's paradigm, based on the concept of a *generative* model (one that maps from the feature space to the data space) [Bishop, Svensén, and Williams 1996]. An adaptation of this model offers a potential for application to a Sammon-like mapping, and one that would permit a reduction in computation through the use of a constrained mixture-modelling approach. Indeed, an additional advantage would be the defining of a density approximation in the data space, a measure that is absent in Sammon's algorithm. Because of the importance of the scaling problem, this would perhaps be the most relevant, interesting and profitable direction for future research.