Journal of Data Mining and Digital Humanities

How to visualize high-dimensional data: a roadmap

Hermann Moisl

Newcastle University, UK

hermann.moisl@ncl.ac.uk

Abstract

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 multivariate data provide a more complete description. Where the data are high-dimensional, however, their complexity can defy analysis using traditional philological methods. The first step in dealing with such data is to visualize it using graphical methods in order to identify any latent structure. If found, such structure facilitates formulation of hypotheses which can be tested using a range of mathematical and statistical methods. Where, however, the dimensionality is greater than 3, direct graphical investigation is impossible. The present discussion presents a roadmap of how this obstacle can be overcome, and is in three main parts: the first part presents some fundamental data concepts, the second describes an example corpus and a high-dimensional data set derived from it, and the third outlines two approaches to visualization of that data set: dimensionality reduction and cluster analysis.

keywords

Data visualization, multivariate data, high dimensionality, dimensionality reduction, cluster analysis.

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 provide a more complete description. Where the multivariate data are high-dimensional, however, their complexity can defy analysis using traditional philological methods. The first step in dealing with such data is to visualize it using graphical methods in order to identify any latent structure. If found, such structure facilitates formulation of hypotheses which can be tested using a range of mathematical and statistical methods. Where, however, the dimensionality is greater than 3, direct graphical investigation is impossible. The present discussion presents a roadmap of how this obstacle can be overcome. Exemplification is based on data abstracted from a corpus of English historical texts with a known temporal distribution, allowing the efficacy of the methods covered in the discussion to be readily verified by the reader. The discussion is in three main parts. The first part presents some fundamental data concepts - its nature, its representation using vectors and matrices, and its interpretation in terms of concepts of vector space and manifold, the second part describes the corpus and a high-dimensional data set abstracted from it, and the third outlines approaches to visualization of that data set using the concepts from (1) applied to (2). These approaches are of two types.

- The first, dimensionality reduction, reduces high-dimensional data to dimensionality 3 or less to enable graphical representation; the methods presented are (i) variable selection based on variance and (ii) principal component analysis.
- The second, cluster analysis, represents the structure of data in high-dimensional space directly without dimensionality reduction.

1. FUNDAMENTAL DATA CONCEPTS

1.1 Data

'Data' is the plural of 'datum', the past participle of Latin 'dare', to give, and means things that are given. A datum is therefore something to be accepted at face value, a true statement about the world. What is a true statement about the world? That question has been debated in philosophical metaphysics and epistemology since Antiquity and probably before, and, in our own time, has been intensively studied by the disciplines that comprise cognitive science ([Audi, 2010]). The issues are complex, controversy abounds, and the associated academic literatures are vast – saying what a true statement about the world might be is anything but straightforward. We can't go into all this, and so will adopt the attitude prevalent in most areas of science: data are abstractions of what we perceive using our senses, often with the aid of instruments.

Data are ontologically different from the world. The world is as it is; data are an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data – measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as lexical frequency are. Data are constructed from observation of things in the world, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the analytical results. The importance of understanding such data issues to visualization and numerical analysis can hardly be overstated ([Jain, 2010]). On the one hand, nothing can be discovered that is beyond the limits of what the data say about the world. On the other, failure to understand and where necessary to emend relevant characteristics of data can lead to results and interpretations that are distorted or even worthless; for general discussions of data and data transformation see [Tan et al., 2006]; [Izenman, 2008]; [Moisl, 2015].

Given that data are an interpretation of some domain of interest, what does such an interpretation look like? It is a description of objects in the domain in terms of variables. A variable is a symbol, that is, a physical entity to which a meaning is assigned by human interpreters. The variables chosen to describe a domain constitute the conceptual template in terms of which the domain is interpreted and on which the proposed analysis is based. If the analysis is to be valid with respect to the domain, therefore, it is crucial that the set of selected variables be adequate in relation to the research question, where adequacy is understood as follows:

- The variables should represent all and only those aspects of the domain which are relevant to the research question, that is, relevant aspects of the domain should not be unrepresented in the set of variables, and irrelevant aspects should not be represented. Failure to include relevant aspects in the data renders the description of the domain incomplete and thereby self-evidently compromises the validity of analysis based on it; inclusion of irrelevant aspects is less serious but introduces potentially confounding factors into an analysis.
- Each variable should be independent of all the others in terms of what it represents in the domain, that is, the variables should not overlap with one another in what they describe in the domain because such overlap describes the same thing multiple times and can thereby skew the analysis by overemphasizing the importance of some aspects of the domain over others.

In general, adequacy so defined cannot be guaranteed in any given research application because neither relevance nor independence is always obvious. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an adequate set of variables.

Once variables have been selected, a value is assigned to each of them for each of the objects of interest in the domain. This value assignment is what makes the link between the researcher's

conceptualization of the domain in terms of the variables s/he has chosen and the actual state of the world, and allows the resulting data to be taken as a valid representation of the domain based on empirical observation. The type of value assigned to any given variable depends on its meaning. The fundamental distinction of types is between quantitative, that is, numerical values, and qualitative ones such as binary 'yes / no' or categorical 'poor / adequate / good / excellent' ([Kaufman and Rousseeuw, 1990]; [Jain and Dubes, 1988]; [Jain et al., 1999]; [Gan et al., 2007]; [Xu and Wunsch 2009]). This discussion concentrates on quantitative variables because the majority of analytical applications are defined relative to them.

1.2 Data representation

A standard way of representing data for numerical analysis is via the mathematical concepts of vector and matrix ([Strang, 2016]).

- Vector

  If they are to be analyzed using computational methods, the descriptions of the entities in the domain of interest must be mathematically represented. A standard way of doing this is the vector. A vector is a sequence of n numbers each of which is indexed by its position in the sequence.

$$v = \underset{1}{\boxed{2.5}} \quad \underset{2}{\boxed{5.1}} \quad \underset{3}{\boxed{9.3}} \quad \underset{4}{\boxed{0.2}} \quad \underset{5}{\boxed{1.5}} \quad \underset{6}{\boxed{6.8}}$$

Figure 1. Vector

  Figure 1 shows $n = 6$ real-valued numbers, where the first number $v(1)$ is 2.1, the second $v(2)$ is 5.1, and so on.

- Matrix

  Given some number m of objects, each of which is described by an n-component vector, the set of vectors is organized as an m x n matrix.

Speaker	ɔ1	ɔ2	o:	ɔ3	ɪ	eɪ	n	a:1	a:2	aɪ	r	w
tlsg01	3	1	55	101	33	26	193	64	1	8	54	96
tlsg02	8	0	11	82	31	44	205	54	64	8	83	88
tlsg03	4	1	52	109	38	25	193	60	15	3	59	101
tlsn01	100	116	5	17	75	0	179	64	0	19	46	62
tlsg04	15	0	12	75	21	23	186	57	6	12	32	97
tlsg05	14	6	45	70	49	0	188	40	0	45	72	79
tlsg06	5	0	40	70	32	22	183	46	0	2	37	117
tlsn02	103	93	7	5	87	27	241	52	0	1	19	72
tlsg07	5	0	11	58	44	31	195	87	12	4	28	93
tlsg08	3	0	44	63	31	44	140	47	0	5	43	106
tlsg09	5	0	30	103	68	10	177	35	0	33	52	96
tlsg10	6	0	89	61	20	33	177	37	0	4	63	97
tlsn03	142	107	2	15	94	0	234	15	0	25	28	118
tlsn04	110	120	0	21	100	0	237	4	0	61	21	62
tlsg11	3	0	61	55	27	19	205	88	0	4	47	94
tlsg12	2	0	9	42	43	41	213	39	31	5	68	124
tlsg52	11	1	29	75	34	22	206	46	0	29	34	93
tlsg53	6	0	49	66	41	32	177	52	9	1	68	74
tlsn05	145	102	4	6	100	0	208	51	0	22	61	104
tlsn06	109	107	0	7	111	0	220	38	0	26	19	70
tlsg54	3	0	8	81	22	27	239	30	32	8	80	116
tlsg55	7	0	12	57	37	20	187	77	41	4	58	101
tlsg56	12	0	21	59	31	40	164	52	17	6	45	103
tlsn07	104	93	0	11	108	0	194	5	0	66	33	69

Table 1. Matrix

  Table 1 shows a data matrix in which each of $m = 24$ rows, labelled tlsg 01 - tlsn07, is a speaker in a sociolinguistic study, and each speaker is described by his or her frequency of

usage of $n = 12$ phonetic segments which label the columns.

