Topological mapping for visualisation of high-dimensional historical linguistic data

Hermann Moisl
Newcastle University, UK
hermann.moisl@ncl.ac.uk

Introduction

Discovery of the chronological or geographical distribution of collections of historical text can be more reliable when based on multivariate rather than on univariate data because, assuming that the variables describe different aspects of the texts in question, multivariate data necessarily provides a more complete description. Where the multivariate data is high-dimensional, however, its complexity can defy analyis using traditional philological methods. Increasingly, the first step in interpreting such complexity is cluster analysis because it gives insight into structure latent in the data, thereby facilitating hypotheses which can then be tested using a range of other mathematical and statistical methods (Moisl 2015).

The present discussion addresses an issue in cluster analysis whose importance in quantitative and corpus linguistics has, in my view, thus far not been sufficiently well appreciated: the possibility that the data is nonlinear. Most applications of cluster analysis in these fields use linear proximity measures which simply ignore any nonlinearity, and, if the data really is significantly nonlinear, can give misleading results.

The discussion is in three main parts: the first part outlines the nature of nonlinearity in data generally and in linguistic data specifically, the second shows why nonlinearity is a problem for linear clustering methods, and the third shows how topological mapping can be used to cluster high-dimensional data in a way that takes any nonlinearity into account.

1. Nonlinearity

For greater detail on what follows, see (Moisl 2015, chapters 3 and 4).

1.1 *Nonlinearity in natural processes*

In natural processes there is a fundamental distinction between linear and nonlinear behavior. Linear processes have a constant proportionality between cause and effect. If a ball is kicked *x* hard and it goes *y* distance, then a 2*x* kick will appear to make it go 2*y*, a 3*x* kick 3*y*, and so on. Nonlinearity is the breakdown of such proportionality. In the case of our ball, the linear relationship increasingly breaks down as it is kicked harder and harder. Air and rolling resistance become significant factors, so that for, say, 5*x* it only goes 4.9 *y*, for 6*x* 5.7*y*, and again so on until eventually it bursts and goes hardly any distance at all. Such nonlinear effects pervade the natural world and gives rise to a wide variety of complex and often unexpected --including chaotic—behaviours (Strogatz 2000; Bertuglia & Vaio 2005).

1.2 *Nonlinearity in data*

Data is a description of objects from a domain of interest in terms of a set of variables such that each variable is assigned a value for each of the objects. Given *m* objects described by *n* variables, a standard representation of data for computational analysis is a matrix M in which each of the *m* rows represents a different object, each of the *n* columns represents a different variable, and the value at $M_{i,j}$ describes object *i* in terms of variable *j*, for $i = 1..m$, $j = 1..n$. The matrix thereby makes the link between the researcher's conceptualization of the domain in terms of the semantics of the variables s/he has chosen and the actual state of the world, and allows the resulting data to be taken as a representation of the domain based on empirical observation.

M is linear when the functional relationships between all its variables, that is, the values in its columns, conform to the mathematical definition of linearity. In mathematics, a linear function *f* is one that satisfies the following properties, where *x* and *y* are variables and *a* is a constant (Lay 2010):

- Additivity: $f(x+y) = f(x) + f(y)$ -- adding the results of *f* applied to *x* and *y* separately is equivalent to adding *x* and *y* and then applying *f* to the sum.
- Homogeneity: $f(ax) = af(x)$ -- multiplying the result of applying *f* to *x* by a constant is equivalent to multiplying *x* by the constant and then applying *f* to the result.

A function which does not satisfy these two properties is nonlinear, and so is a data matrix in which the functional relationships between two or more of its columns are nonlinear.

Matrices have a geometrical interpretation. For each row vector of M:

- The dimensionality of the vector, that is, the number of its components *n*, defines an *n*-dimensional Euclidean space.
- The sequence of *n* numbers comprising the vector specifies the coordinates of the vector in the space.
- The vector itself is a point at the specified coordinates

The set of row vectors in M defines a configuration of points in the *n*-dimensional space called the data manifold.

Linear manifolds are shapes consisting of straight lines and flat planes and represent linear data, whereas nonlinear manifolds consist of curved lines and surfaces and represent nonlinear data; examples are given in Figure 1.

