In: Proceedings of the Fourth Neural Computation and
Psychology Workshop: Connectionist Representations
Edited by John A. Bullinaria, David G. Glasspool
and George Houghton
London Springer-Verlag (1998)

Objective Functions for Topography: A Comparison of Optimal Maps

Geoffrey J. Goodhill
Georgetown Institute for Cognitive and Computational Sciences
Georgetown University Medical Center
Washington DC 20007, USA

Terrence J. Sejnowski
The Salk Institute for Biological Studies
La Jolla, CA 92037, USA

Topographic mappings are important in several contexts, including data visualization, connectionist representation, and cortical structure. Many different ways of quantifying the degree of topography of a mapping have been proposed. In order to investigate the consequences of the varying assumptions that these different approaches embody, we have optimized the mapping with respect to a number of different measures for a very simple problem - the mapping from a square to a line. The principal results are that (1) different objective functions can produce very different maps, (2) only a small number of these functions produce mappings which match common intuitions as to what a topographic mapping "should" actually look like for this problem, (3) the objective functions can be put into certain broad categories based on the overall form of the maps, and (4) certain categories of objective functions may be more appropriate for particular types of problem than other categories.

1 Introduction

Problems of mapping occur frequently in understanding biological processes, in designing connectionist representations, and in formulating abstract methods of data analysis. An important concept in all these domains is that of a "neighbourhood preserving" map, also sometimes referred to as a topographic, topological, topology-preserving, orderly, or systematic map. Intuitively speaking, such maps take points in one space to points in another space such that nearby points map to nearby points (and sometimes in addition far-away points map to far-away points). Such maps are useful in data analysis and data visualization, where a common goal is to represent data from a high-dimensional space in a low-dimensional space so as to preserve as far as possible the "internal structure" of the data in the high dimensional space (see e.g. [11]). In psychology, topographic mappings have been used to understand mental representations: for instance the idea that similar features of the world are represented close together in some internal semantic space [18]. In neurobiology there are many examples of neighbourhood-preserving mappings, for instance between the retina and more central structures [20]. Another type of neighbourhood-

preserving mapping in the brain is that, for instance, from the visual world to cells in the primary visual cortex which represent a small line segment at a particular position and orientation in the visual scene [8]. A possible goal of such biological maps is to represent nearby points in some sensory "feature space" by nearby points in the cortex [4]. This could be desirable since sensory inputs are often locally redundant: for instance in a visual scene pixel intensities are highly predictable from those of their neighbours. In order to perform "redundancy reduction" [1], it is necessary to make comparisons between the output of cells in the cortex that represent redundant inputs. Two ways this could be achieved are either by making a direct connection between these cells, or by constructing a suitable higher-order receptive field at the next level of processing. In both cases, the total length of wire required can be made short when nearby points in the feature space map to nearby points in the cortex (see [3, 4, 16, 14] for further discussion).

A number of different objective functions have been proposed to measure the degree of topography of a particular mapping (for reviews see [6, 7]). Given the wide variety of quantification choices available, it is important to understand what impact these choices have on the form of the maps that each measure best favors. This gives insight into which measures are most appropriate for particular types of applications. This paper addresses this question for a very simple problem: the mapping of 10×10 points in a square array to 1×100 points in a linear array (see figure 1). Our approach is to explicitly optimize several different objective functions from the topographic mapping literature for this case, and thus gain insight into the type of representation that each measure forms.

2 Objective functions

The objective functions investigated are as follows (for more details see [6]). Define the similarities in the input space (square) as $F(i,j)$, and in the output space (line) as $G(p,q)$ (figure 1), where i and j are points in the input space and p and q are points in the output space. Let there be N points in total, and M be a 1-1 mapping from points in the input space to points in the output space. For the first three of the measures considered, both F and G are taken to be euclidean distances in the two spaces, with distance between neighbouring points in each space taken as unity.

