Proceedings of the 3rd Joint
Symposium on Neural
Computation, La Jolla, CA
Vol. 6, 61-82 (1996)

Quantifying neighbourhood preservation in topographic mappings

Geoffrey J. Goodhill[1] & Terrence J. Sejnowski[2]

1. The Salk Institute for Biological Studies
10010 North Torrey Pines Road
La Jolla, CA 92037, USA

2. The Howard Hughes Medical Institute
The Salk Institute for Biological Studies
10010 North Torrey Pines Road
La Jolla, CA 92037, USA
&
Department of Biology
University of California San Diego
La Jolla, CA 92037, USA

Abstract

Mappings that preserve neighbourhood relationships are important in many contexts, from neurobiology to multivariate data analysis. It is important to be clear about precisely what is meant by preserving neighbourhoods. At least three issues have to be addressed: how neighbourhoods are defined, how a perfectly neighbourhood preserving mapping is defined, and how an objective function for measuring discrepancies from perfect neighbourhood preservation is defined. We review several standard methods, and using a simple example mapping problem show that the different assumptions of each lead to non-trivially different answers. We also introduce a particular measure for topographic distortion, which has the form of a quadratic assignment problem. Many previous methods are closely related to this measure, which thus serves to unify disparate approaches.

1 Introduction

Problems of mapping occur frequently both in understanding biological processes and in formulating abstract methods of data analysis. An important concept in both 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. [Krzanowski 1988]). Just a few of the algorithms that have found application in this context are principal components analysis (PCA), multidimensional scaling [Torgerson 1952, Shepard 1962a, Shepard 1962b, Kruskal 1964a, Kruskal 1964b], Sammon mappings [Sammon 1969], and neural network algorithms such as the self-organizing map (SOM) [Kohonen 1982, Kohonen 1988] and the elastic net [Durbin & Willshaw 1987, Durbin & Mitchison 1990]. One hope is that by preserving neighbourhoods in the mapping it will be possible to see more clearly structure in the high-dimensional data, such as clusters, or that this type of dimension-reduction will reveal that the data occupies a lower-dimensional subspace than was originally apparent.

In neurobiology there are many examples of neighbourhood-preserving mappings, for instance between the retina and more central structures [Udin & Fawcett 1988]. 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 [Hubel & Wiesel 1977]. A possible goal of such biological maps is to represent nearby points in some sensory "feature space" by nearby points in the cortex [Durbin & Mitchison 1990]. 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" (e.g. [Barlow 1989]), 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 [Cowey 1979, Durbin & Mitchison 1990, Nelson & Bower 1990, Mitchison 1991, Mitchison 1992] for further discussion).

So far we have discussed neighbourhood preservation in intuitive terms. However, it is vital to ask what this intuitive idea might mean more precisely, i.e. exactly what computational principles such mappings are addressing. Without a clear set of principles it is impossible to decide whether a particular mapping has achieved "neighbourhood preservation", whether one mapping algorithm has performed better than another on a particular problem, or what computational goals mappings in the brain might be pursuing. A large number of choices have to be made to reach a precise mathematical measure of neighbourhood preservation, as we discuss below. Different combinations of choices will in general give different answers for the same mapping problem, and the combination of choices that is most appropriate will vary from problem to problem. Several measures of neighbourhood preservation have recently been proposed which implement particular sets of choices. In many cases there are few *a priori* grounds for choosing between different formulations: rather each may be useful for different types of problem. From a biological perspective, an interesting question is to investigate which combinations of choices yield mappings closest to those seen experimentally in various contexts, and how such choices could be implemented in the brain. From a practical perspective, it is desirable to understand more about the choices available and the degree to which they are appropriate for different types of problems.

We adopt the distinction made in [Marr 1982] between the "computational" and "algorithmic" levels of analysis. The former concerns computational goals, while the latter concerns how these computational goals are achieved. These two levels can sometimes be difficult, or inappropriate, to disentangle when addressing biological problems [Sejnowski et al 1988]. However, for discussing topographic mappings from an abstract perspective, it is important to be clear about this distinction. As an example, minimal wiring and minimal path length (discussed later) are clear computational level principles. However, the SOM [Kohonen 1982] exists only at the algorithmic level since it is not following the gradient of an objective function [Erwin et al 1992]. In particular, given a map, it does not provide a number measuring its quality.

This paper reviews a number of different measures of neighbourhood preservation. We show that several of them can, under certain circumstances, be economically described as particular special cases of a more general objective function. In order to explore the consequences of the different assumptions embodied in each measure, we examine in detail their application to a very simple mapping problem. This is the mapping from a square to a line, or more particularly a regular array of 10×10 points to a regular array of 1×100 points. This is one of the simplest examples of a dimension mismatch between two spaces. The investigation proceeds in three stages. Firstly we consider four exemplar maps, and calculate how each measure ranks these in terms of achieving neighbourhood preservation. Secondly, the sensitivity of some of these rankings to the measurement of similarity in each space is investigated. Thirdly, simulated annealing is used to calculate the (close to) optimal square-to-line map for each measure. We discuss the implications of these results for the appropriate choice of measure for a particular problem.

In order to make a direct comparison of measures purely in terms of their topographic properties,

only a restricted class of mapping problems is discussed. To eliminate issues of clustering or vector quantization, we assume that there are the same number of points in each space, and that the mapping is bijective (i.e. one-to-one). In addition, we assume that in each space there exists a fixed "similarity structure", that specifies for every pair of points its degree of similarity. In a simple case, this similarity is just euclidean distance between points in a geometric space. However, the similarity structure need not have a geometric interpretation. For example, for purely "nearest neighbour" structure, similarities are binary. These restrictions give a simple enough framework so that several different approaches to neighbourhood preservation can be compared.

2 Defining perfection and measures of discrepancy

There are several choices for defining a mapping that "perfectly" preserves the neighbourhood structure of data in one space in another space. The strongest is to say that the mapping must *preserve similarities*; that is, for each pair of points in one space, its similarity should be equal to the similarity of the images of those points in the other space. A slighter weaker one is that the similarity values between pairs of points should be *perfectly correlated*. Weaker still is that the mapping should only *preserve similarity orderings*; that is, rather than comparing the absolute values of the similarity between pairs of points in one space and the similarity between their images in the other, one is concerned only that the relative ordering of similarities within the two sets is the same. If similarity values in the output space are plotted against similarity values in the input space for a particular map (as in for example figure 3), the first criterion specifies that all points should lie on a straight line at 45 degrees to the x axis, the second that points lie on a straight line of arbitrary angle, and the third that the points lie on a line that is not necessarily straight, but is monotonically increasing. It is important to be clear about which of these three goals (or perhaps some fourth goal) any particular mapping principle implies. Different goals will be appropriate for different applications.

However, in most practical applications none of these varieties of perfect map will be achievable, and a *measure of discrepancy* is required which assesses the degree to which perfection has been achieved. Given a definition of perfection there are many ways to measure discrepancy, each of which will compare different non-perfect maps differently. Before surveying some of the measures that have been proposed, we first introduce the C measure. This is simply the correlation coefficient between similarities in the two spaces. Several previously proposed measures turn out to be equivalent (under the restrictions described above) to C, for particular choices of similarity function.

2.1 The C measure

Consider an input space V_{in} and an output space V_{out}, each of which contains N points (see figure 1). Let M be the mapping from points in V_{in} to points in V_{out}. We use the word "space" in a general sense: either or both of V_{in} and V_{out} may not have a geometric interpretation. Assume that for each space there is a symmetric "similarity" function which, for any given pair of points in the space, specifies by a non-negative scalar value how similar (or dissimilar) they are. Call these functions F for V_{in} and G for V_{out}. Then we define a cost functional C as follows [Goodhill et al 1996][1]

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

where i and j label points in V_{in}, and M(i) and M(j) are their respective images in V_{out}. The sum is over all possible pairs of points in V_{in}. Since we have assumed that M is a bijection it is therefore

[1]Readers may notice that under particular assumptions C can be interpreted as a discrete form of the continuous mapping functional introduced in [Luttrell 1990, Mitchison 1995]. This connection is discussed in more detail in section 3.1.5.