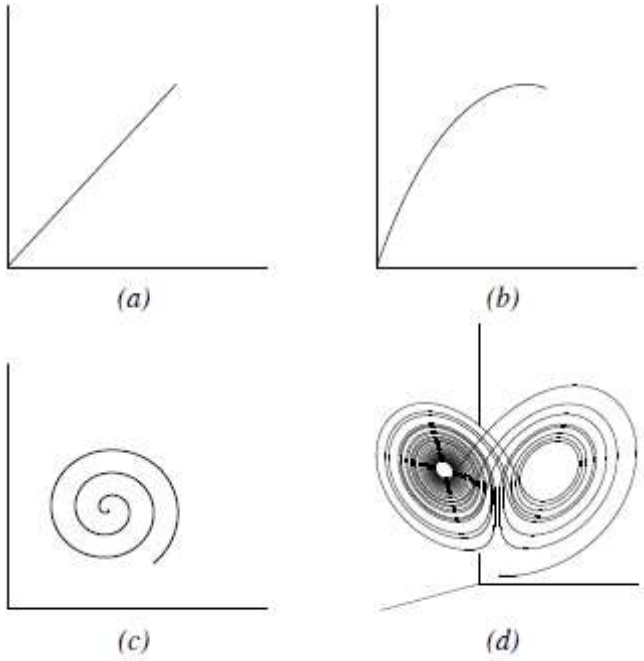
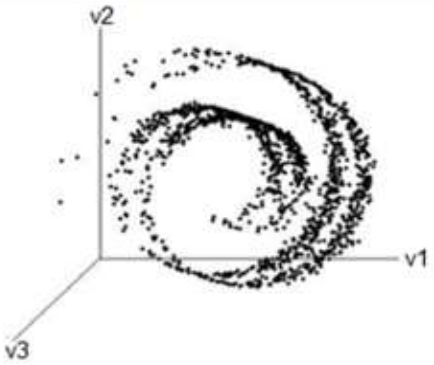


Figure 1: Linear and nonlinear manifolds in two and three dimensional space

An essentially unlimited range of nonlinear manifolds is possible in any dimensionality. Figure 2 gives another example of a nonlinear manifold in three-dimensional space.

	V1	V2	V3
1	5.214	7.087	6.729
2	0.421	8.426	7.896
3	6.134	5.686	6.088
...	...	...	...
1600	0.945	11.501	10.867

a: 1600 3-dimensional vectors



b: The vectors in (a) plotted in 3-dimensional space

Figure 2: Nonlinear manifold in three dimensional space

1.3 *Nonlinearity in linguistic data*
Data abstracted from a natural process known to be linear is itself guanteed to be linear. Data abstracted from a known nonlinear process is not necessarily nonlinear, but may be. The human brain - the generator of language - is a nonlinear dynamical system that exhibits highly complex physical behaviour in which nonlinearity arises on account of latency and saturation effects in individual neuron and neuron assemblies. One must, therefore, always reckon with the possibility that data abstracted from speech or text will be nonlinear.

2. **The problem**
The problem that nonlinearity poses for cluster analysis of high-dimensional multivariate data is easily seen. Commonly-used methods such as PCA for projection into two- or three-dimensional space for graphical display, or hierarchical analysis using proximity measures like the Euclidean, are linear: they take no account of any curvature in the manifold, and can thereby introduce distortions into visualization results in some proportion to the degree of nonlinearity in the manifold. This is shown in Figure 3 for three-dimensional data, but the situation extends to any dimensionality.

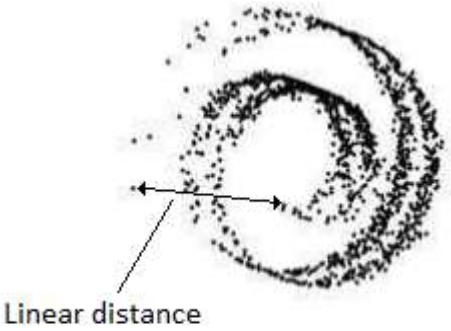


Figure 3: Linear distance between points on a nonlinear manifold

3. **Topological mapping**

3.1 *Topology*
Topology is an aspect of mathematics that grew out of the vector space geometry we have been using so far in the discussion. Its objects of study are manifolds, but these are studied as spaces in their own right, topological spaces, without reference to any embedding vector space and associated coordinate system (Lee 2010). Topology would, for example, describe the points which constitute the manifold embedded in the vector space of Figure 4a independently of the three-dimensional coordinates, as in Figure 4b.