- **Metric Multidimensional Scaling** [19]: minimize

$$\sum_{i=1}^{N} \sum_{j<i} (F(i,j) - G(M(i), M(j)))^2$$

- **Sammon measure** [17]: minimize

$$\frac{1}{\sum_{i=1}^{N} \sum_{j<i} F(i,j)} \sum_{i=1}^{N} \sum_{j<i} \frac{(F(i,j) - G(M(i), M(j)))^2}{F(i,j)}$$

- **Spearman coefficient** [2]: maximize

$$\frac{\sum_i (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2} \sqrt{\sum_i (S_i - \overline{S})^2}}$$

  where R_i and S_i are the corresponding rankings in the ordered lists of the F's and and G's.

For the other four measures we consider, similarities are nonlinear functions of euclidean distance. They are all cases of the C measure [5, 7]:

$$C = \sum_{i=1}^{N} \sum_{j<i} F(i,j) G(M(i), M(j)),$$

for different choices of similarity function.

- **Minimal path length** [4]: $F(i,j)$ = euclidean distance, $G(p,q) = 1$ if p, q are neighbouring on the line and 0 otherwise.
- **Minimal wiring** [4]: $G(p,q)$ = euclidean distance, $F(i,j) = 1$ if i, j are neighbouring in the square and 0 otherwise.
- **Minimal distortion** [13]: $F(i,j)$ = squared euclidean distance, $G(p,q) = e^{-d^2/\sigma^2}$, where d = euclidean distance between p and q, and σ is the length scale in the output space over which nearby output points should represent similar input points. This is related to the minimal path length measure, but with a broader neighbourhood function in the output space.
- **Inverted minimal distortion** [15]: $G(p,q)$ = squared euclidean distance, $F(i,j) = e^{-d^2/\sigma^2}$, where d = euclidean distance between i and j, and σ is now the equivalent length scale in the input space. This is related to the minimal wiring measure, but with a broader neighbourhood function in the input space.

3 Minimization procedure

There are of the order of 100! possible mappings for this problem, and thus optimization by exhaustive search is clearly impractical. Instead we used simulated annealing, a heuristic optimization method [9]. This performs gradient descent (or ascent, as appropriate) in the objective function, but allows occasional steps in the wrong direction so that the solution is less likely to get stuck in a local optimum. The probability of taking a step in the wrong direction is controlled by a "temperature" parameter that is gradually reduced. The parameters used were as follows [12]. The initial map between points in the square and points on the line was random. At each step, a candidate move consisted of interchanging a random pair of points in the map. This move was accepted with 100%

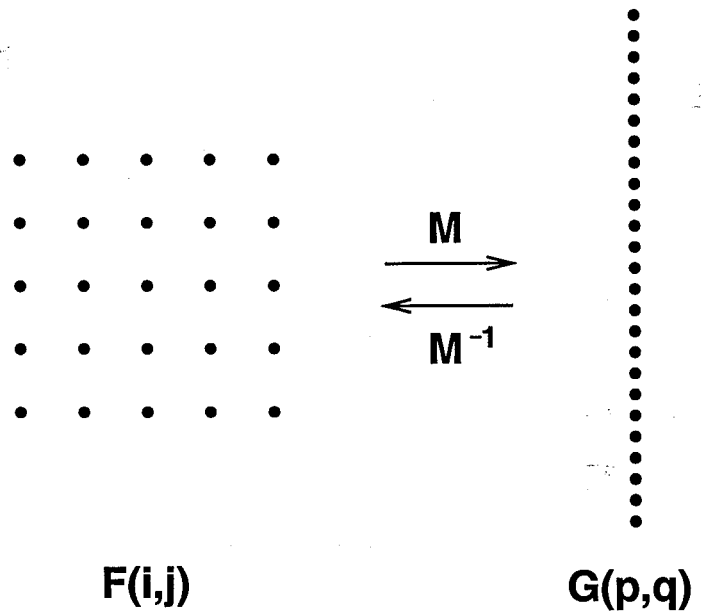

Figure 1: The example mapping problem (only 25 points are shown). The matrix $F(i,j)$ defines similarities in the input space (square), the matrix $G(p,q)$ defines similarities in the output space (line), and M is the 1-1 map between the two spaces.