1.3 Data interpretation

- Vector spaces

  In colloquial usage, the word 'space' denotes a fundamental aspect of how humans understand their world: that we live our lives in a three-dimensional space, that there are directions in that space, that distances along those directions can be measured, that relative distances between and among objects in the space can be compared, that objects in the space themselves have size and shape which can be measured and described. The earliest geometries were attempts to define these intuitive notions of space, direction, distance, size, and shape in terms of abstract principles which could, on the one hand, be applied to scientific understanding of physical reality, and on the other to practical problems like construction and navigation. Basing their ideas on the first attempts in ancient Mesopotamia and Egypt, Greek philosophers from the sixth century BCE onwards developed such abstract principles systematically, and their work culminated in the geometrical system attributed to Euclid (*floruit ca*. 300 BCE), which remained the standard for more than two millennia thereafter ([Tabak, 2011]).

  In the nineteenth century CE the validity of Euclidean geometry was questioned for the first time both intrinsically and as a description of physical reality. It was realized that the Euclidean was not the only possible geometry, and alternative ones were proposed in which, for example, there are no parallel lines and the angles inside a triangle always sum to less than 180 degrees. Since the nineteenth century these alternative geometries have continued to be developed without reference to their utility as descriptions of physical reality, and as part of this development 'space' has come to have an entirely abstract meaning which has nothing obvious to do with the one rooted in our intuitions about physical reality. A space under this construal is a mathematical set on which one or more mathematical structures are defined, and is thus a mathematical object rather than a humanly-perceived physical phenomenon ([Lee, 2010]). The present discussion uses 'space' in the abstract sense; the physical meaning is often useful as a metaphor for conceptualizing the abstract one, though it can easily lead one astray.

  Vectors have a geometrical interpretation ([Strang, 2016]).

  - The dimensionality n of a vector defines an abstract vector space: two-component vectors define a two-dimensional space, three-component vectors a three-dimensional space, and so on.
  - The values in the vector components are coordinates in the space.
  - The vector itself is a point at those coordinates.

  This is exemplified in Figure 2.

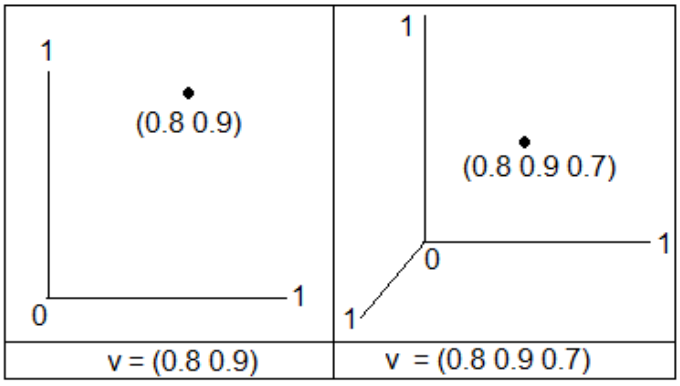

Figure 2. Two and three dimensional vector spaces

- Manifolds

  It is possible to have more than one vector in a vector space. For a matrix with m rows, there will be m points in the space, and those m points define a shape in the space. That shape is a manifold. For example, Figure 3 shows a manifold in 2-dimensional space:

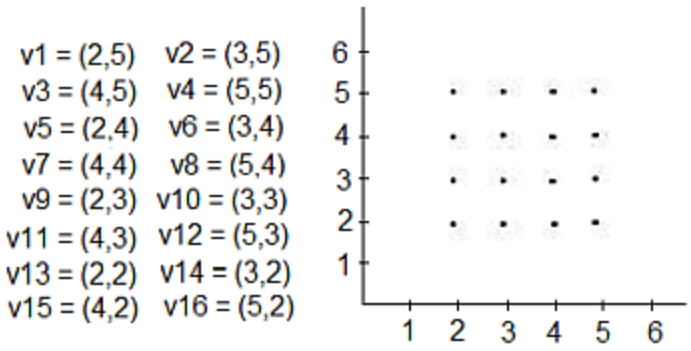

Figure 3. Data manifold in 2-dimensional space

The shape in the case of Figure 3 is a plane, but many other shapes are possible; an example is given in Figure 4

	v(1)	v(2)	v(3)
1	-7.82	11.63	4.99
2	-6.41	-3.83	-2.88
3	15.52	12.96	27.82
...	...	...	...
2000	5.85	12.38	12.19

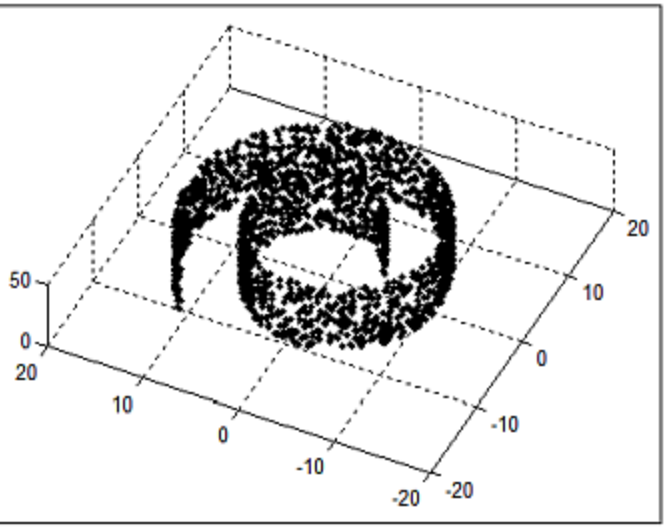

Figure 4. Nonlinear manifold in 3-dimensional space

Figures 3 and 4 exemplify a fundamental distinction between two types of manifold shape, linear and nonlinear, which reflects a corresponding distinction in the characteristics of natural processes and the data that describe them. As their names indicate, linear manifolds are straight lines and flat planes, and nonlinear ones are curved lines and surfaces. In the nonlinear case the complexity of curvature can range from simple curved lines and planes to highly convoluted fractal shapes.

The concepts of vector, matrix, vector space, and manifold are standardly used across the sciences to represent and interpret data. For a domain of interest containing *m* objects such as historical texts, each vector represents a single text. Each text is described in terms of one or more variables, and every variable is represented by a different vector component. The collection of *m n*-dimensional text vectors constitutes the data. Geometrically, those data exist in an *n*-dimensional vector space and constitute a data manifold. The manifold is the shape of the data.

2. CORPUS AND DATA

The corpus used for exemplification here is a collection of Old English, Middle English, and Early Modern English texts, listed in Table 2. These texts were chosen on account of their familiarity within the English historical linguistics research community, which will facilitate interpretation of the results presented in what follows.

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, Richard III
Genesis A	Morte Arthure	Shakespeare, Hamlet
Beowulf	King Horn	Jonson, Alchemist
	Alliterative Morte Arthure	
	Bevis of Hampton	
	Chaucer, Troilus	
	Langland, Piere Plowman	
	York Plays	
	Cursor Mundi	

Table 2. Corpus.

All the above texts were downloaded from online digital collections:

- Old English: Internet Sacred Text Archive (https://www.sacred-texts.com/neu/ascp/).
- Middle English: Corpus of Middle English Prose and Verse (https://quod.lib.umich.edu/c/cme/).
- Early Modern English: Early English Books Online (https://quod.lib.umich.edu/e/eebogroup/).