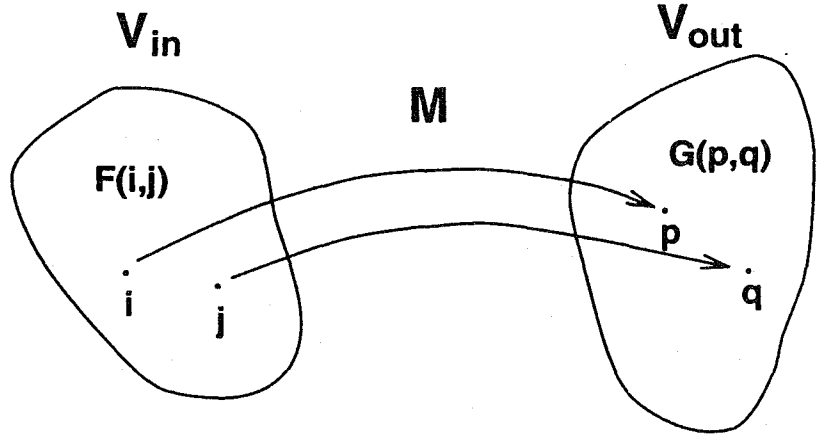

Figure 1: The mapping framework.

invertible, and C can equivalently be written

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

where now i and j label points in V_{out}, and M^{-1} is the inverse map. A good mapping is one with a high value of C. However, if one of F or G is given as a dissimilarity function (i.e. increasing with decreasing similarity) then a good mapping has a low value of C. How F and G are defined is problem-specific. They could be euclidean distances in a geometric space, some (possibly non-monotonic) function of those distances, or they could just be given, in which case it may not be possible to interpret the points as lying in some geometric space. Clearly C measures the correlation between the F's and the G's, and thus falls into the second of the three categories described above for defining mapping perfection. It is also straightforward to show that if a mapping that preserves orderings exists, then maximizing C will find it. This is equivalent to saying that for two vectors of real numbers, their inner product is maximized over all permutations within the two vectors if the elements of the vectors are identically ordered, a proof of which can be found in [Hardy et al 1934, page 261].

2.2 Relation of C to quadratic assignment problems

Formulating neighbourhood preservation in terms of the C measure sets it within the well-studied class of quadratic assignment problems (QAPs). These occur in many different practical contexts, and take the form of finding the minimal or maximal value of an equation similar to C (see [Burkard 1984] for a general review, and [Lawler 1963, Finke et al 1987] for more technical discussions). An illustrative example is the optimal design of typewriter keyboards [Burkard 1984]. If $F(i,j)$ is the average time it takes a typist to sequentially press locations i and j on the keyboard, while $G(p,q)$ is the average frequency with which letters p and q appear sequentially in text of a given language (note that in this example F and G are not necessarily symmetrical), then the keyboard that minimizes average typing time will be the one that minimizes the product

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

(cf equation 1), where $M(i)$ is the letter that maps to location i. The substantial theory developed for QAPs is directly applicable to the C measure. As a concrete example, QAP theory provides several different ways of calculating bounds on the minimum and maximum values of C for each problem. This could be very useful for the problem of assessing the quality of a map relative to the unknown best

possible (rather than simply making a comparison between two maps as we do later). One particular case is the eigenvalue bound [Finke et al 1987]. If the eigenvalues of symmetric matrices F and G are λ_i and μ_i respectively, such that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ and $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$, then it can be shown that $\sum_i \lambda_i \mu_i$ gives a lower bound on the value of C.

QAPs are in general known to be of complexity NP-hard. A large number of algorithms for both exact and heuristic solution have been studied (see e.g. [Burkard 1984], the references in [Finke et al 1987], and [Simić 1991]). However, particular instantiations of F and G may make possible very efficient algorithms for finding good local optima, or alternatively may beset C with many bad local optima. Such considerations provide an additional practical constraint on what choices of F and G are most appropriate.

3 A survey of measures

3.1 Measures equivalent to C for particular choices of similarity functions

3.1.1 Metric Multidimensional Scaling

Metric multidimensional scaling (metric MDS) is a technique originally developed in the context of psychology for representing a set of N "entities" (e.g. subjects in an experiment) by N points in a low- (usually two-) dimensional space. For these entities one has a matrix which gives the numerical dissimilarity between each pair of entities. The aim of metric MDS is to position points representing entities in the low-dimensional space so that the set of distances between each pair of points matches as closely as possible the given set of dissimilarities. The particular objective function optimized is the summed squared deviations of distances from dissimilarities. The original method was presented in [Torgerson 1952]; for reviews see [Shepard 1980, Young 1987].

In terms of the framework presented earlier, the MDS dissimilarity matrix is F. Note that there may not be a geometric space of any dimensionality for which these dissimilarities can be represented by distances (for instance if the dissimilarities do not satisfy the triangle inequality), in which case V_{in} does not have a geometric interpretation. V_{out} is the low-dimensional, continuous, space in which the points representing entities are positioned, and G is euclidean distance in V_{out}. Metric MDS selects the mapping M, by adjusting the positions of points in V_{out}, which minimizes

$$\sum_{i=1}^{N} \sum_{j<i} (F(i,j) - G(M(i), M(j)))^2 \tag{3}$$

If a value of zero can be achieved, the resulting map clearly satisfies all three definitions of perfection above.

Under our assumptions, this objective function is identical to the C measure. Expanding out the square in 3 gives

$$\sum_{i=1}^{N} \sum_{j<i} (F(i,j)^2 + G(M(i), M(j))^2 - 2F(i,j)G(M(i), M(j))). \tag{4}$$

The last term is twice the C measure. The entries in F are fixed, so the first term is independent of the mapping. In metric MDS the sum over the G's varies as the entities are moved in the output space. If one instead considers the case where the positions of the points in V_{out} are fixed, and the problem is to find the assignment of entities to positions that minimizes equation 3, then the sum over the G's is also independent of the map. In this case the metric MDS measure becomes exactly equivalent to C. One effect of varying the G^2 term in metric MDS is to match the scale of the representation to that of the entries in the F matrix.

3.1.2 Minimal wiring

In minimal wiring [Mitchison & Durbin 1986, Durbin & Mitchison 1990], a good mapping is defined to be one that maps points that are nearest neighbours in V_{in} as close as possible in V_{out}, where closeness in V_{out} is measured by for instance euclidean distance raised to some power. The motivation here is the idea that it is often useful in processing e.g. sensory data to perform computations that are local in some space of input features V_{in}. To do this in V_{out} (e.g. the cortex) the images of neighbouring points in V_{in} need to be connected; the similarity function in V_{out} is intended to capture the cost of the wire (e.g. axons) required to do this. Minimal wiring is equivalent to the C measure for

$$F(i,j) = \begin{cases} 1 & : \ i,j \text{ neighbouring} \\ 0 & : \ \text{otherwise} \end{cases}$$

$$G(M(i), M(j)) = ||M(i) - M(j)||^p$$

For the cases of 1-D or 2-D square arrays investigated in [Mitchison & Durbin 1986, Durbin & Mitchison 1990], neighbours are taken to be just the 2 or 4 adjacent points in the array respectively.

3.1.3 Minimal path length

In this scheme, a good map is one such that, in moving between nearest neighbours in V_{out}, one moves the least possible distance in V_{in}. This is for instance the mapping required to solve the Traveling Salesman Problem (TSP) where V_{in} is the distribution of cities and V_{out} is the one-dimensional tour. This goal is implemented by the elastic net algorithm [Durbin & Willshaw 1987, Durbin & Mitchison 1990, Goodhill & Willshaw 1990], which measures similarity in V_{in} by squared distances:

$$F(i,j) = ||v_i - v_j||^2$$

$$G(p,q) = \begin{cases} 1 & : \ p,q \text{ neighbouring} \\ 0 & : \ \text{otherwise} \end{cases}$$

where v_k is the position of point k in V_{in} (we have only considered here the regularization term in the elastic net energy function, which also includes a term matching input points to output points). Thus minimal wiring and minimal path length are symmetrical cases under equation 1. Their relationship is discussed further in [Durbin & Mitchison 1990], where the abilities of minimal wiring and minimal path length are compared with regard to reproducing the structure of the map of orientation selectivity in primary visual cortex (see also [Mitchison 1995]).

3.1.4 The approach of Jones et al

[Jones et al 1991] investigated the effect of the shape of the cortex (V_{out}) relative to the lateral geniculate nuclei (V_{in}) on the overall pattern of ocular dominance columns in the cat and monkey, using an optimization approach. They desired to keep both neighbouring cells in each LGN (as defined by a hexagonal array), and anatomically corresponding cells between the two LGNs, nearby in the cortex (also a hexagonal array). Their formulation of this problem can be expressed as a maximization of C when

$$F(i,j) = \begin{cases} 1 & : \ i,j \text{ neighbouring, corresponding} \\ 0 & : \ \text{otherwise} \end{cases}$$

and

$$G(p,q) = \begin{cases} 1 & : \ p,q \text{ first or second nearest neighbours} \\ 0 & : \ \text{otherwise} \end{cases}$$

For 2-D V_{in} and V_{out} they found a solution such that if $F(i,j) = 1$ then $G(M(i), M(j)) = 1$, $\forall i,j$. Alternatively this problem could be expressed as a minimization of C when $G(p,q)$ is the stepping distance (see below) between positions in the V_{out} array. They found this gave appropriate behaviour for the problem addressed.

3.1.5 Minimal distortion

Luttrell and Mitchison have introduced mapping functionals for the continuous case. Under appropriate assumptions to reduce them to the discrete case, these are equivalent to C. Expression in this restricted form helps to connect these functionals with the other measures we have discussed.