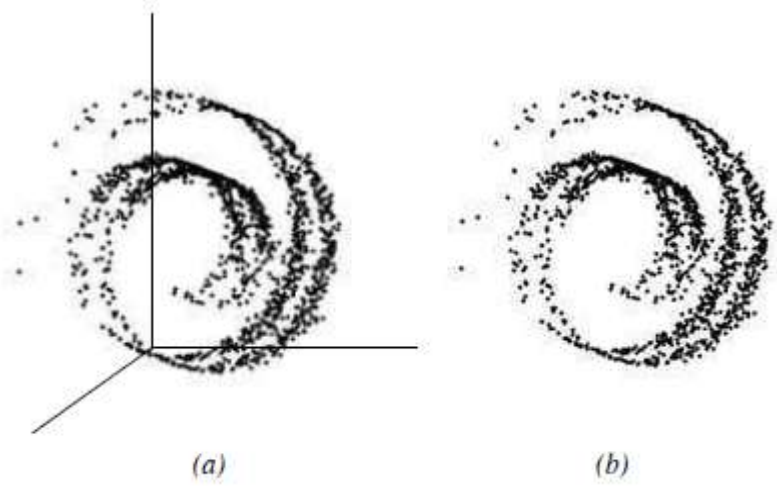


Figure 4: A manifold embedded in a three-dimensional doordinate system and as a topological object

Topology replaces the concept of vector space and associated coordinate system with relative nearness of points to one another in the manifold as the mathematical structure defined on the underlying set; relative nearness of points is determined by a function which, for any given point *p* in the manifold, returns the set of all points within some specified proximity *e* to *p.* The set of all points in proximity *e* to *p* constitute the neighbourhood of *p*, as in Figure 5.

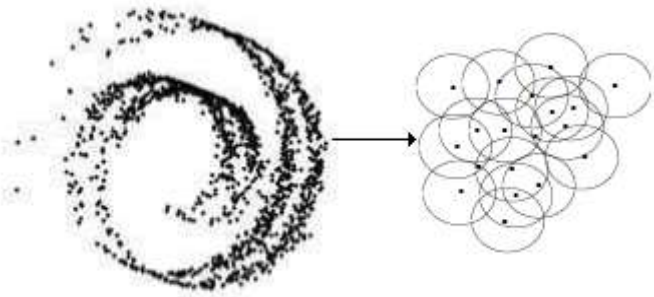


Figure 5: Overlapping topological neighborhoods on a manifold

3.2 *Projection of topological structure into low-dimensional space*
High-dimensional manifolds can be visualized as low-dimensional ones by means of projection in which the topology of the high-dimensional manifold, that is, the neighbourhood structure, is preserved in the low-dimensional one, so that points close to one another in high dimensions are close to one another in the low-dimensional projection. This can be conceptualized as in Figure 6, where a three-dimensional manifold is projected onto a two-dimensional surface.

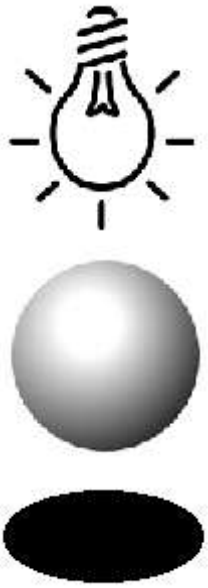


Figure 6: Projection from three to two dimensions

3.3 *Preservation of nonlinearity*
The set of neighbourhoods which constitutes the topology of a manifold by definition follows that surface of the manifold, whatever its shape. Because a projection preserves the topology, that shape is preserved - in other words, nonlinearity is preserved in the projection.

3.4 *Example*
The aim of this section is to show how topological mapping can be used to discover structure in high-dimensional multivariate data abstracted from a multi-document corpus. It does this by using a particular topological mapping method, the self-organizing map (SOM), to infer the relative chronology of a collection of Old English, Middle English, and Early Modern English texts from spelling data abstracted from them.