No attempt was made at textual optimality via currently definitive editions of the various texts because the aim of this discussion is methodological, as noted in the Introduction: to exemplify the application of visualization methods rather than to elicit philologically reliable results.

Spelling is the basis of data abstraction from the texts. Specifically, the variables used to represent spelling 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 and assembled in a matrix, henceforth M. A fragment of M is shown in Table 3, where the rows are the texts and the columns the letter-pair variables. Note that it is very high-dimensional, and cannot be graphically represented directly.

	1. <[space]h>	2. <hw>	3. <wæ>	...	841. <jm>
Exodus	338	35	94	...	0
Sawles Warde	629	52	0	...	0
...	...	...	...	...	...
Cursor Mundi	15219	20	0	...	0

Table 3. Fragment of frequency matrix M abstracted from the corpus.

Before applying visualization methods to M, a problem that often arises when working with multi-text corpora must be dealt with: variation in text length. Why this can be a problem is easy enough to see. If the data are based on frequency of some textual feature such as words or, as here, letter pairs,

then for whatever feature one is counting it is in general probable that a relatively longer text will contain more instances of it than a shorter one - a novel inevitably contains more instances of 'the' than a typical email, for example. If frequency vector profiles for varying-length texts are constructed, then the profiles for longer texts can generally be expected to have larger frequency values than the profiles for shorter ones, and when visualization methods are applied to a collection of such vectors the results will be distorted by these disparities of magnitude. The texts listed in Table 2 vary substantially in length, and so our data matrix M poses this problem.

The solution is to transform or 'normalize' the values in the data matrix so as to mitigate or eliminate the effect of text length variation. Various normalization methods exist; for discussion and references see [Moisl 2015, Ch.3]. The one used here on account of its intuitive simplicity is normalization by mean document length. Restricting the discussion to M, this involves transformation of its row vectors in relation to the mean number of pairs per text across all texts in the corpus, as in Equation1.

$$M_i = M_i \left(\frac{\mu}{\text{NrPairs}(T_i)}\right) \quad (1)$$

Equation 1. Mean document length normalization

M_i is the *i*'th matrix row (for i = 1...the number of texts / number of matrix rows), T_i is the *i*'th text, NrPairs(T_i) is the number of letter pairs in T_i, and μ is the mean number of letter pairs per text across all the texts in the corpus, which is calculated as in Equation 2.

$$\mu = \sum_{i=1..m} \frac{\text{NrPairs}(T_i)}{m} \quad (2)$$

Equation 2. Mean number of letter pairs per text across all texts

Here m is the number of texts in the corpus / number of matrix rows and the Σ term with subscript stands for summation over all m rows.

In words, first calculate the mean number of letter pairs per text across all texts in the corpus by adding the number of pairs in each of the texts and then dividing by the number of texts. This quantity μ is then used to transform the matrix so as to compensate for the variation in text length: for each row M_i multiply every value in M_i by the ratio of μ to the total number of pairs in text T_i: if T_i is larger than μ then multiplication by the ratio will decrease the values in M_i in proportion to how much larger T_i is than μ; if T_i is smaller than μ then multiplication by the ratio will increase the values in M_i in proportion to how much smaller T_i is than μ; and if T_i and μ are equal then the ratio is 1 and there will be no change.

M was normalized as above, and the fragment of the normalized version M' shown in Table 4 is the one to which the visualization methods described in the next section are applied. The numerical quantities are now real-valued and no longer represent observed frequencies on account of the normalization procedure; note how the large observed disparities between the relatively short *Exodus* and the relatively long *Cursor Mundi* in column 1 of M have been brought much closer together in M' by the procedure.

	1. [space]h	2. hw	3. wæ	...	841. jm
Exodus	2728.55	282.54	758.82	...	0
Sawles Warde	3824.59	316.18	0	...	0
...	...	...	...	...	...
Cursor Mundi	3005.92	3.95	0	...	0

Table 4. Fragment of M' derived from M by mean document length normalization

III. VISUALIZATION OF HIGH-DIMENSIONAL DATA

3.1 Dimensionality reduction

One approach to visualization of high-dimensional data is to reduce the dimensionality to 3 or less so that the distribution of data objects can be graphically displayed directly by plotting. There are various ways to do this. What follows describes two of the most intuitively accessible.

3.1.1 Variable selection

The simplest way of reducing dimensionality is to select the two or three most important variables in the data, discarding the rest. This involves throwing away the information contained in the discarded variables, which may or may not be justifiable since it may be that the two or three retained variables are not sufficient adequately to describe the domain of interest.

To select the most important variables, a criterion of importance is required. A useful criterion is variability. Consider two limiting cases: (i) all the objects in the domain of interest are the same, and (ii) all the objects in the domain of interest are completely different. In neither case is the domain structured, and there is nothing further to say about it. There is interesting structure when there is some sort of pattern of similarities and differences between and among its constituent objects. The limiting cases rarely obtain in real-world domains, that is, there is usually some structure, and the similarities and differences are reflected in the variables that describe the domain. Variables whose values show little or no variability do little or nothing to distinguish similarities and differences among objects. Variables with large variability do. The most important variables are therefore those with the greatest variability. So:

- Calculate the variance of each column in the data matrix
- Sort the columns on decreasing order of variance magnitude.
- Select the two or three columns with the largest variances.
- Plot them.

Applying this to M', Figure 5 is a plot of sorted variances.

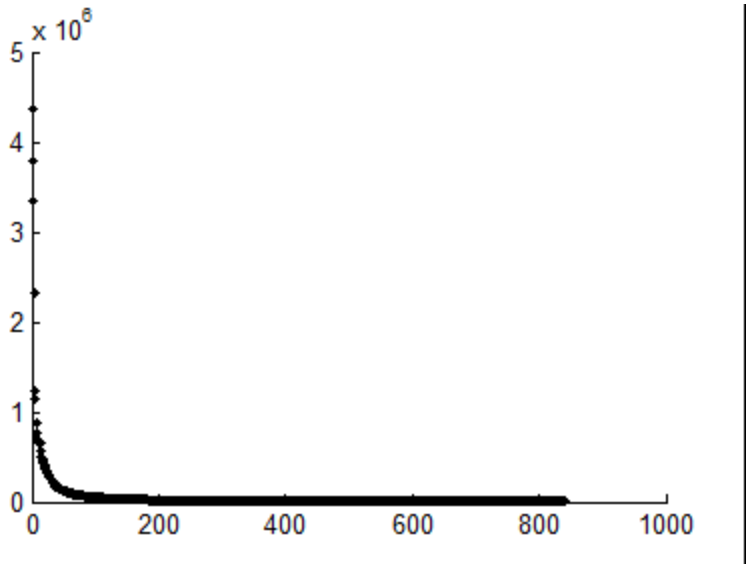

Figure 5. Variances of the columns of M', sorted in descending order of magnitude

Most of the variance is concentrated in the relatively few variables on the left. For a clearer view, a plot of the column variances of the 100 highest-variance columns of M' in descending order is given in Figure 6.