[Luttrell 1990, Luttrell 1994] defined a "minimal distortion principle" that can be interpreted as a measure of mapping quality. He defined "distortion" D as

$$D = \int dx\, P(x)\, d\{x, x'[y(x)]\}$$

x and y are vectors in the input and output spaces respectively. y is the map in the input to output direction, x' is the (in general different) map back again, and $P(x)$ is the probability of occurrence of x. x' and y are suitably adjusted to minimize D. An augmented version of D can be written which includes additive noise in the output space:

$$D = \int dx\, P(x) \int dn\, \pi(n)\, d\{x, x'[y(x) + n]\}$$

where $\pi(n)$ is the probability density of the noise vector n. Intuitively, the aim is now to find the forward and backward maps so that the reconstructed value of the input vector is as close as possible to its original value after being corrupted by noise in the output space. In e.g. [Luttrell 1990], the d function was taken to be $\{x - x'[y(x) + n]\}^2$. In this case, [Luttrell 1990] showed that the MD measure can be differentiated to produce a learning rule that is almost the SOM rule.

For the discrete version of this, it is necessary to assume that appropriate quantization has occurred so that $y(x)$ defines a 1-1 map, and $x'(y)$ defines the same map in the opposite direction. In this case the minimal distortion measure becomes equivalent to the C measure, with $F(i, j)$ the squared euclidean distance between the positions of vectors i and j (assuming these to lie in a geometric space) and $G(p, q)$ the noise process in the output space.

The minimal distortion principle was generalized in [Mitchison 1995] by allowing the noise process and reconstruction error to take arbitrary forms. For instance they can be reversed, so that F is a gaussian and G is euclidean distance. In this case the measure can be interpreted as a version of the minimal wiring principle, establishing a connection between minimal distortion (and hence the SOM algorithm) and minimal wiring. This identification also yields a self-organizing algorithm similar to the SOM for solving minimal wiring problems, the properties of which are briefly explored in [Mitchison 1995].

[Luttrell 1994] generalized minimal distortion to allow a probabilistic match between points in the two spaces, which in the discrete case can be expressed as

$$D = \sum_i \sum_j F(i, j) \sum_k P(k|i) P(j|k)$$

where D is the distortion to be minimized, F is euclidean distance in the input space, $P(k|i)$ is the probability that output state k occurs given that input state i occurs, and $P(j|k)$ is the corresponding Bayes' inverse probability that input state j occurs given that output state k occurs. This reduces to C in the special case that $P(k|i) = P(i|k)$, which is true for bijections. A great deal of theory is developed in [Luttrell 1994] pointing out the usefulness of this revised functional for converting problems with hard constraints to those with soft constraints.

3.2 Other metric measures

These measures try to match similarities, but are not equivalent to C.

3.2.1 Sammon mappings

The Sammon approach [Sammon 1969] tries to match similarities exactly, by moving points around in V_{out}, but uses normalization terms which make the mapping nonlinear. The objective function is

$$\frac{1}{\sum_{i=1}^{N}\sum_{j<i} F(i,j)} \sum_{i=1}^{N}\sum_{j<i} \frac{(F(i,j) - G(M(i),M(j)))^2}{F(i,j)}. \tag{5}$$

This is similar to metric MDS, except that now the discrepancies for small F's are weighted more than for large F's. If the scale of the discrepancies between the F's and the G's is roughly the same as the size of the F's, this serves to even up the contributions to the objective function from different size F's and G's. It has recently been argued that this approach produces better maps than for instance the SOM [Bezdek & Pal 1995, Mao & Jain 1995, Lowe & Tipping 1996]. Note that the Sammon formula is not symmetric to interchange of F and G.

3.2.2 Demartines and Hérault

A measure related is the Sammon approach is that of [Demartines & Hérault 1995, Demartines & Hérault 1996]:

$$\sum_{i=1}^{N}\sum_{j<i} (F(i,j) - G(M(i),M(j)))^2 \; h(G(M(i),M(j))). \tag{6}$$

h is some decreasing function, therefore if M(i) and M(j) are very dissimilar, the measure is not concerned about whether i and j are similar or not. Demartines and Hérault argue that this measure has a more efficient minimization algorithm than the Sammon measure, and produces better representations of "strongly folded structures".

3.3 Nonmetric measures

These measures aim to match only similarity orderings.

3.3.1 Nonmetric Multidimensional Scaling

Nonmetric MDS (NMDS) aims to position points as in metric MDS, except that now it is attempted to match only the ordering of similarities between the input and output spaces, rather than the absolute values [Shepard 1962a, Shepard 1962b]. The mathematical measure used to quantify deviations from perfect ordering is somewhat ad hoc, and in fact often varies between different packaged software routines that implement NMDS. The first measure proposed was called STRESS [Kruskal 1964a, Kruskal 1964b], and is given in our notation by

$$\text{STRESS} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j<i}(G(i,j) - D(i,j))^2}{\sum_{i=1}^{N}\sum_{j<i} G(i,j)^2}} \tag{7}$$

where the D(i,j)'s are "disparities". These are target values for each G(i,j) such that if the G's achieved these values, then ordering would be preserved and STRESS would be zero. An algorithm for calculating the disparities is given in [Kruskal 1964b]. Another commonly used version of NMDS is ALSCAL [Takane et al 1977], which uses an objective function called SSTRESS ("squared stress"):

$$\text{SSTRESS} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j<i}(G(i,j)^2 - D(i,j)^2)^2}{\sum_{i=1}^{N}\sum_{j<i} D(i,j)^4}} \tag{8}$$

Note the normalization by D's, rather than G's as for STRESS. A different algorithm is used in ALSCAL for calculating disparities than the Kruskal method. Note that STRESS and SSTRESS do not purely measure ordinal discrepancies, since they are expressed in terms of absolute similarity values rather than orderings.

3.3.2 The approach of Villmann et al

In [Villmann et al 1994] the primary concern is with many-to-one mappings of data in a geometric, continuous input space V_{in} to a square array of points. We consider the situation after some process of, for instance, vector quantization has occurred, so that there are now the same number of points in both spaces, the positions of points in the input space are fixed, and we can talk about the degree of neighbourhood preservation of the bijective mapping between these two sets of points. [Villmann et al 1994] give a way of defining neighbourhoods, in terms of "masked Voronoi polyhedra" (see also [Martinetz & Schulten 1994]). This defines a neighbourhood structure where for any two points (two centers of masked Voronoi polyhedra), there is an integer dissimilarity which defines the "stepping distance" between them (cf [Kendall 1971]). They define a series of measures $\Phi(k)$ which give the number of times points which are neighbours in one space are mapped stepping distance k apart in the other space (they consider all indices for both directions of the map). If all the $\Phi(k)$ are zero they call the mapping "perfectly topology preserving".

The distribution of non-zero $\Phi(k)$ gives useful information about the scale of "discontinuities" in the map. However, no rule is specified in [Villmann et al 1994] for combining the $\Phi(k)$ into a single number that specifies the overall quality of a particular mapping. A simple way to do this would be to take a sum of the $\Phi(k)$ weighted by some function of k. If this function were increasing with k (and good mappings were defined to be the minima of the product), this would express a desire to minimize large scale discontinuities at the expense of small scale ones. Whether this or some other function is most appropriate depends on the problem.

3.3.3 The Topographic Product

The "topographic product" was introduced in [Bauer & Pawelzik 1992], based on ideas first discussed in the context of nonlinear dynamics. It is somewhat similar to the approach of [Villmann et al 1994],[2] in that they define a series of measures $Q(i,j)$ which give information about the preservation of neighbourhood relations at all possible scales. Briefly, $Q_1(i,j)$ is the distance between point i in the input space and its jth nearest neighbour as measured by distance orderings of their images in the output space, divided by the distance between point i in the input space and its jth nearest neighbour as measured by distance orderings in the input space. $Q_2(i,j)$ gives analogous information where i and j are points in the output space. The Q's are then combined to yield a single number P, the "topographic product", which defines the quality of the mapping:

$$P = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1}^{N-1} \log\left(\prod_{l=1}^{k} Q_1(j,l)Q_2(j,l)\right)^{\frac{1}{2k}}$$

Although originally expressed in terms of geometric spaces, the distance orderings in this definition could equally well be replaced by abstract similarity orderings that do not have a geometric interpretation. Again the concern is with orderings: $P = 0$ for a perfectly order-preserving map. [Bauer & Pawelzik 1992] show the application of the topographic product to dimension-reducing mappings of speech recognition data.

[2] The topographic product was introduced first; however it is more convenient for exposition purposes to explain it second.

3.3.4 The approach of Bezdek and Pal

[Bezdek & Pal 1995] also argue for a criterion that preserves similarity orderings rather than actual similarities. They call such a transformation "metric topology preserving" or MTP. The method they propose for calculating the discrepancy from an MTP map is to use a rank correlation coefficient between similarities in the two spaces, in particular Spearman's ρ. This is defined as the linear correlation coefficient of the ranks (see e.g. [Press et al 1988]):

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

where R_i and S_i are the corresponding rankings in the ordered lists of the F's and and G's. ρ has the useful property that it is bounded in the range $[-1, 1]$, and [Bezdek & Pal 1995] prove that $\rho = 1$ corresponds to an MTP map. Interesting comparisons are made in [Bezdek & Pal 1995] between the performance as measured by ρ of PCA, Sammon mappings and the SOM applied to various mapping problems. They conclude that the SOM generally performs significantly worse.