3.4.1 *The text collection*

Old English	Middle English	Early Modern English
Exodus	Sawles Warde	King James Bible
Phoenix	Henryson, Testament of Cressid	Campion, Poesie
Juliana	The Owl and the Nightingale	Milton, Paradise Lost
Elene	Malory, Morte Darthur	Bacon, Atlantis
Andreas	Gawain and the Green Knight	More, RichardIII
Genesis	Morte Arthure	Shakespeare, Hamlet
Beowulf	KingHorn	Jonson, Alchemist
	Alliterative Morte Arthure	
	Bevis Of Hampton	
	Chaucer, Troilus	
	Langland, Piers Plowman	
	York Plays	
	Cursor Mundi	

3.4.2 *Spelling data*

Spelling is used as the basis for inference of the relative chronology of the above texts on the grounds that it reflects the phonetic, phonological, and morphological development of English over time. The variables used to represent spelling in the texts are letter pairs: for 'the cat sat', the first letter pair is (t,h), the second (h,e), the third (e,<space>), and so on. All distinct pairs across the entire text collection were identified, and the number of times each occurs in each text was counted. A fragment of the resulting data matrix exemplifies this.

	1. hw	2. we	3. fe	...	841. jm
Exodus	35	149	125	...	0
Sawles Warde	52	147	45	...	0
...	...	...	...	...	...
King James	0	42	36	...	0

The matrix was normalized to compensate for variation in document length, and truncated to the most important 100 letter pairs. Details of normalization and truncation are available in (Moisl 2015, ch. 3).

3.4.3 *The self-organizing map (SOM)*

The self-organizing map (Kphonen 2001) is a topological mapping method. It is an artificial neural network that was originally invented to model a particular kind of biological brain organization, but that can also be used without reference to neurobiology as a way of visualizing high-dimensional data manifolds by projecting and displaying them in low-dimensional space. It has been extensively and successfully used for this purpose across a wide range of disciplines. Figure 7 shows the architecture of the SOM.

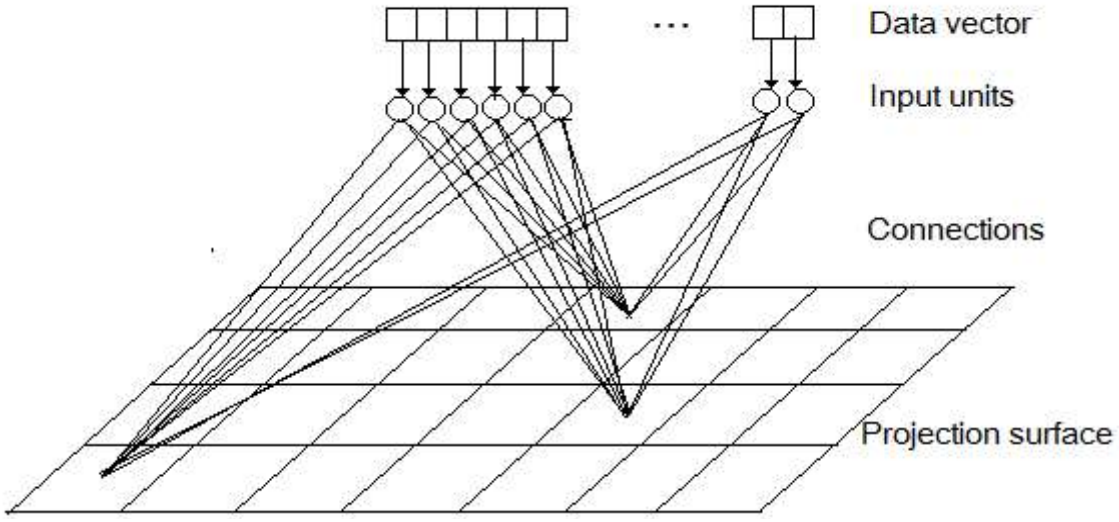
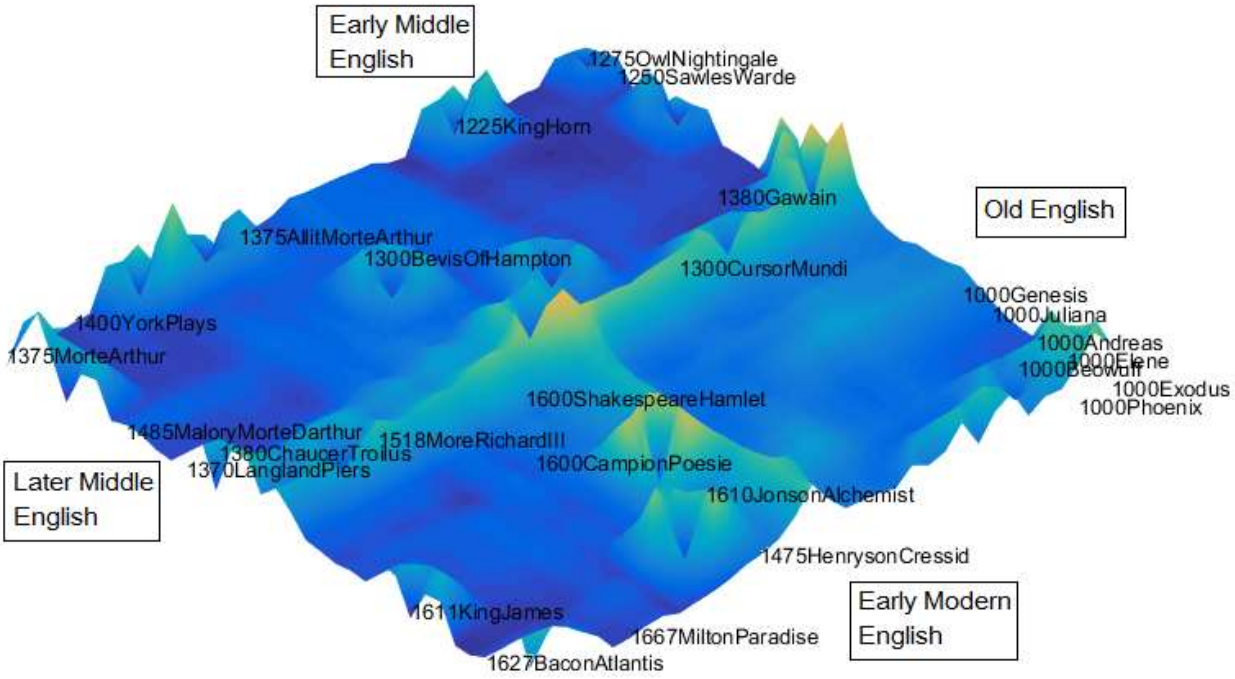