probability if it improved the value of the objective function, or with a probability determined by the temperature if it did not. Once the sooner of 10,000 candidate moves had been generated or 1000 moves accepted, the temperature was multiplied by 0.998. The procedure was terminated when no moves were accepted out of 10,000 candidates at the same temperature. Empirically, these values were found to produce close to optimal solutions for cases where the optimal solution is explicitly known (see figure 2).

4 Results

Figure 3 shows the maps found for the metric MDS, Sammon and Spearman measures. The illusion of multiple ends to the line is due to the map frequently doubling back on itself. For instance, consider the fifth column of the square for the optimal Sammon map (figure 3(b)). Initially the line meets this column at the point (5,1), counting from the bottom left corner of the square. However, the next point in the map is actually (5,10), followed by (5, 6), (5,-8), (5, 7), (5,4), (5,9), (5,5), (5,3), and (5,2), where the line then proceeds on to the sixth column. This strong local discontinuity is the result of the more global optimization concerns that dominate these measures.

Figure 4 shows minimal distortion solutions for varying σ. For small σ,

(a) (b)

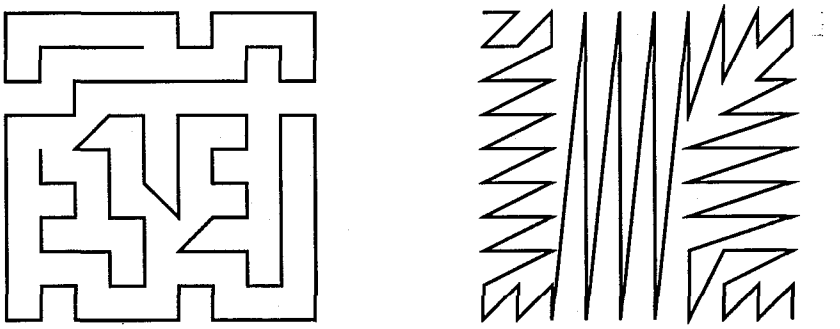

Figure 2: Testing the minimization algorithm for cases where the optima are explicitly known. (a) Mininal path length solution, length = 100.243, 1.3% longer than the optimal of 99.0. (b) Minimal wiring solution, length = 917.0, 0.3% longer than the optimal of 914.0 [4]. An optimal minimal path length solution was found when the cooling rate was increased to 0.9999 and the upper bound increased to 100,000; however it was computationally impractical to run all the simulations this slowly.

the solution resembles the minimal path optimum of figure 2(a), since the contribution from more distant neighbours than nearest neighbours is negligible. However, as σ increases the map changes form. Local continuity becomes less important compared to continuity at the scale of σ, the map becomes more spiky, and the number of large-scale folds in the map gradually decreases until at $\sigma = 20$ there is just one. This last map also shows some of the frequent doubling back behaviour seen in figure 3.

Figure 5 shows analogous results for reversed minimal distortion. For small σ the map somewhat resembles the minimal wiring map of figure 2(b), as expected. However, as σ increases, the map rapidly takes on a form reminiscent of figure 3.

In terms of general appearance, the optimal maps we have calculated can be placed into four classes.

1: Metric MDS, Sammon, reversed minimal distortion for $\sigma = 4.0$ (figs 3(a), 3(b), 5(d)). These maps are very locally discontinuous but have a characteristic overall form. This is because they all take into account neighbourhood preservation *at all scales.* Thus local continuity is not privileged over global continuity, and global concerns dominate.