4 Comparing measures for the square to line problem

Comparisons between the performance of various mapping algorithms on particular problems have been illuminating [Durbin & Mitchison 1990, Bezdek & Pal 1995, Mao & Jain 1995, Lowe & Tipping 1996]. Here we ask a different set of questions. How do different measures compare in rating the same maps? Do the measures generally give a consistent ordering for different maps? How well does this ordering compare with intuitive assessments? How sensitive are these orderings to the measure of similarity employed? Answers to these questions reveal more about the relationships between different measures, and aid the choice of appropriate topography and similarity measures for particular problems.

We consider the very simple case of mapping between a regular 10×10 square array of points (V_{in}) and a regular 1×100 linear array of points (V_{out}). This is a paradigmatic example of dimension reduction, also used in [Durbin & Mitchison 1990, Mitchison 1995]. It has the virtue that solutions are easily represented. Figure 2 shows the four alternative maps considered (labelled A-D). Map A is an optimal minimal path solution. Maps B and C are taken from [Durbin & Mitchison 1990]: map B is an optimal minimal wiring solution [Mitchison & Durbin 1986], and map C is that found by the elastic net algorithm (for parameters see [Durbin & Mitchison 1990]).[3] Map D is a random mapping. Visual inspection suggests that this corpus should provide a range of different quality values.

4.1 Euclidean dissimilarities

First we take the F and G dissimilarity functions to be euclidean distances in each array, except for appropriate modifications for the minimal path and minimal wiring measures as described earlier. Later we investigate gaussian similarities. Figure 3 shows scatter plots of euclidean dissimilarities in the output space versus euclidean dissimilarities in the input space: note the various types of deviation from a straight or monotonically increasing line. The question is now how each measure trades off the different types of non-monotonicity apparent in figure 3. The actual numbers yielded by some of the measures discussed above are shown in table 1, and the relative ordering assigned to the different maps for all measures is shown in table 2. There are several points to note. As expected, all of the measures rate the random map D the worst, and $\rho \approx 0$ for this map. The Sammon, C, ρ and STRESS measures show remarkable agreement in their orderings, rating map A the best, despite being each formulated on quite different mathematical principles. It is a common intuition in the mapping literature that for a

[3] Qualitatively similar maps would be found by suitable variants of the SOM, such as [Angeniol et al 1988].

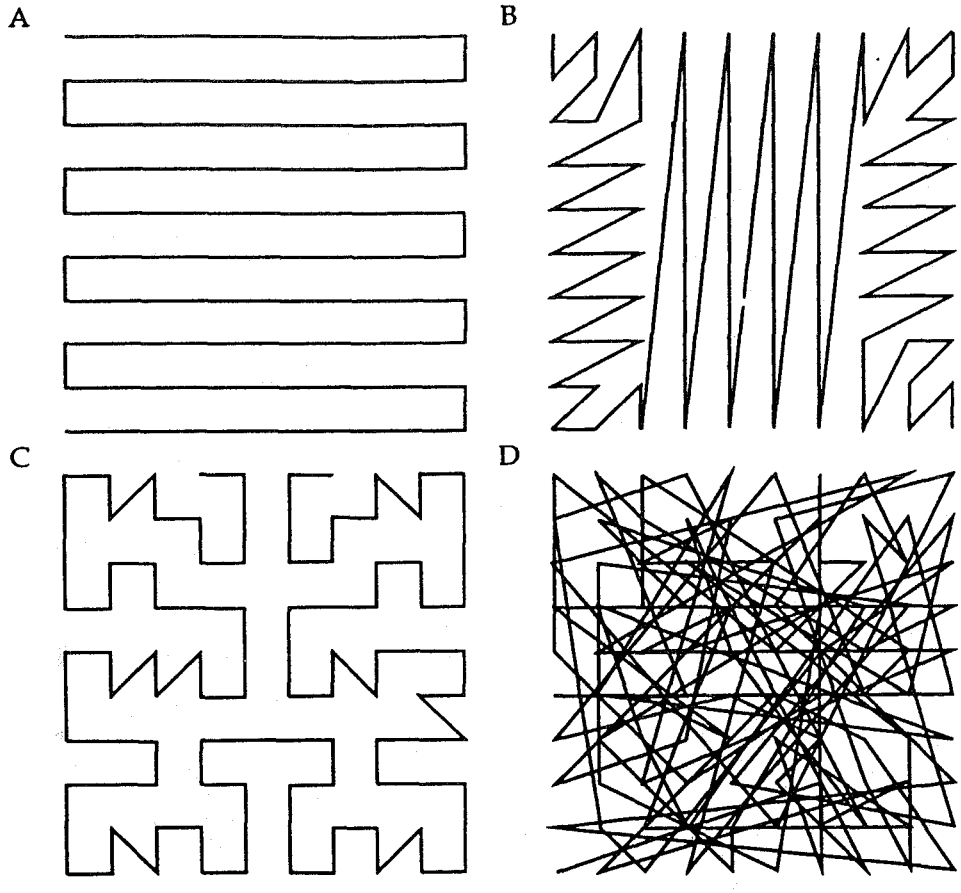

Figure 2: Some alternative mappings between a square array and a line.

case such as shown in figure 2 the best map is one that resembles a Peano curve, i.e. map C. However, for euclidean dissimilarities only the minimal distortion and topographic product measures agree with this assessment. As a control, we repeated the calculations using dissimilarities in both spaces that were euclidean distances raised to various powers between 0.5 and 2.0. In all cases the ordering for each measure was unchanged, except for minimal wiring, where the orderings obtained for powers 0.5 and 0.6 (in addition to power 1.0) are shown in table 2 (see also [Durbin & Mitchison 1990]).[4] With these additions, every possible ordering of the maps is represented in the table, given that D is always last. Spearman's ρ has the attractive feature that it is unchanged by any such monotonic transformation of the F's or G's. It also has the advantage of having a predictable value for random maps. More generally, it appears that the output of most measures is best treated as being at the ordinal level of measurement [Stevens 1951].

In terms of absolute costs, there is little discrimination between the three non-random maps given by the Sammon measure. This is because the range of F is from 1 to $\sqrt{200}$, whereas the range of G is from 1 to 100, and this inherent, map-independent mismatch dominates. Greater discrimination can be obtained by multiplying all G values by some number $\alpha < 1$ (e.g. $\alpha = \frac{\sqrt{200}}{100}$). For the minimum path measure the cost for map C is greater than that for map A by $8(\sqrt{2}-1)$, since there are 8 diagonal segments in map C. Although a minimum of the elastic net energy function is at map A, in practice (for unknown reasons) the algorithm tends to find slightly longer, more "folded" maps. This is analogous to the effect observed for a different mapping problem, where striped solutions are obtained even though these are not the global optimum [Goodhill & Willshaw 1990, Dayan 1993, Goodhill et al 1996].

[4] In a neurobiological context this exponent is intended to capture the "cost" of a length of axon. It is easy to construct biological arguments for a wide range of exponents, so it is interesting to consider the sensitivity of minimal wiring to this parameter.