Figure 7: SOM architecture

An *n*-dimensional data vector is loaded into the input units. These values are propagated along the connections in such a way that the data vector is assigned to one of the cells on the two-dimensional projection surface; only a selection of connections is shown, and in reality every input unit is connected to every projection surface cell. Once every input vector has been projected onto the surface, the topology of the *n*-dimensional data manifold has been mapped onto the two-dimensional projection space, where it is available for visual inspection.

The variation in connection strengths is key to successful topological mapping, and these connections are learned from the data. Details of how this learning proceeds are fairly complex, and are not presented here.

3.4.4 *Result*
The result of the SOM projection is shown in Figure 8.



The labels are anchored on the left, that is, '1600ShakespeareHamlet' is, for example, located at the initial '1' on the map. As can be seen, the projection from the 100-dimensional data matrix onto a two-dimensional surface has clustered the texts in accordance with what is independently known of their dates.

Conclusion
Topological mapping is widely applicable to data abstracted from multi-text historical linguistic corpora:

- Where the characteristics of the corpus language are well known, as for the well-studied European languages, topological mapping can be used to bestow the fundamental scientific characteristics of objectivity and replicability on them.
- Where they are less well known, as for corpora in non-European languages, it can be used to identify objective, replicable geographical and relative chronological distributions.

References

Bertuglia, C., Vaio, F. (2005) *Nonlinearity, Chaos, and Complexity: The Dynamics of Natural and Social Systems*, Oxford University Press

Kohonen, T. (2001) *Self Organizing Maps,* 3rd ed., Springer

Lay, D. (2010) *Linear Algebra and its Applications*, 4th ed., Pearson

Lee, J. (2010) *Introduction to Topological Manifolds*, 2nd ed., Springer

Moisl, H. (2015) *Cluster Analysis for Corpus Linguistics*, de Gruyter

Strogatz, S. (2000) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*, Perseus Books.