2: Minimal distortion for $\sigma \leq 4.0$ (fig 4(a-c)). Only local neighbourhoods on the line are of interest. For $\sigma \sim 1$ this means in effect only nearest neighbours, and so the line meanders randomly through the square. As σ increases, the line

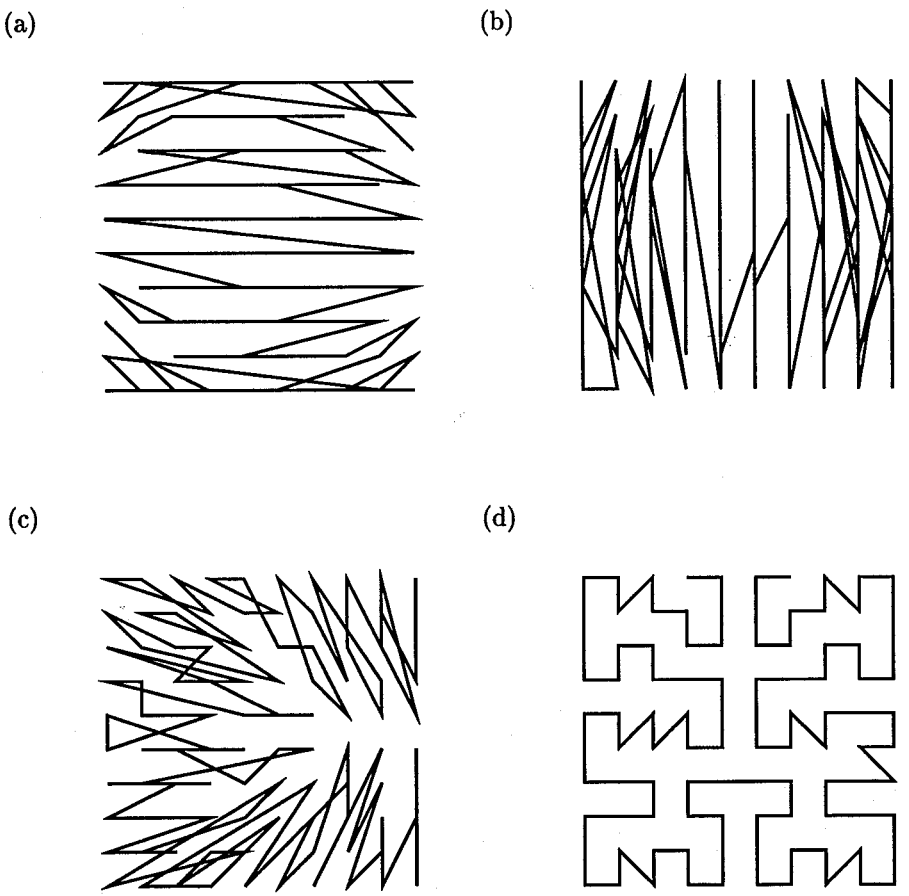

Figure 3: Solutions found by simulated annealing for the square to line problem. (a) Metric MDS measure, cost = 9570087.8 (b) Sammon measure, cost = 38.5. (c) Spearman measure, cost = 0.698. For these measures, global topography dominates local topography. (d) For comparison, a map found by the elastic net algorithm [4]. This is less optimal than any of the maps shown in this paper with respect to the objective functions for which they were optimized.