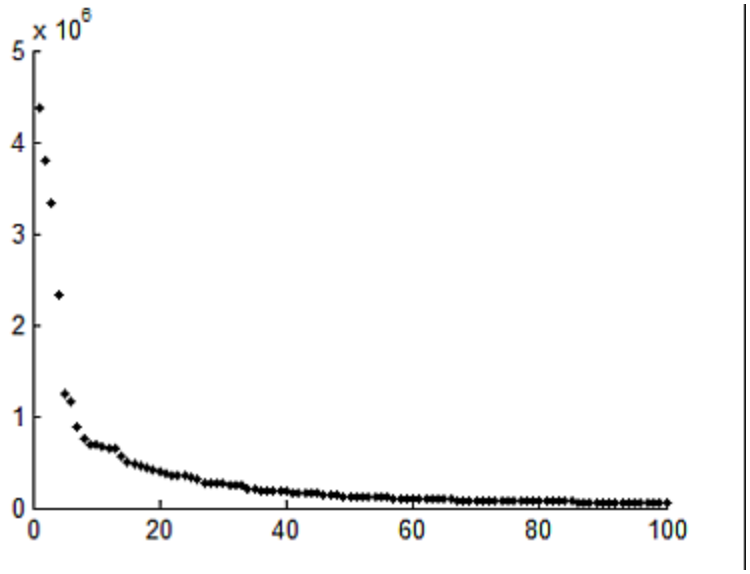

Figure 6. 100 highest-variance columns of M' in descending order of magnitude

	1. th	2. e[space]	3. [space]t	...	841. jm
Exodus	8.07	5142.27	339.05	...	0
Sawles Warde	0	8238.99	1143.12	...	0
...	...	...	...	...	...
Cursor Mundi	770.09	6210.94	1751.33	...	0
Variance	4367519.5	3793490.25	3335804.5		0

Table 5. Fragment of M' with columns sorted in decreasing order of variance

A fragment of M' with its columns sorted by variance in Table 5 shows the highest-variance variables.

The two highest-variance variables are selected for scatter plotting. That plot is based on 21% of the variability in the data; this percentage is calculated as follows:

1. Calculate the total variance in the data by adding the variances of all 841 columns of M.
2. Calculate the variance of the selected variables by adding them.
3. (2-variable variance / total data variance) x 100 ≈ 21%

Using just two selected variables representing only 21% of the variability in the data, scatter-plot them, as in Figure 7:

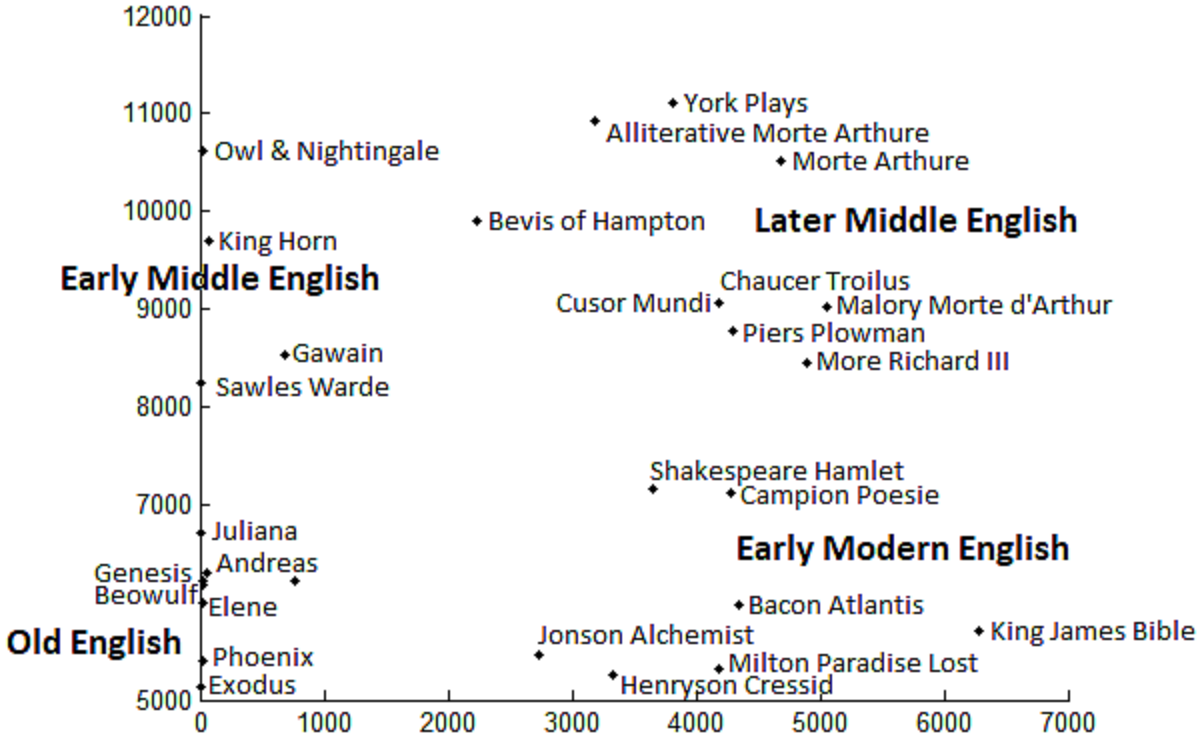

Figure 7. Plot of the two highest-variance variables in M', with corresponding text labels

Relative spatial distances among the data points in the plot reflect the relative similarities of the two-dimensional vectors abstracted from M' to represent the texts. Four distinct clusters of points are visually identifiable, and these correspond to four conventional periods in the development of English, as labelled; dimensionality reduction via selection of the highest-variance variables has yielded a good indication of the known chronological structure of the example corpus. There is, however, no guarantee that so low a percentage of the total data variance will always give accurate results. In the present case one could extend to a 3-dimensional scatter plot, based on a higher percentage of the variance, but with that the scope of variable selection is exhausted. What's required is a way of using more of the data variability while retaining the graphability of two or three dimensions. A method for doing this follows.

3.1.2 Variable redefinition

The foregoing discussion of data creation noted that, because selection of variables is at the discretion of the researcher, it is possible that the selection in any given application will be suboptimal in the sense that there is redundancy among them, that is, that they overlap with one another to greater or lesser degrees in terms of what they represent in the research domain. Where there is such redundancy, dimensionality reduction can be achieved by eliminating the repetition of information which redundancy implies, and more specifically by replacing the researcher-selected variables with a

smaller set of new, non-redundant variables. Slightly more formally, given an *n*-dimensional data matrix, dimensionality reduction by variable redefinition assumes that the data can be described, with tolerable loss of information, by a manifold in a vector space whose dimensionality is lower than that of the data, and proposes ways of identifying that manifold.

For example, data for a study of student performance at university might include variables like personality type, degree of motivation, score on intelligence tests, scholastic record, family background, class, ethnicity, age, and health. For some of these there is self-evident redundancy: between personality type and motivation, say, or between scholastic record and family background, where support for learning at home is reflected in performance in school. For others the redundancy is less obvious or controversial, as between class, ethnicity, and score on intelligence tests. The researcher-defined variables personality type, motivation, scholastic record, and score on intelligence tests might be replaced by a 'general intelligence' variable based on similarity of variability among these variables in the data, and family background, class, ethnicity, age, and health with a 'social profile' one, thereby reducing data dimensionality from nine to two.

Clearly, if there is little or no redundancy in the user-defined variables then there is little or no point to variable redefinition. The first step must, therefore, be to determine the level of redundancy in the data of interest to see whether variable extraction is worth undertaking; for details on how to do this see [Moisl, 2015, ch. 3].

The standard method for variable redefinition is principal component analysis (PCA). The global variance of a data matrix is the sum of the variances of all the variable columns. Dimensionality reduction using PCA is based on the idea that most of the global variance of data with *n* variables can be expressed by a smaller number $k < n$ of newly-defined variables. These *k* new variables are found using the shape of the manifold in the original *n*-dimensional space. The mathematics of how PCA works are too complex for detailed exposition here; for the technicalities see the standard accounts by [Jolliffe, 2002] and [Jackson, 2003] and, more briefly, [Bishop, 1995, 310 ff.], [Everitt and Dunn, 2001, 48 ff.], [Tabachnick and Fidell, 2007, Ch.13], [Izenman, 2008, Ch. 7], [Moisl, 2015, ch. 3]. Instead, an intuitive graphical explanation is given.

Figure 8a shows a manifold of data points plotted in a 2-dimensional space defined by the researcher-specified variables *v*1 and *v*2.