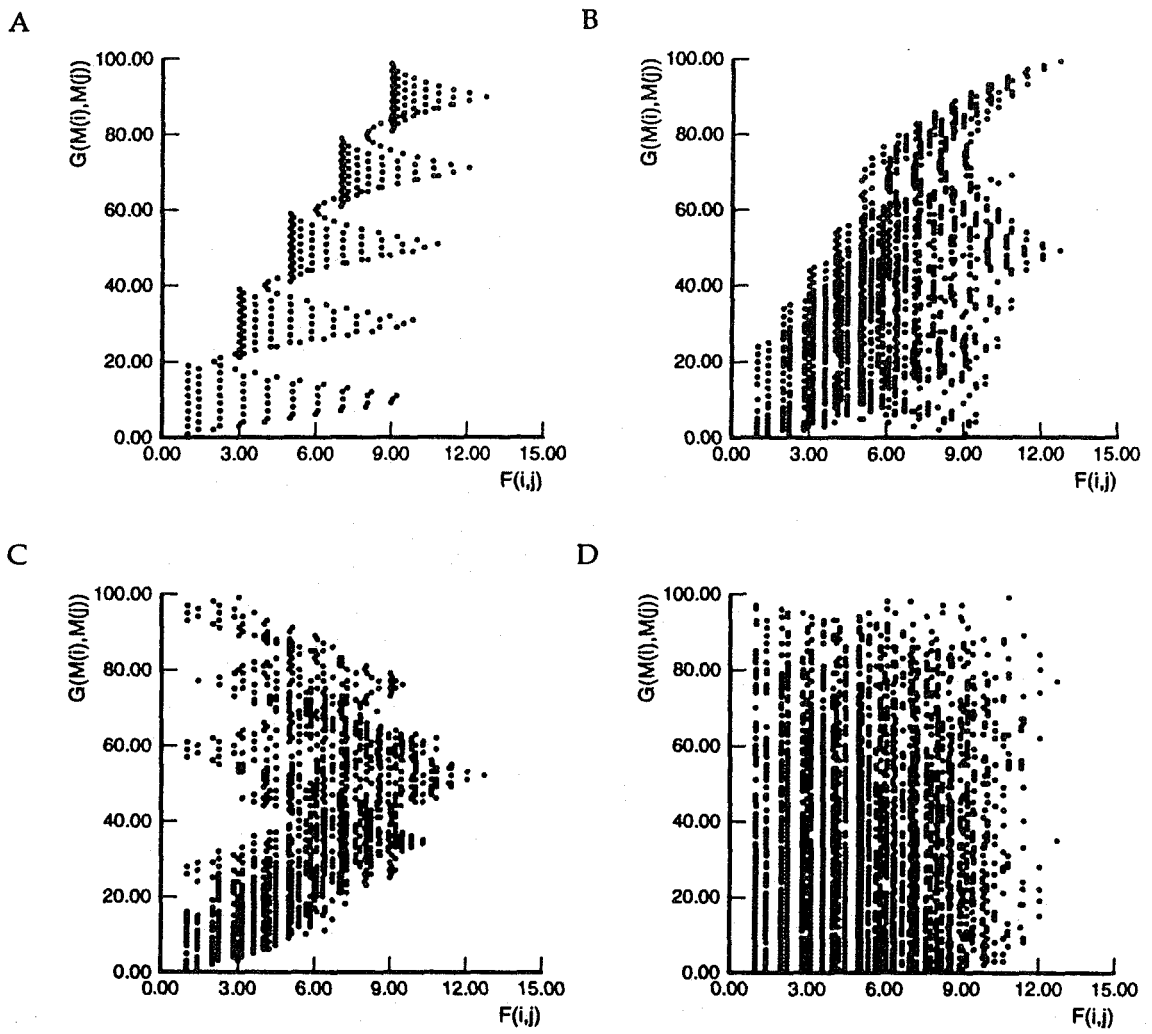

Figure 3: Scatterplots of G(M(i), M(j) against F(i, j) for the four maps shown in figure 2.

A general issue we have not considered is how ease of implementation and efficient optimization might bias a choice of measure.

Why is the "intuitively appealing" map C so rarely rated the best for euclidean dissimilarities? The answer lies in figure 3 and the way ordering violations at different scales are assessed by each measure (for further discussion see [Bauer & Pawelzik 1992]). For map C, the majority of points in the square map close on the line to their 4 neighbours in the square. However, for some points near the middle of the square, one neighbour is mapped to a point on the line a very large distance away (of the order of half the length of the line). In figure 3(c) these points are in the upper left region of the graph. These distances are far greater than the maximum equivalent distances in maps A or B, and accounts for the poor rating given to map C by several of the measures.

4.2 Gaussian similarities

In order to investigate the sensitivity of the ordering of maps to the similarity measure employed, we now consider the case where one of F or G is a gaussian function of euclidean distance. This implies that one is now concerned only with local neighbourhoods in one of the spaces, on a scale determined by the width of the gaussian. One trivial consequence is that increasing measures become decreasing

	Min wiring	Min path	MD	TP	Sammon	MDS	ρ	STRESS
Type	Decrease	Decrease	Decrease	See caption	Decrease	Decrease	Increase	Decrease
Map A	990	99.0	309.9	−0.04608	38.94	6366242	0.6498	0.423
Map B	914	159.9	940.9	−0.04843	39.98	6383306	0.6413	0.442
Map C	1140	102.3	214.7	−0.03244	49.93	6455352	0.6332	0.519
Map D	5458	484.9	3931.7	−0.1638	70.02	6716864	0.0502	0.594

Table 1: Costs for the four maps given by measures discussed in the text. "Type" tells how the measure changes with increasing mapping quality, so that for type "decrease" small numbers mean better maps. The topographic product (TP) is decreasing in absolute value, and the negative values reflect the fact the dimension of the output space (the line) is "too small". Spearman's ρ was calculated using the routine **spear** of [Press et al 1988], and STRESS by the program ALSCAL [Young & Harris 1990]. For the minimal distortion measure, σ was taken to be 2.0.

Ranking	MW 1.0	MW 0.5	MW 0.6	Min path	MD	TP	Sammon	MDS	ρ	STRESS
1	B	C	B	A	C	C	A	A	A	A
2	A	B	C	C	A	A	B	B	B	B
3	C	A	A	B	B	B	C	C	C	C
4	D	D	D	D	D	D	D	D	D	D

Table 2: Relative ordering of the 4 maps given by each measure. First three columns show minimum wiring (MW) for different powers (see section 4.1).

and vice versa (cf table 1). For instance, a good value of ρ is now one close to −1. A more complex issue is what happens more generally to measures based on orderings. The gaussian transformation is monotonic, therefore these measures should be unaffected (given the increasing/decreasing reversal). However, from a practical perspective, all values of G(p, q) for the euclidean distance between p and q greater than a few standard deviations will be effectively zero. If the standard deviation is very much less than the maximum distance, this implies a large degree of degeneracy in the orderings. Measures which require rank orderings to be calculated, such as ρ, are not well-equiped to deal with the majority of the G's being equal, and give unreliable results. This is illustrated in figure 4. The issue of tied ranks has been much discussed in the NMDS literature [Kruskal 1964a]; however, it has been shown that if there are few distinct levels in the similarity matrix then spurious results can be produced [Simmen et al 1994, Goodhill, Simmen & Willshaw 1995]. Since the Topographic Product, ρ, and NMDS measures all involve calculating rank orderings, we do not consider them in this section. Measures that match F's and G's are also clearly inappropriate when F or G is a gaussian function of distance, which discounts the metric MDS and Sammon measures. From table 1 this leaves just the C measure, since this includes minimal path and minimal wiring. We now explore this case in more detail.

4.2.1 Gaussian G

The case of gaussian G and euclidean F is closely related to Luttrell's original minimal distortion measure (see section 3.1.5). It can also be thought of as a generalized version of the minimal path measure: besides keeping nearest neighbours in the output space nearby in the input space, it is attempted to also keep further neighbours nearby. This has profound consequences. Consider moving along the line in map A. In just a few steps one has moved a substantial distance across the square. However in map C, moving a few steps along the line generally causes a much smaller displacement across the square. Thus, when more than first neighbours are considered, it becomes favourable to "turn corners" and form a Peano-type curve, as in map C. For instance, using a G of one times nearest-neighbours plus only 0.1 times second nearest neighbours gives map C as the best. This behavior is

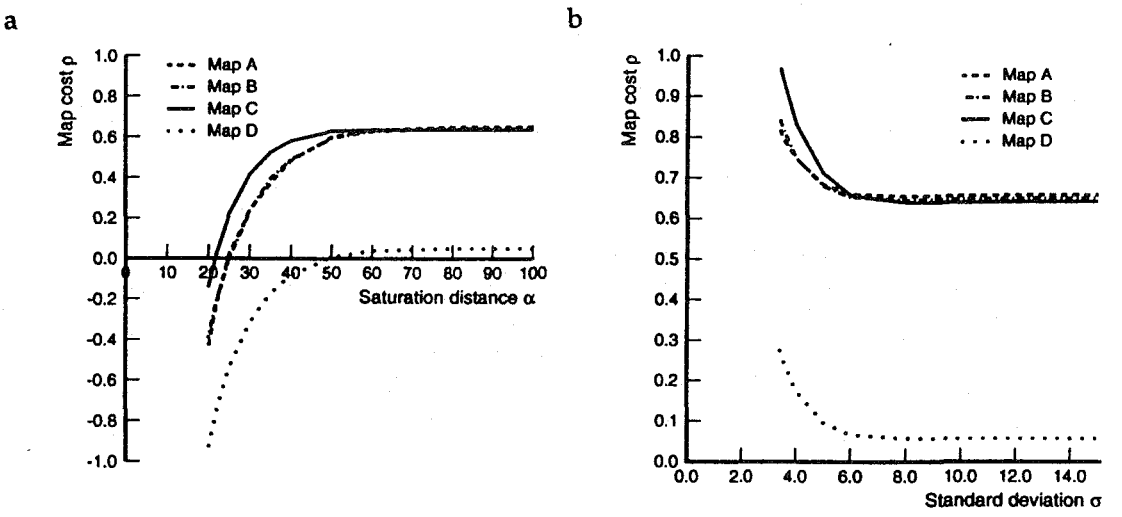

Figure 4: The effect on measure ρ of increasing degeneracy in the G values. (a) G is given by euclidean distance up to α, and then saturates at α for all greater values. ρ is stable for $\alpha > 60$, but as α decreases and the amount of degeneracy increases, ρ starts to change rapidly. Although the orderings of the maps are roughly preserved, the actual values of ρ cease to be meaningful. (b) G is given by e^{-d^2/σ^2}, d = euclidean distance (sign of y axis is reversed). In theory, ρ should have the same value for all σ. In practice, when σ is very much less than the maximum d (100 in this case), the calculation of ρ fails to yield meaningful vaules. In both (a) and (b), for smaller values of α and σ than those plotted, the **spear** routine of [Press et al 1988] gave a run-time error.

illustrated more generally in figure 5. Thus the intuition that C is a good map is correct for the case of euclidean F and gaussian G. Although this case is closely related to the SOM, in the SOM algorithm the size of the neighbourhood function is constantly shrinking. This lack of a fixed neighbourhood scale makes it hard to pin down exactly what mapping problem the algorithm is attempting to solve.