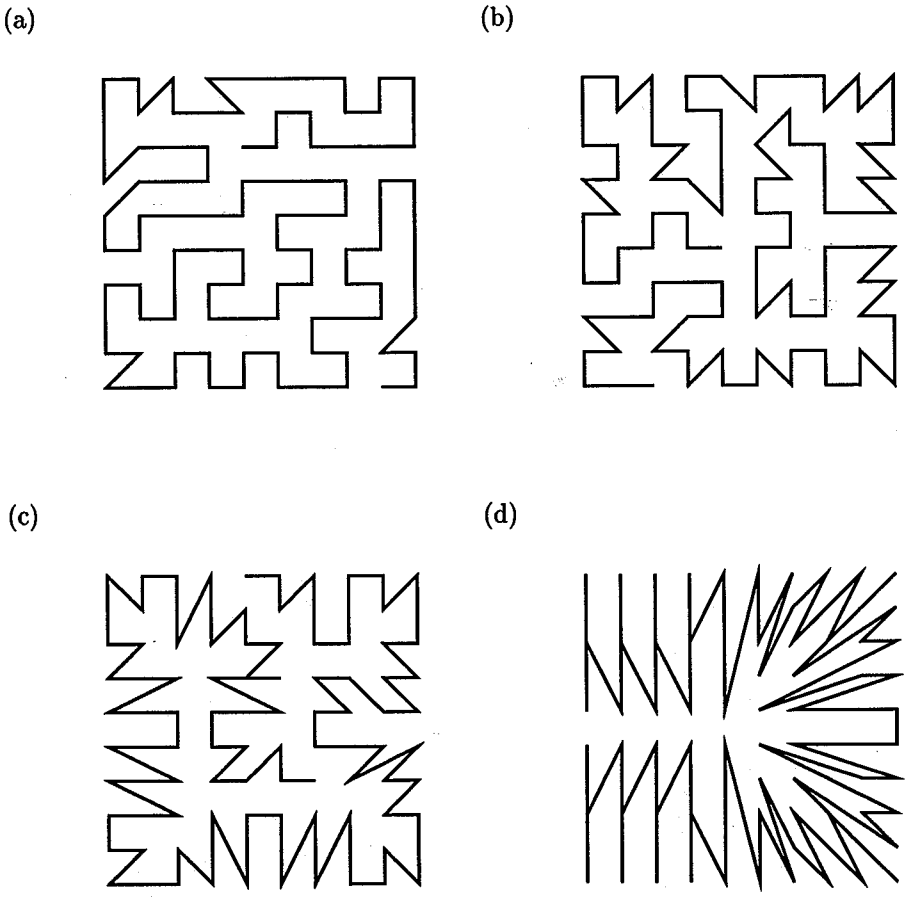

Figure 4: Minimal distortion solutions found by simulated annealing for the square to line problem. (a) $\sigma = 1.0$, cost $= 43.3$. (b) $\sigma = 2.0$, cost $= 214.7$. (c) $\sigma = 4.0$, cost $= 833.2$. (d) $\sigma = 20.0$, cost $= 18467.1$. Note how the scale of the folding of the map changes with σ.