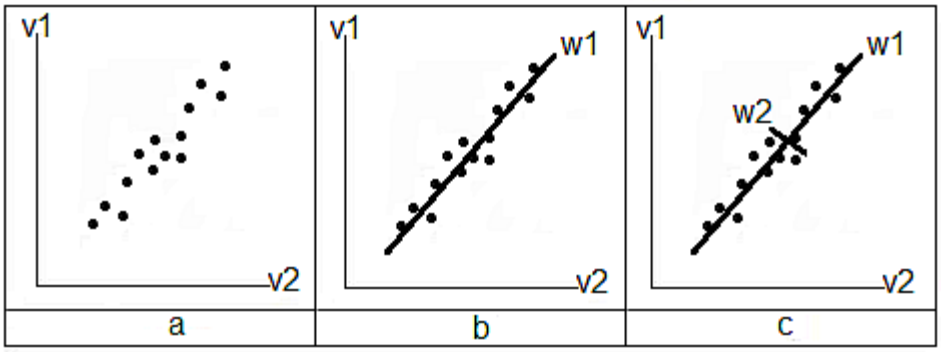

Figure 8. Manifold and two principal components

The main direction of variability in 8a can be visually identified; the line of best fit drawn through the manifold in that direction, as in 8b, is the first new variable *w*1: it captures most of the variance in the manifold, and its length is the amount of variance that it represents. A second line of best fit is now drawn at right angles to the first, as in 8c, to capture the remaining variance. This is the second new variable *w*2, and its length is again the amount of variance it represents. We now have two new variables in addition to the two original ones. What about dimensionality reduction? Note the

disparity in the lengths of $w1$ and $w2$. It's clear that $w1$ captures almost all the variance in the manifold and $w2$ very little, and one might conclude that $w2$ can simply be omitted with minimal loss of information. Doing so reduces the dimensionality of the original data from 2 to 1.

This idea extends to any dimensionality. Using it, PCA is a general method for dimensionality reduction of n-dimensional data, where n is any integer. The first variable represents the largest direction of variability in the data manifold, the second variable represents the second-largest direction of variability in the manifold, the third represents the third-largest direction, and so on; the first two or three variables can be visualized by scatter plotting, and that visualization will represent the most important directions of variability in the original data.

PCA was applied to the full 841-dimensional data matrix M', yielding a new matrix M'' containing 841 new variables. The variances of these new variables are plotted in descending order of magnitude in Figure 9.

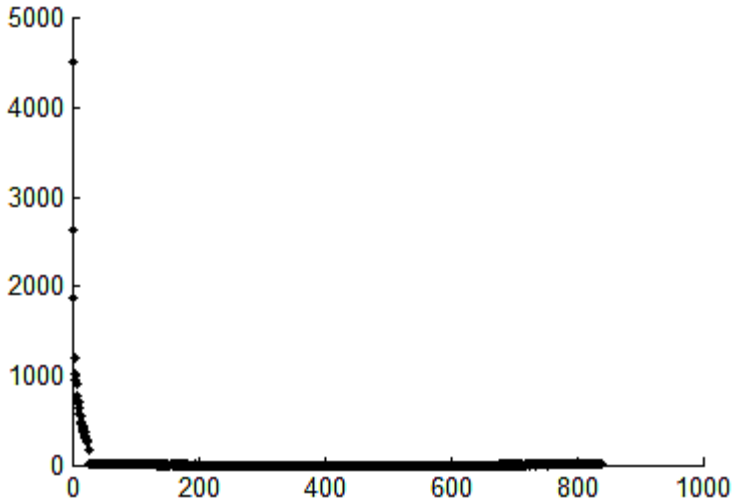

Figure 9. Variances of the 841 PCA-transformed variables in descending order of magnitude

To see this distribution more clearly, the 12 highest-variance variables are shown in Figure 10.

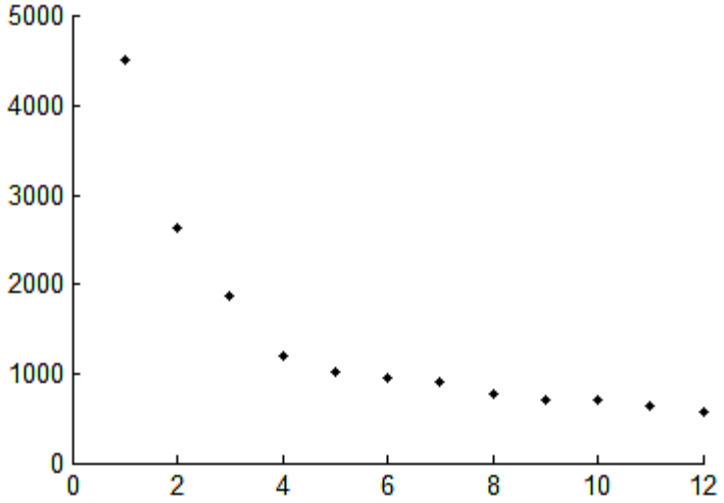

Figure 10. The 12 highest-variance variables in M'' in descending order of magnitude

Almost all the variability is in the first 4 variables. In fact, the first two variables contain 69% of the variability, as compared to 21% using variable selection. Figure 11 is a scatter plot of these two, with associated text names.

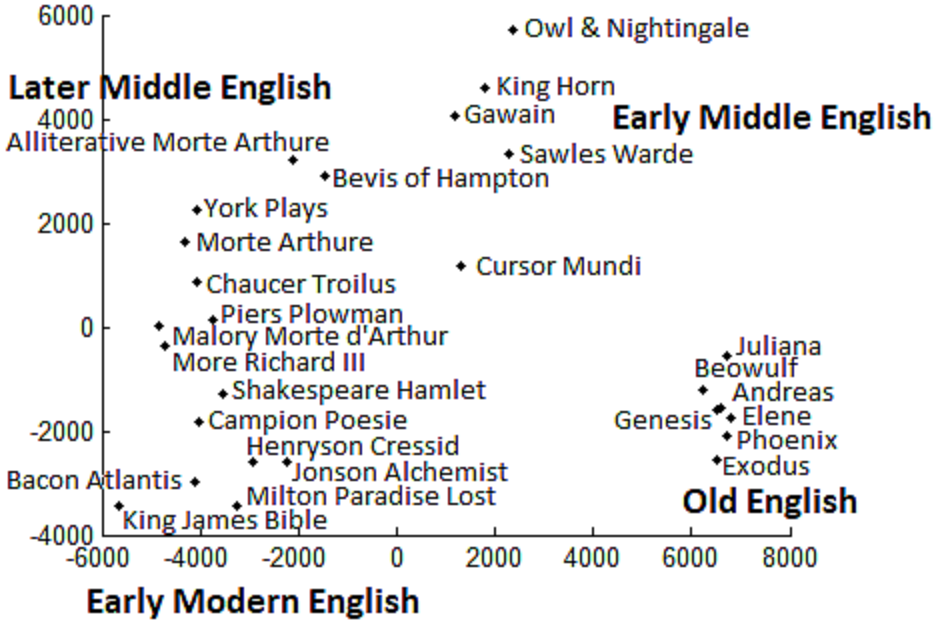

Figure 11: Scatter plot of the two highest-variance variables derived via PCA, with associated text name.

As with variable selection, the spatial distribution of the texts is here consistent with their independently-known chronological structure, but, because it is based on PCA-transformed data which has preserved a substantially greater amount of the variance in the original data, its reliability is enhanced.

3.2 Cluster analysis

The alternative to dimensionality reduction of high-dimensional data for visualization is cluster analysis. The following account is in three main parts: the first attempts to define the notion of 'cluster', the second briefly surveys the varieties of cluster analysis, and the third describes the most frequently used variety. The clustering literature is extensive. The more important references include [Jain and Dubes, 1988], [Kaufman and Rousseeuw, 1990], [Arabie and Hubert, 1996], [Gordon, 1999], [Gan et al., 2007], [Xu and Wunsch, 2009], [Everitt et al., 2011], [Mirkin, 2013]; for further references see [Moisl, 2015, Ch. 4].