4.2.2 Gaussian F

Gaussian F is equivalent to minimal wiring extended to a larger neighbourhood. The value of C for the four maps as the size of this neighbourhood changes in shown in figure 6. Map B eventually loses the lead in favor of map A.

5 Near-optimal maps for the square to line problem

Above we quantitatively compared different measures for four given maps for the square to line problem. A more qualitative insight into the relationship between measures can be obtained by comparing the *optimal* map for each measure. Calculation of the global optimum by exhaustive search is clearly impractical: there are of the order of 100! possible mappings for this problem, and the continuous optimization algorithms that exist for some of the measures (e.g. Sammon) cannot be used in this discrete case. Instead we employ simulated annealing [Kirkpatrick et al 1983] to find a solution close to optimal. The parameters used are shown in table 3. We tested the efficacy of this technique by first applying it to the minimal path and minimal wiring problems, where optimal solutions are explicitly known. Maps with costs within about 1% of the optimal value were found (figure 7).

The objective functions optimized were metric MDS, Sammon, Spearman, and Minimal Distortion for

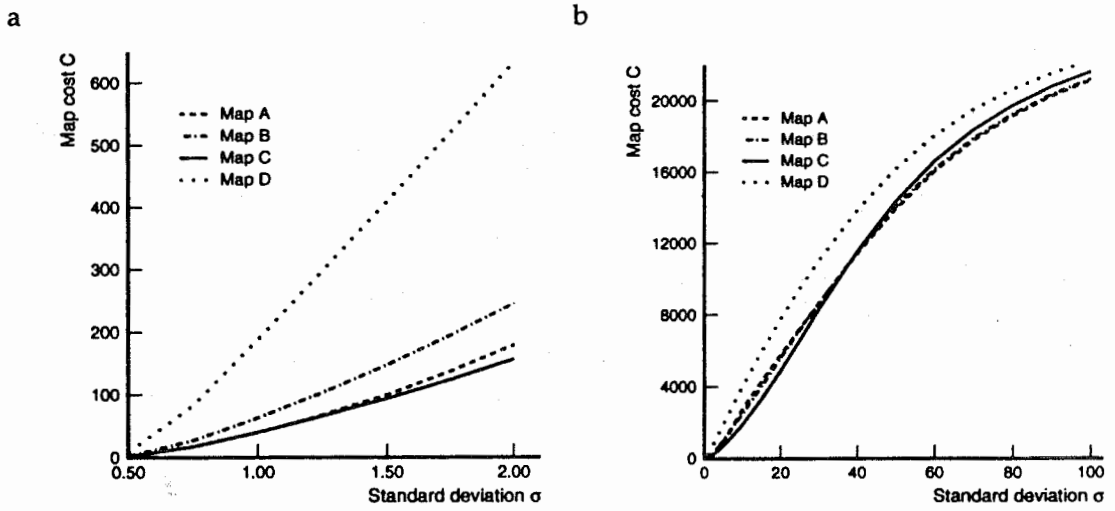

Figure 5: Map costs for $G = e^{-d^2/\sigma^2}$. (a) Small σ. Map A is initially the best, but is quickly beaten by map C as σ increases (crossing point $\sigma \sim 1$). (b) Large σ. Map C eventually relinquishes the lead back to map A.

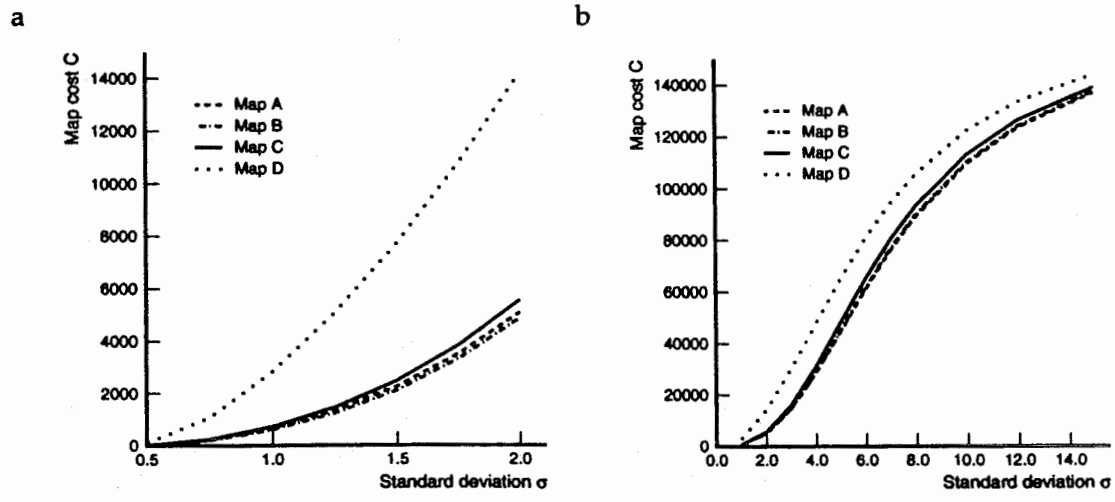

Figure 6: Map costs for $F = e^{-d^2/\sigma^2}$. (a) Small σ. Map B is initially the best, but is beaten by map A for σ greater than about 4.0. (b) Large σ. The same overall profile is seen as in the gaussian G graph.

Parameter	Value
Initial map	Random
Move set	Pairwise interchanges
Initial temperature	3× mean energy difference over 10,000 moves
Cooling schedule	Exponential
Cooling rate	0.998
Acceptance criterion	1000 moves at each temperature
Upper bound	10,000 moves at each temperature
Stopping criterion	Zero acceptances at upper bound

Table 3: Parameters used in simulating annealing runs: for further explanation see e.g. [van Laarhoven & Aarts 1987].

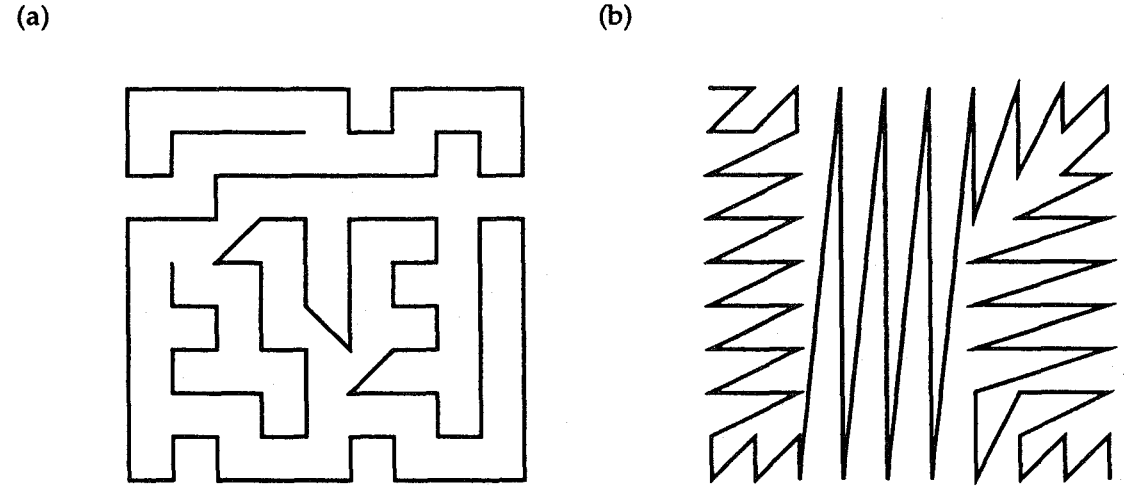

Figure 7: Testing the minimization algorithm for cases where the optima are known. (a) Mininal path length solution, cost = 100.243, 1.3% larger than the optimal of 99.0. (b) Minimal wiring solution, cost = 917.0, 0.3% larger than the optimal of 914.0 [Durbin & Mitchison 1990]. An optimal minimal path length solution was found when the cooling rate was increased to 0.9999 and the upper bound increased to 100,000 moves; however it was computationally impractical to run all the simulations this slowly.