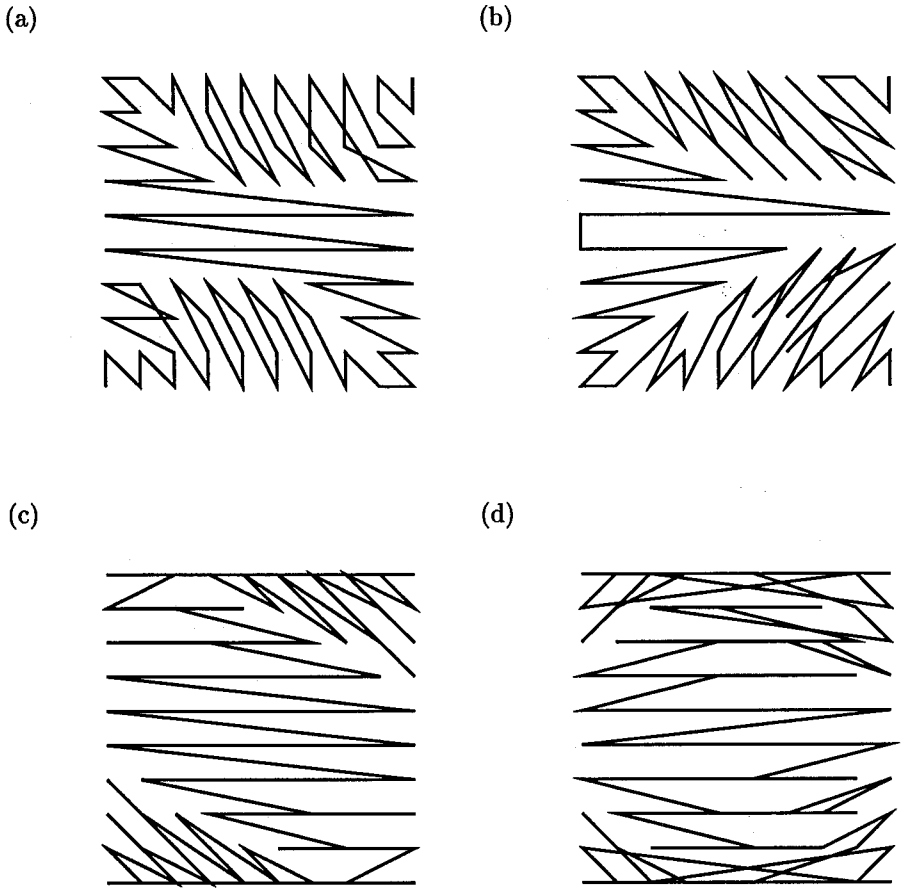

Figure 5: Reversed minimal distortion solutions found by simulated annealing for the square to line problem. (a) $\sigma = 1.0$, cost = 6250.2. (b) $\sigma = 2.0$, cost = 86469.1. (c) $\sigma = 3.0$, cost = 349926.5. (d) $\sigma = 4.0$, cost = 851763.4.

is encouraged to fold to try to keep more distant neighbours close: the scale of the folding depends on σ.

3: Spearman, minimal distortion for $\sigma = 20$ (figs 3(c), 4(d)). Although these share with class 1 the property of having very global concerns, they both have a characteristic horseshoe shape.

4: Reversed minimal distortion with $\sigma \leq 3.0$ (fig 5(a-c)). These have long stretches in the middle, with rapid zig-zags at the edge.

5 Discussion

Of the measures we have considered, only minimal distortion produces intuitively appealing maps for this problem. An interesting point is that the minimal distortion measure is almost an objective function for the SOFM algorithm [10], with σ determining the size of the neighbourhood function [13]. In the SOFM algorithm however σ decreases with time, making it hard to draw direct analogies. It could be that the intuitive appeal of the maps produced by minimal distortion is precisely because of wide familiarity with the behaviour of the SOFM, rather than for any reason more firmly rooted in the mathematics of neighbourhood preservation.

What do these results tell us about which measures are appropriate for different problems? If it is desired that generally nearby points should always map to generally nearby points as much as possible in both directions, and one is not concerned about very local continuity, then measures in class 1 are useful. This may be appropriate for some data visualization applications where the overall structure of the map is more important than its fine detail. If, on the other hand, one wants a smooth progression through the output space to imply a smooth progression through the input space, one should choose from class 2. This may be important for data visualization where it is believed the data actually lies on a lower-dimensional manifold in the high-dimensional space. However, an important weakness for this representation is that some neighbourhood relationships between points in the input space may be completely lost in the resulting representation. For understanding the structure of cortical mappings, self-organizing algorithms that optimize objectives in class 2 have proved useful [4]. Very few other objectives have been applied to this problem though, so it is still an open question which are most appropriate. Classes 3 and 4 represent pathologies that have been hitherto unappreciated. There may be some applications for which they are worthwhile, but for brain maps they are unsuitable.

6 Conclusions

This paper has attempted to impose some order on the space of popular measures of neighbourhood preservation, in order to better understand topographic mapping methods in data analysis, connectionism and neurobiology. We considered a mapping problem that represents an extremely simple example of a

mismatch between the dimensions of the input space and the output space. By examining the maps given by optimizing each measure, we tried to group together different types of optimal maps and thus the measures that generated them. The main conclusions are as follows.