3.2.1 Cluster definition

In cluster analytical terms, identification and visualization of structure in data is identification of clusters. To undertake such identification it is necessary to have a clear idea of what a cluster is, and this is provided by an innate human cognitive capability. Human perception is optimized to detect patterning in the environment ([Köppen, 2000]; [Peissig and Tarr, 2007]), and clusters are a kind of pattern. Contemplation of a rural scene, for example, reveals clusters of trees, of farm buildings, of sheep. Looking up at the night sky reveals clusters of stars. And, closer to present concerns, anyone looking at the data plot in Figure 12 immediately sees the clusters. A casual observer looking at the scatterplots would say that 12a shows a few small concentrations of points but is essentially random, that 12b has three clearly identifiable clusters of roughly equal size, that 12c has two clusters of unequal, size the smaller of which is in the lower-left corner of the plot and the larger elongated one in the upper right, and that 12d has two intertwined, roughly semi-circular clusters, all embedded in a random scatter of points. That casual observer would, moreover, have been able to make these identifications solely on the basis of innate pattern recognition capability and without recourse to any explicit definition of the concept ‘cluster’.

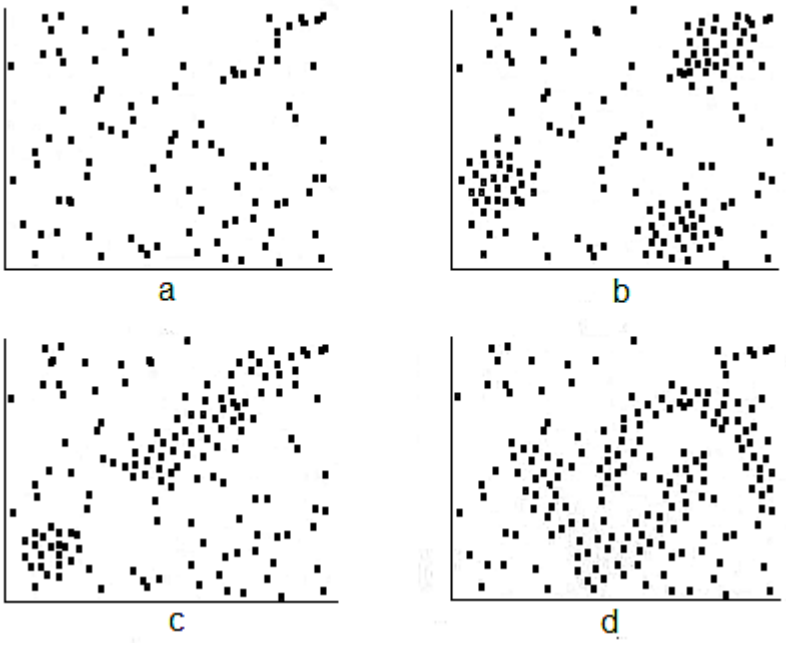

Figure 12. A selection of clusters in 2-dimensional space

Direct perception of pattern is the intuitive basis for understanding what a cluster is, and is fundamental in identifying the cluster structure of data, but it has two main limitations. One limitation is subjectivity and consequent unreliability. Apart from the obvious effect of perceptual malfunction in the observer, this subjectivity stems from the cognitive context in which a given data distribution is interpreted: the casual observer brings nothing to the observation but innate capability, whereas the researcher who compiled the data and knows what the distribution represents brings prior knowledge which potentially and perhaps inevitably affects interpretation. In Figure 12c, for example, does the larger cluster on the upper right contain two subclusters or not?

What would the answer be if it were known that the points represent cats in the upper part of the cluster and dogs in the lower? The other limitation is that reliance on innate perceptual capability for cluster identification is confined to what can be perceived, and in the case of data this means dimensionality of 3 or less for graphical representation; there is no way of perceiving clusters in data with dimensionality higher than that directly. The obvious way to address these limitations is by formal and unambiguous definition of what a cluster is, relative to which criteria for cluster membership can be stated and used to test perceptually-based intuition on the one hand and to identify non-visualizable clusters in higher-dimensional data spaces on the other. Textbook and tutorial discussions of cluster analysis uniformly agree, however, that it is difficult and perhaps impossible to give such a definition, and, if it is possible, that no one has thus far succeeded in formulating it. In principle, this lack deprives cluster analysis of a secure theoretical foundation. In practice, the consensus is that there are intuitions which, when implemented in clustering methods, give conceptually useful results, and it is on these intuitions and implementations that contemporary cluster analysis is built. The fundamental intuition underlying cluster analysis is that data distributions contain clusters when the data objects can be partitioned into groups on the basis of their relative similarity such that the objects in any group are more similar to one another than they are to objects in other groups, given some definition of similarity.

3.2.2 Varieties of cluster analysis

Given an *m* x *n* data matrix D, cluster analysis works by partitioning the *m* row vectors into disjoint subsets in accordance with their relative similarity in *n*-dimensional space. Numerous methods are available, and they are standardly divided into two categories in accordance with the kind of output

they generate: hierarchical and nonhierarchical. Nonhierarchical methods partition the *m* row vectors of D into a set of clusters C = *c*(1), *c*(2)...*c*(*k*) such that the members of cluster *c*(*i*), for *i* = 1...*k*, are more similar to one another than they are to any member of any other cluster, on the basis of some definition of similarity. Hierarchical methods regard the *m* row vectors of D as a single cluster C and recursively divide each cluster into two subclusters each of whose members are more similar to one another than they are to members of the other on the basis, again, of some definition of similarity, until no further subdivision is possible: at the first step C is divided into subclusters *c*1 and *c*2, at the second step *c*1 is divided into two subclusters *c*1.1, *c*1.2, and *c*2 into *c*2.1, *c*2.2, at the third step each of *c*1.1, *c*1.2, *c*2.1, *c*2.2 is again subdivided, and so on. The succession of subdivisions can be and typically is represented as a binary tree, and this gives the hierarchical methods their name. Both hierarchical and nonhierarchical methods partition the data; the difference is that the nonhierarchical ones give only a single partition into *k* clusters, where *k* is either pre-specified by the user or inferred from the data by the method, whereas the hierarchical ones offer a succession of possible partitions and leave it to the user to select one of them. These two categories therefore offer complementary information about the cluster structure of data.

3.2.3 Hierarchical cluster analysis

Hierarchical analysis is the most frequently used variety, and is described in what follows. Construction of a hierarchical cluster tree is a two-step process: given an *m* x *n* data matrix, the first step abstracts a table of proximities of each possible pairing of the *m* rows in the *n*-dimensional space, and the second then constructs the tree by successive transformations of the proximity table.

Numerous distance metrics exist ([Deza and Deza, 2009, chs. 17, 19]). For present purposes these are divided into two types: (i) linear metrics, where the distance between two points in a manifold is taken to be the length of the straight line joining the points, or some approximation to it, without reference to the shape of the manifold, and (ii) nonlinear metrics, where the distance between the two points is the length of the shortest line joining them along the surface of the manifold and where this line can but need not be straight. This categorization is motivated by the earlier observation that manifolds can have shapes which range from perfectly flat to various degrees of curvature. Where the manifold is flat, as in Figure 13a, linear and nonlinear measures are identical. Where it is curved, however, linear and nonlinear measurements can differ to varying degrees depending on the nature of the curvature, as shown in Figures 13b and 13c.