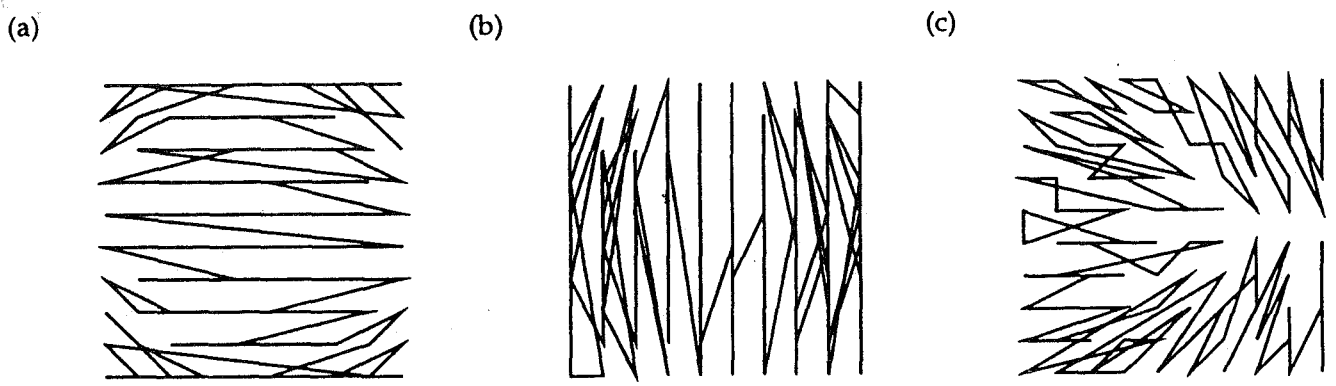

Figure 8: (a) Metric MDS measure, cost = 6352324. (b) Sammon measure, cost = 38.5. (c) Spearman measure, cost = 0.698. Note that, as expected, for each measure the value is better than the best value given in table 1 for the four exemplar maps.

a range of σ.[5] Figure 8 shows the maps found by this procedure 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, in the fifth column of the square for the optimal Sammon map, the vertical component for neighbouring points in the line progresses in the order 2, 3, 5, 9, 4, 7, 8, 6, 10, 1. This local discontinuity arises because these measures take into account neighbourhood preservation at all scales: local continuity is not privileged over global continuity, and global concerns dominate.

Figure 9 shows minimal distortion solutions for varying σ. For small σ, the solution resembles the minimal path optimum of figure 7(a), since the contribution from more distant neighbours compared to 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 8.

5.1 Classes of mappings

Three qualitatively different types of map are apparent from figures 8 and 9, which we will refer to as classes 1, 2 and 3. For class 1, the MDS and Sammon measures (figure 8(a,b)), the line progresses through the square by a series of locally discontinuous jumps along one axis of the square (horizontal in the MDS case, vertical in the Sammon case; which axis is chosen is the result of random symmetry breaking), and a comparatively smooth progression in the orthogonal direction. One consequence of this is that nearby points in the square never map too far apart on the line.[6] For class 2, the Spearman measure and minimal distortion for $\sigma = 20$ (figures 8(c) and 9(d)), the strategy is similar except that there is now one fold to the map. This decreases the size of local jumps in following the line through the square, at the cost of introducing a "seam" across which nearby points in the square map a very long way apart on the line. For class 3, minimal distortion for small σ (figure 9(a-c)), the strategy is to introduce several folds. This gives strong local continuity, at the cost that now many points that are nearby in the square map to points that are far apart on the line.

[5] We did not optimize the topographic product or STRESS measures due to the large amount of computation involved in calculating these measures.

[6] One might expect that for a rectangle mapping to a line, the optimal maps for these measures would have discontinuous jumps parallel to the short axis of the rectangle and a relatively smooth progression parallel to the long axis. We optimized the Sammon measure for a 20 × 5 rectangle mapping to a 1 × 100 array, and obtained this result.

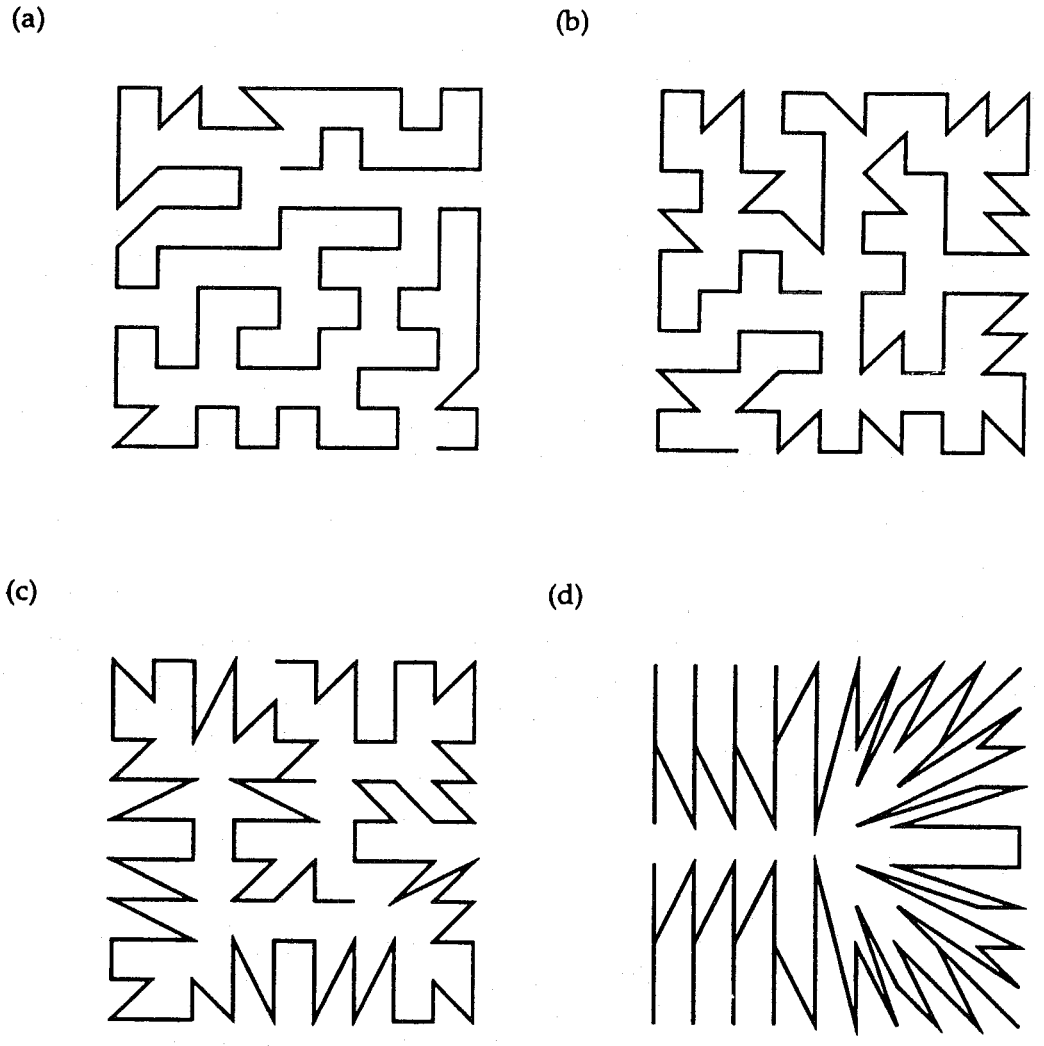

Figure 9: Minimal distortion solutions. (a) $\sigma = 1.0$, cost = 43.3. (b) $\sigma = 2.0$, cost = 214.7. (c) $\sigma = 4.0$, cost = 833.2. (d) $\sigma = 20.0$, cost = 18467.1.

6 Discussion

6.1 Lessons from comparison of the four exemplar maps

We have shown that, of the four maps, the one considered best depends on both the measure of topography and the measure of similarity employed. Thus, comparisons of maps should be qualified by precise statements about which measures are being used. For instance, self-organizing algorithms such as the elastic net and the SOM produce maps of the same form as map C. This is a good mapping only under particular interpretations of what the mapping problem actually is (e.g. euclidean F and gaussian G similarity measures, minimal distortion topography measure). When such assumptions are justified for various mapping problems needs to be addressed. In particular, these assumptions favor the preservation of *local* structure in the output to input direction. It is generally not discussed why this should be an appropriate goal for any particular problem (though see [Durbin & Mitchison 1990]).

6.2 Lessons from the calculation of optimal maps

What do the optimal maps tell us about which measures are most 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 the 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 3. 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 3 have proved useful [Durbin & Mitchison 1990]. However, very few other objectives have been applied to this problem, so it is still an open question which are most appropriate. Class 2 represents a form that has been hitherto unappreciated. There may be some applications for which such maps are worthwhile, perhaps in a neurobiological context for understanding why different input variables are sometimes mapped into different areas rather than interdigitated in the same area.