1. The optimal maps span a surprisingly broad subspace of possible maps, and include maps lacking local continuity.

2. This subspace is much larger than the space of maps that are often referred to as topographic. This suggests that great caution should be used in relying on visual inspection to judge degrees of topography.

3. The subspace of optimal maps, and thus the measures that generated them, can be divided into four main classes based on the general form of the maps produced.

4. The structure of this subspace can provide guidance in choosing the most appropriate mapping measure to apply to more complex mapping problems. For instance, finding a highly curved manifold in a high dimensional space requires preservation of local but not global topography, whereas forming a low dimensional representation of the relationships between clusters in a high dimensional space (ignoring structure within a cluster) requires preservation of global but not local topography. In general, the sensible use of topographic mapping techniques requires a good understanding of the nature of the particular application.

References

[1] Barlow, H.B. (1989). Unsupervised learning. *Neural Computation*, **1**, 295-311.

[2] Bezdek, J.C. & Pal, N.R. (1995). An index of topological preservation for feature extraction. *Pattern Recognition*, **28**, 381-391.

[3] Cowey, A. (1979). Cortical maps and visual perception. *Qua. Jou. Exper. Psychol.*, **31**, 1-17.

[4] Durbin, R. & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, **343**, 644-647.

[5] Goodhill, G. J., Finch, S. & Sejnowski, T. J. (1995). Quantifying neighbourhood preservation in topographic mappings. Institute for Neural Computation Technical Report Series, No. INC-9505, November 1995. Available from http://www.giccs.georgetown.edu/~geoff

[6] Goodhill, G.J. & Sejnowski, T.J. (1996) Quantifying neighbourhood preservation in topographic mappings. In: "Proceedings of the 3rd Joint Symposium on Neural Computation", University of California, San Diego and California Institute of Technology, Vol. 6, Pasadena, CA: California Institute of Technology, 61-82. Available from http://www.giccs.georgetown.edu/~geoff

[7] Goodhill, G.J. & Sejnowski, T.J. (1997). A unifying objective function for topographic mappings. *Neural Computation*, **9**, 1291-1304.

[8] Hubel, D.H. & Wiesel, T.N. (1977). Functional architecture of the macaque monkey visual cortex. *Proc. R. Soc. Lond. B*, **198**, 1-59.

[9] Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671-680.

[10] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43**, 59-69.

[11] Krzanowski, W.J. (1988). Principles of multivariate analysis: a user's perspective. Oxford statistical science series; v. 3. Oxford University Press.

[12] van Laarhoven, P.J.M. & Aarts, E.H.L. (1987). Simulated annealing: theory and applications. Reidel, Dordrecht, Holland.

[13] Luttrell, S.P. (1990). Derivation of a class of training algorithms. *IEEE Trans. Neural Networks*, **1**, 229-232.

[14] Mitchison, G. (1991). Neuronal branching patterns and the economy of cortical wiring. *Proc. Roy. Soc. B.*, **245**, 151-158.

[15] Mitchison, G. (1995). A type of duality between self-organizing maps and minimal wiring. *Neural Computation.*, **7**, 25-35.

[16] Nelson, M.E. & Bower, J.M. (1990). Brain maps and parallel computers. *Trends Neurosci.*, **13**, 403-408.

[17] Sammon, J.W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, **18**, 401-409.

[18] Shepard, R.N. (1980). Multidimensional scaling, tree-fitting and clustering. *Science*, **210**, 390-398.

[19] Torgerson, W.S. (1952). Multidimensional Scaling, I: theory and method. *Psychometrika*, **17**, 401-419.

[20] Udin, S.B. & Fawcett, J.W. (1988). Formation of topographic maps. *Ann. Rev. Neurosci.*, **11**, 289-327.