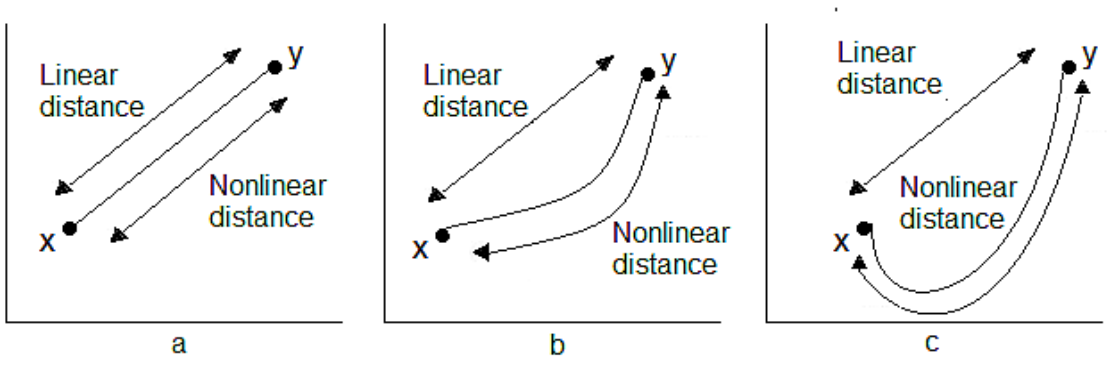

Figure 13. Linear and nonlinear distances on flat and curved manifolds

The usual metric used for hierarchical cluster analysis is Euclidean distance, which is linear, but where a data manifold is known to be curved, more accurate results can be expected from a nonlinear metric. Euclidean distance is the shortest distance between two points *i* and *j* calculated by the formula in Equation 3.

$$d_{i,j} = \sqrt[2]{\sum_{k=1...n} (M_{i,k} - M_{j,k})^2} \quad (3)$$

Equation 3. Euclidean distance

This is just the Pythagorean rule known, one hopes, to all schoolchildren, that the length of the hypotenuse of a right-angled triangle is the square root of the sum of the squares of the lengths of the other two sides. For n = 2, M(i) = (6,8) and M(j) = (2,4), the Euclidean distance d(M(i), M(j)) is calculated as in Figure 14, and it is the shortest distance between the two points.

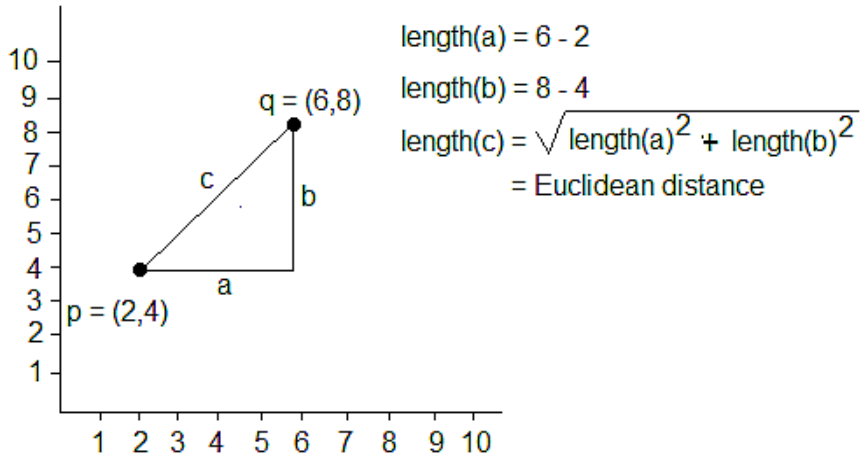

Figure 14. Calculation of Euclidean distance

A Euclidean distance matrix is abstracted from an m x n matrix by applying Formula 1 to each possible pairing of the m rows. Applied to M' in Table 4, the formula generates the distance matrix D in Table 6.

	Exodus	**Sawles Warde**	**...**	**Cursor Mundi**
Exodus	0	9924.9	...	11400
Sawles Warde	9924.9	0	...	9778.5
...	...	...	...	...
Cursor Mundi	11400	9778.5	...	0

Table 6. Fragment of proximity matrix D abstracted from M'

The distance from *Exodus* to *Sawles Warde* is 9924.9, from *Exodus* to *Cursor Mundi* is 11400, from *Sawles Warde* to *Cursor Mundi* 9778.5, and so on for the other texts not shown. Note that the distance from a text to itself is 0, which accounts for the zero-diagonal, and that the matrix is symmetrical about that diagonal, since the distance from any text A to any text B is the same as the distance from B to A. The absolute values of the distances do not matter. What matters is their relative magnitude: in terms of spelling, one expects the distance from the Old English *Exodus* to the early Middle English *Sawles Warde,* for example, to be smaller than to the later Middle English *Cursor Mundi*.

The proximity matrix is now used to construct a cluster tree that describes the relative distances of all the row vectors / data objects from one another. The construction algorithm is too complicated to describe here; for details see [Moisl, 2015, ch. 4]. The result for the proximity matrix in Table 6 is

shown in Figure 15.

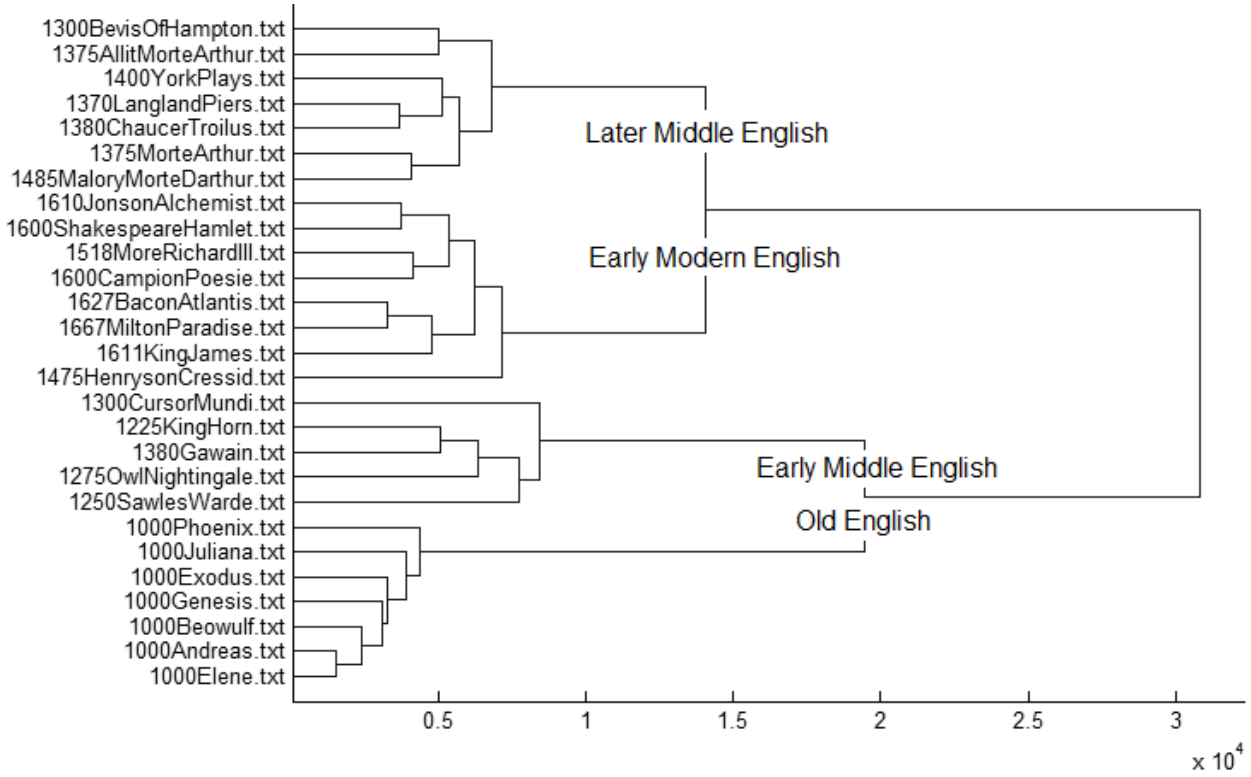

Figure 15. Hierarchical analysis of M' based on proximity matrix D in Table 6

Interpretation of the tree is based on the total distance in D between any two texts, as represented graphically by the length of the path joining them: the path joining *Andreas* and *Elene* is relatively short because they are both Old English texts and therefore very similar in terms of spelling; the path joining *Elene* and *Cursor Mundi* is much longer, reflecting substantial disparity in chronologically-conditioned spelling; the path joining *Elene* and *Hamlet* is longer still. As with the foregoing methods, the cluster tree gives a visual representation consistent with what is independently known of the chronological structure of the corpus.