6.3 Many-to-one mappings

We have discussed only the case of one-to-one mappings. In many practical contexts there are many more points in V_{in} than V_{out}, and it is necessary to also specify a many-to-one mapping from points in V_{in} to N "exemplar" points in V_{in}, where N = number of points in V_{out}. It may be desirable to do this adaptively while simultaneously optimizing the form of the map from V_{in} to V_{out}. For instance, shifting a point from one cluster to another may increase the "clustering cost", but by moving the positions of the cluster centers decrease the sum of this and the "continuity cost". The elastic net, for instance, trades off these two contributions explicitly with a ratio that changes during the minimization, so that finally each cluster contains only one point and the continuity cost dominates ([Durbin & Willshaw 1987]; for discussions see [Simic 1990, Yuille 1990]). The SOM trades off these two contributions implicitly. [Luttrell 1994] discusses allowing each point in the input space to map to many in the output space in a probabilistic manner, and vice-versa.

6.4 Further biological considerations

Minimal wiring considerations are a powerful tool for understanding brain connectivity [Mitchison 1991, Mitchison 1992]. Here we have discussed the effect of including neighbourhoods on a larger scale than only nearest neighbours. A particular biological problem that has been much discussed in the context of mappings is to understand the pattern of interdigitation in the primary visual cortex of cat and monkey of attributes of the visual image such as spatial position, orientation, ocular dominance, and disparity. It has been proposed several times that the form of this map is a result of a desire to preserve neighbourhoods between a high-dimensional space of features and the two-dimensional cortex (e.g. [Durbin & Mitchison 1990, Goodhill & Willshaw 1990, Swindale 1992, Obermayer et al 1992]). Usually geometric distance in the high-dimensional space is taken to represent dissimilarity. The approach outlined in the present paper suggests that one way to proceed in addressing this and related biological mapping problems is to (1) formulate some reasonable way of specifying the similarity functions for the space of input features and the cortex (this could for instance be in terms of the correlations between different input variables, and intrinsic cortical connections, respectively), then (2) explore which sets of mapping choices give maps resembling those seen experimentally, taking account of biological constraints. A first step in this direction for interdigitated maps such as ocular dominance columns in primary visual cortex is taken in [Goodhill et al 1996]. A consideration of how to optimize some measure of topography by, in addition to adapting the mapping, allowing the similarity function in the cortex to change (within some constrained range), could provide insight into the development of lateral connections in the cortex (cf [Katz & Callaway 1992]).

However, one-to-one mappings are rare in biological contexts. Rather more frequently axonal arbors form many-to-many connections with dendritic trees. There are several conceivable ways in which for instance the C measure could be generalized to allow for a weighted match of each point in V_{in} to many in V_{out}, and vice-versa (cf [Luttrell 1994]).

Acknowledgements

We thank Christopher Longuet-Higgins for original inspiration, and Steve Finch for very useful discussions at an earlier stage of this work. We are especially grateful to Graeme Mitchison, Steve Luttrell and Hans-Ulrich Bauer for helpful discussions and comments on the manuscript. In addition, we thank Peter Dayan, Klaus Pawelzik, Martin Simmen and Paul Viola for stimulating discussions of these ideas, and Hans-Ulrich Bauer for very kindly supplying us with code for calculating the Topographic Product. Also we thank Eric Mjolsness and Dimitris Tsioutsias for helpful pointers to the quadratic assignment problem literature. Research was supported by the Sloan Center for Theoretical Neurobiology at the Salk Institute and the Howard Hughes Medical Institute.

Bibliography

Angeniol, B., de la Croix Vaubois, G. & Le Texier, J. (1988). Self-organizing feature maps and the traveling salesman problem. *Neural Networks*, **1**, 289-293.

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

Bauer, H.U. & Pawelzik, K.R. (1992). Quantifying the neighbourhood preservation of self-organizing feature maps. *IEEE Trans. Neural Networks*, **3**, 570-579.

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

Burkard, R.E. (1984). Quadratic assignment problems. *Europ. Journ. Oper. Res.*, **15**, 283-289.

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

Dayan, P.S. (1993). Arbitrary elastic topologies and ocular dominance. *Neural Computation*, **5**, 392-401.

Demartines, P. & Hérault, J. (1995). CCA: "Curvilinear Component Analysis". In: *Proc. 15th Workshop GRETSI*, Juan-Les-Pins, France.

Demartines, P. & Hérault, J. (1996). Curvilinear component analysis: a self-organizing neural network for non-linear mapping of data sets. Manuscript submitted for publication.

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

Durbin, R. & Willshaw, D.J. (1987). An analogue approach to the traveling salesman problem using an elastic net method. *Nature*, **326**, 689-691.

Erwin, E., Obermayer, K. & Schulten, K. (1992). Self-organizing maps: convergence properties and energy functions. *Biol. Cybern.*, **67**, 47-55.

Finke, G., Burkard, R.E. & Rendl, F. (1987). Quadratic assignment problems. *Annals of Discrete Mathematics*, **31**, 61-82.

Goodhill, G.J., Finch, S. & Sejnowski, T.J. (1996). Optimizing cortical mappings. To appear in *Advances in Neural Information Processing Systems*, **8**, eds. David S. Touretzky, Michael C. Mozer & Michael E. Hasselmo, MIT Press: Cambridge, MA.

Goodhill, G.J., Simmen, M., & Willshaw, D.J. (1995). An evaluation of the use of Multidimensional Scaling for understanding brain connectivity. *Phil. Trans. Roy. Soc. B*, **348**, 265-280.

Goodhill, G.J. & Willshaw, D.J. (1990). Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network*, **1**, 41-59.

Hardy, G.H., Littlewood, J.E. & Pólya, G. (1934). Inequalities. Cambridge University Press.

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

Jones, D.G., Van Sluyters, R.C. & Murphy, K.M. (1991). A computational model for the overall pattern of ocular dominance. *J. Neurosci.*, **11**, 3794-3808.

Katz, L.C. & Callaway, E.M. (1992). Development of local circuits in mammalian cortex. *Ann. Rev. Neurosci.*, **15**, 31-56.

Kendall, D.G. (1971). Construction of maps from "odd bits of information". *Nature*, **231**, 158-159.

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

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

Kohonen, T. (1988). Self-organization and associative memory. Springer, Berlin.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-27.

Kruskal, J.B. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115-129.

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

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

Lawler, E.L. (1963). The quadratic assignment problem. *Management Science*, **9**, 586-599.

Lowe, D. & Tipping, M. (1996). Feed-forward Neural Networks and Topographic Mappings for Exploratory Data Analysis. *Neural Computing and Applications*, **4**, 83-95.

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

Luttrell, S.P (1994). A Bayesian analysis of self-organizing maps. *Neural Computation*, **6**, 767-794.

Mao, J. & Jain, A.K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Networks*, **6**, 296-317.

Marr, D. (1982). Vision. W.H. Freeman and Company, New York.

Martinetz, T. & Schulten, K. (1994). Topology representing networks. *Neural Networks*, **7**, 507-522.

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

Mitchison, G. (1992). Axonal trees and cortical architecture. *Trends Neurosci.*, **15**, No. 4, 122-126.

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

Mitchison, G. & Durbin, R. (1986). Optimal numberings of an N × N array. *SIAM J. Alg. Disc. Meth.*, **7**, 571-581.

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

Obermayer, K., Blasdel, G.G. & Schulten, K. (1992). Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Phys. Rev. A*, **45**, 7568-7589.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1988). Numerical recipes in C: the art of scientific computing. Cambridge University Press: Cambridge.

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

Sejnowski, T.J., Koch, C. & Churchland, P.S. (1988). Computational Neuroscience. *Science*, **241**, 1299-1306.

Shepard, R.N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, 125-140.

Shepard, R.N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, 219-246.

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

Simić, P.D. (1990). Statistical mechanics as the underlying theory of 'elastic' and 'neural' optimizations. *Network*, **1**, 89-103.

Simić, P.D. (1991). Constrained nets for graph matching and other quadratic assignment problems. *Neural Computation*, **3**, 268-281.

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

Stevens, S.S. (1951). Mathematics, measurement and psychophysics. In *Handbook of experimental psychology*, ed S.S. Stevens, 1-49. New York: Wiley.

Swindale, N.V. (1992). Elastic nets, travelling salesmen and cortical maps. *Current Biology*, **2**, 429-431.

Takane, Y., Young, F.W. & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.

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

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

Villmann, T., Der, R., Herrmann, M. & Martinetz, T. (1994). Topology preservation in SOFMs: general definition and efficient measurement. In: *Informatik Aktuell - Fuzzy-Logik*, Ed. B. Reusch, Springer-Verlag, 159-166.

Young, F. W. (1987). Multidimensional scaling: history, theory, and applications. Hillsdale, New Jersey: Lawrence Erlbaum.

Young, F. W. & Harris, D.F. (1990). Multidimensional scaling: procedure ALSCAL. In *SPSS base system user's guide* (ed. M. Norusis), 397-461. Chicago: SPSS.

Yuille, A.L. (1990). Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, **2**, 1-24.