The main and considerable advantage of hierarchical clustering is that it provides an exhaustive and intuitively accessible description of the proximity relations among data objects, and thereby provides more information than a simple partitioning of the data generated by the non-hierarchical methods covered thus far. It has also been extensively and successfully used in numerous applications, and is widely available in software implementations. There are, however, associated problems. Two of the most important are:

- Tree selection: There are various definitions of cluster membership which, in the literature, go by names like 'single linkage', 'complete linkage', 'average linkage', and 'Ward's method', among others. When used for tree construction, these definitions can and often do generate different structures with respect to the same data. Where this is found to be the case, the obvious questions are: which tree should be preferred, and why? The traditional answer is that an expert in the domain from which the data was taken should select the analysis which seems most reasonable in terms of what s/he knows about the research area. The problem with this is that it is subjective. It runs the risk of reinforcing preconceptions and discounting the unexpected and potentially productive insights which are the prime motivation for use of cluster analysis in hypothesis generation. The alternative is to resort to cluster validation methods, for details of which see [Moisl, 2015, Ch. 4].
- How many clusters? Given that a hierarchical cluster tree provides an exhaustive description of the proximity relations among data objects, how many clusters do the data 'really' contain?

As already noted, it is up to the user to decide. Looking at the tree in Figure 15 the answer seems obvious: there are two main clusters, one containing the Old and Early Middle English texts, and the other the Later Middle English and Early Modern English ones, with further subdivision in each. What about a structure like the one in Figure 16, however?

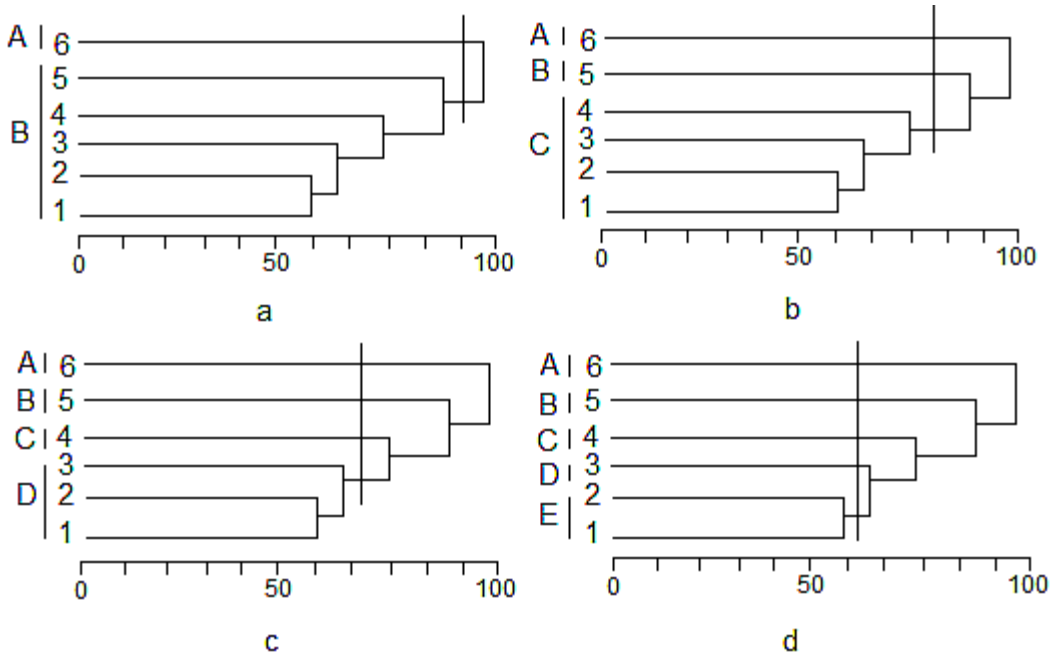

Figure 16. Different 'cuts' of the same tree

- There are no obvious main clusters, and depending on where one 'cuts' the tree, two, three, four or more clusters can be identified. In 16a the cut is placed so that subclusters below a threshold of 90 are not distinguished, yielding two clusters. In 16b the cut is at about 80, yielding three clusters, in 16c there are four clusters for a threshold of about 70, and in 16d there are five. Which is the best cut, that is, the one that best captures the cluster structure of the data? There have been attempts to formalize selection of a best cut, but the results have been mixed, and the current position is that the best cut is the one that makes most sense to experts in the subject from which the data comes.

Conclusion

The foregoing discussion has presented three methods for visualization of high-dimensional textual data as a basis for hypothesis generation. Many other dimensionality reduction and clustering methods are currently available, details of which can be found in the list of references that follows. An important topic that was briefly mentioned in the course of discussion but whose implications have hardly been addressed in corpus-based linguistics is the possibility that the data manifold is nonlinear.

References

Arabie P. and Hubert L. An overview of combinatorial data analysis. In: *Clustering and Classification*, ed. P. Arabie, L. Hubert, and G. De Soete. World Scientific (Singapore), 1996: 5–63.

Audi R. Epistemology: *A contemporary introduction to the theory of knowledge*. Routledge (London), 2010.

Bishop C. *Neural networks for pattern recognition*. Clarendon Press (Oxford), 1995.

Deza M. and Deza, E. (2009) *Encyclopedia of distances*. Springer (Berlin), 2009.

Everitt B. and Dunn G. *Applied multivariate data analysis*, 2nd ed. Arnold (London), 2001.

Everitt B., Landau S., Leese M., and Stahl D.(2011): *Cluster analysis*, 5th ed. Wiley (Hoboken NJ), 2011.

Gan G., Ma C., and Wu J. *Data clustering. Theory, algorithms, and applications*. American Statistical Association (Alexandria VA), 2007.

Gordon A. *Classification*, 2nd ed. Chapman and Hall (London), 1999.

Izenman A. *Modern multivariate statistical techniques. Regression, classification, and manifold learning.* Springer (Berlin), 2008.

Jackson J. *A user's guide to principal components*. Wiley-Interscience (Hoboken NJ), 2003.
Jain A. and Dubes R. *Algorithms for clustering data*. Prentice Hall (London), 1988.
Jain A., Murty M. and Flynn P. Data clustering: a review. *ACM Computing Surveys*. 1999, 31: 264–323.
Jolliffe I. *Principal component analysis*, 2nd ed. Springer (Berlin), 2002.
Kaufman L. and Rousseeuw P. *Finding groups in data*. Wiley–Interscience (Hoboken NJ), 1990.
Köppen M. The curse of dimensionality. *5th Online World Conference on Soft Computing in Industrial Applications* (WSC5), 2000.
Lee J. *Introduction to topological manifolds*, 2nd ed. Springer (Berlin), 2010.
Mirkin B. *Core concepts in data analysis: Summarization, correlation, and visualization*. Springer (Berlin), 2011.
Moisl H. *Cluster analysis for corpus linguistics*. De Gruyter (Berlin), 2015.
Peissig J. and Tarr M. Visual object recognition: Do we know more than we did 20 years ago? *Annual Review of Psychology*. 2007, 58: 75–96.
Strang G. *Introduction to linear algebra*, 5th ed., Wellesley-Cambridge Press (Cambridge), 2016.
Tabachnick B. and Fidell L. *Using multivariate statistics*. Pearson Education (London), 2007.
Tabak J. *Geometry: The language of space and form*. Facts on File (New York), 2011.
Tan P., Steinbach, M. and Kumar, V. *Introduction to data mining*. Pearson Addison Wesley (London), 2006.
Xu R. and Wunsch, D. *Survey of clustering algorithms*. IEEE Transactions on Neural Networks. 2005, 16: 645–78.